Electronic Colloquium on Computational Complexity, Report No. 143 (2012)

Direct Products in Communication Complexity

Mark Braverman[*] Anup Rao[†] Omri Weinstein[‡] Amir Yehudayoff[§]

November 4, 2012

Abstract

We give exponentially small upper bounds on the success probability for computing the direct product of any function over any distribution using a communication protocol. Let $\mathsf{suc}(\mu, f, C)$ denote the maximum success probability of a 2-party communication protocol for computing $f(x, y)$ with C bits of communication, when the inputs (x, y) are drawn from the distribution μ. Let μ^n be the product distribution on n inputs and f^n denote the function that computes n copies of f on these inputs.

We prove that if $T \log^{3/2} T \ll C\sqrt{n}$ and $\mathsf{suc}(\mu, f, C) < \frac{2}{3}$, then $\mathsf{suc}(\mu^n, f^n, T) \leq \exp(-\Omega(n))$. When μ is a product distribution, we prove a nearly optimal result: as long as $T \log^2 T \ll Cn$, we must have $\mathsf{suc}(\mu^n, f^n, T) \leq \exp(-\Omega(n))$.

1 Introduction

The *direct sum* question is about quantifying the resources needed to compute n independent copies of a function in terms of the resources needed to compute one copy of it. If one copy can be computed with C resources, then n copies can be computed using nC resources, but is this optimal?

When the inputs are drawn from a distribution (or the computational model is randomized), one can also measure the probability of success of computing the function. The *direct product* question is about understanding what the maximum probability of success of computing n copies of the function is. If there is a way to compute one copy with C resources and success probability ρ, then n copies can be computed using nC resources with success probability ρ^n, but is this optimal?

In this work, we study the direct product question in the model of distributional communication complexity [Yao79]. Direct sum theorems for this model were proved in [BBCR10], and we strengthen their results to give direct product theorems. For a longer introduction to direct sums and direct products in communication complexity and their significance, we refer the reader to the introductions of [BBCR10, JPY12].

[*] Department of Computer Science, Princeton University, mbraverm@cs.princeton.edu, Research supported in part by an Alfred P. Sloan Fellowship, an NSF CAREER award, and a Turing Centenary Fellowship.

[†] Computer Science and Engineering, University of Washington, anuprao@cs.washington.edu. Supported by an Alfred P. Sloan Fellowship, the National Science Foundation under agreement CCF-1016565, an NSF Career award, and by the Binational Science Foundation under agreement 2010089.

[‡] Department of Computer Science, Princeton University, oweinste@cs.princeton.edu.

[§] Department of Mathematics, Technion-IIT, Haifa, Israel, amir.yehudayoff@gmail.com. Horev Fellow – supported by the Taub Foundation. Supported by the Israel Science Foundation and by the Binational Science Foundation under agreement 2010089.

ISSN 1433-8092

Let $\mathsf{suc}(\mu, f, C)$ denote the maximum success probability of a 2-party communication protocol of communication complexity C for computing a function $f(x,y)$ when the inputs are drawn from the distribution μ. Let $f^n(x_1,\ldots,x_n,y_1,\ldots,y_n)$ denote the function that maps its inputs to the tuple $(f(x_1,y_1), f(x_2,y_2),\ldots,f(x_n,y_n))$ and μ^n denote the product distribution on n pairs of inputs, where each pair is sampled independently according to μ. Our goal in this work is to prove new upper bounds on $\mathsf{suc}(\mu^n, f^n, T)$ for $T \gg C$. It is easy to prove that $\mathsf{suc}(\mu^n, f^n, nC) \geq \mathsf{suc}(\mu^n, f^n, C)^n$, and $\mathsf{suc}(\mu^n, f^n, C) \leq \mathsf{suc}(\mu, f, C)$. Shaltiel [Sha03] showed that there exist μ, f, C such that $\mathsf{suc}(\mu^n, f^n, \frac{3}{4}nC) \geq \frac{3}{4}$, even though $\mathsf{suc}(\mu, f, C) \leq \frac{2}{3}$. Roughly, his ideas show that if $T \geq 2(1 - \mathsf{suc}(\mu, f, C))Cn$, there are examples where $\mathsf{suc}(\mu^n, f^n, T) > \mathsf{suc}(\mu, f, C)$.

Much past work has found success in proving upper bounds on $\mathsf{suc}(\mu^n, f^n, T)$ in special cases: for example, when f is the disjointness function [Kla10], or f is known to have small discrepancy [Sha03, LSS08, She11], or have a smooth rectangle bound [JY12], or the protocols computing f^n and f are restricted to using a bounded number of rounds of interaction [JPY12, MWY13], or restricted to behaving somewhat independently on each coordinate of the input [PRW97]. We refer the reader to [BBCR10, JPY12] for more references.

Prior to our work, the only known general upper bounds on $\mathsf{suc}(\mu^n, f^n, T)$, for $T > C$, are a consequence of the direct sum theorem proved in [BBCR10]: If $\mathsf{suc}(\mu, f, C) \leq \frac{2}{3}$, then $\mathsf{suc}(\mu^n, f^n, T) \leq \frac{3}{4}$, as long as $T \log T \ll C\sqrt{n}$. They also proved the same upper bound when $T\mathsf{polylog}(T) \ll Cn$ and μ is a product distribution.

In this work, we give new upper bounds that are exponentially small in n. When $\mathsf{suc}(\mu, f, C) \leq \frac{2}{3}$, we prove that $\mathsf{suc}(\mu^n, f^n, T) \leq \exp(-\Omega(n))$, as long as $T \log^{3/2} T \ll C\sqrt{n}$. By Yao's minimax principle [Yao79], we get an analogous statement for randomized worst case computation. If $\mathsf{suc}(f, C)$ denotes the maximum success probability for the best C-bit public coin randomized protocol computing f in the worst case, and if $\mathsf{suc}(f, C) \leq \frac{2}{3}$, then $\mathsf{suc}(f^n, T) \leq \exp(-\Omega(n))$ as long as $T \log^{3/2} T \ll C\sqrt{n}$. Formally, we prove:

Theorem 1 (Main Theorem)**.** *There is a universal constant $\alpha > 0$ such that if $\gamma = 1 - \mathsf{suc}(\mu, f, C)$, $T \geq 2$, and $T \log^{3/2} T < \alpha\gamma^{5/2}C\sqrt{n}$, then $\mathsf{suc}(\mu^n, f^n, T) \leq \exp\left(-\alpha\gamma^2 n\right)$.*

When μ is a product distribution, we prove an almost optimal result. We show that if $\mathsf{suc}(\mu, f, C) \leq \frac{2}{3}$ and $T \log^2 T \ll Cn$, then $\mathsf{suc}(\mu^n, f^n, T) \leq \exp(-\Omega(n))$.

Theorem 2 (Main Theorem for Product Distributions)**.** *There is a universal constant $\alpha > 0$ such that for every product distribution μ, if $\gamma = 1 - \mathsf{suc}(\mu, f, C)$, $T \geq 2$, and $T \log^2 T \leq \alpha\gamma^6 Cn$, then $\mathsf{suc}(\mu^n, f^n, T) \leq \exp\left(-\alpha\gamma^2 n\right)$.*

Our proofs heavily rely on methods from information theory [Sha48] which have been applied to a variety of problems in communication complexity [Raz92, NW93, Abl96, CSWY01, BYJKS04, BBCR10], and ideas developed to prove the parallel repetition theorem [Raz98, Hol07]. We give an overview of our proofs next.

1.1 Overview of the Proofs

The notation used below is formally defined in Section 2. Before we describe our proof in detail, we give a high level overview of the proof of the direct sum theorem proved in [BBCR10]. The theorem is proved by reduction. For T, C roughly as in the theorems above, they show that any protocol π for computing n copies of f with communication complexity $\|\pi\| = T$ can be used to obtain a protocol for computing one copy, with communication complexity less than C. This proves that computing n copies requires communication complexity more than T. The reduction itself has two steps. In the first step, they show that π can be used to obtain a protocol for computing f with small *information cost* (which we discuss below). In the second step, they show that any protocol with small information cost can be compressed to obtain a protocol that actually has small communication.

[CSWY01] were the first to define the information cost of protocols. Let the inputs to a protocol be X, Y, the messages be M and the public randomness be R. The *external information cost* [CSWY01] of the protocol is the mutual information between the inputs and the messages, conditioned on the public randomness: $I(XY; M|R)$. It is the information that an observer learns about the inputs by watching the execution of the protocol. The *internal information cost* [BYJKS04, BBCR10] of the protocol is defined to be $I(X; M|YR) + I(Y; M|XR)$. It is the information learnt by the parties about each others inputs during the execution of the protocol. The external information is always at least as large as the internal information.

The first step of the reduction in [BBCR10] gives a protocol with internal information cost bounded by $\sim T/n$ and communication bounded by T. In the second step, they show that any protocol with internal information I and communication N can be compressed to get a protocol with communication $\sim \sqrt{I \cdot N}$. Thus one obtains a protocol with communication $\sim T/\sqrt{n}$ for computing f. When μ is a product distribution, the first step of the reduction gives a protocol with external information cost bounded by $\sim T/n$. They show how to compress any protocol with small external information almost optimally, and so obtain a protocol with communication $\sim T/n$ for computing f. In both cases, the intuition for the first step of the reduction is that the T bits of the messages can reveal at most $\sim T/n$ bits of information about an average input coordinate.

To prove our direct product theorems, we modify the approach above using ideas inspired by the proof of the parallel repetition theorem [Raz98]. Let E be the event that π correctly computes f^n. For $i \in [n]$, let W_i denote the event that the protocol π correctly computes $f(x_i, y_i)$. Let $\pi(E)$ denote the probability of E, and let $\pi(W_i|E)$ denote the conditional probability of the event W_i given E. We shall prove that if $\pi(E)$ is not very small, then $(1/n)\sum_i \pi(W_i|E) < 1$, which is a contradiction. In fact, we shall prove that this holds for an arbitrary event W, not just E.

Lemma 3 (Main Lemma)**.** *There is a universal constant $\alpha > 0$ so that the following holds. For every $\gamma > 0$, and event W such that $\pi(W) \geq 2^{-\gamma^2 n}$, if $\|\pi\| \geq 2$, and $\|\pi\| \log^{3/2} \|\pi\| < \alpha\gamma^{5/2} C\sqrt{n}$, then $(1/n)\sum_{i\in[n]} \pi(W_i|W) \leq \mathsf{suc}(\mu, f, C) + \gamma/\alpha$.*

Lemma 4 (Main Lemma for Product Distributions)**.** *There is a universal constant $\alpha > 0$ such that if μ is a product distribution, the following holds. For every $\gamma > 0$, and event W such that $\pi(W) \geq 2^{-\gamma^2 n}$, if $\|\pi\| \geq 2$, and $\|\pi\| \log^2 \|\pi\| \leq \alpha\gamma^6 Cn$, then $(1/n)\sum_{i\in[n]} \pi(W_i|W) \leq \mathsf{suc}(\mu, f, C) + \gamma/\alpha$.*

The proofs of the lemmas proceed by reduction, and can be broken up into two steps as in [BBCR10]. However there are substantial differences in our proof, which are discussed in detail below. First let us see how Lemma 3 implies Theorem 1. Theorem 2 follows from Lemma 4 in the same way.

Proof of Theorem 1. Let E denote the event that π computes f correctly in all n coordinates. So, $(1/n)\sum_{i\in[n]} \pi(W_i|E) = 1$. Set $\gamma = \alpha(1 - \mathsf{suc}(\mu, f, C))/2$ so that $\mathsf{suc}(\mu, f, C) + \gamma/\alpha < 1$. Then by Lemma 3, either $\|\pi\| < 2$, $\|\pi\| \log^{3/2} \|\pi\| < \alpha^{7/2} 2^{-5/2}(1-\mathsf{suc}(\mu, f, C))^{5/2} C\sqrt{n}$, or $\pi(E) \geq 2^{-\gamma^2 n}$. □

We give the formal proofs of the main lemmas in Section 3. At a high level, the proofs of the lemmas are quite similar to each other, though there are some technical differences. We discuss Lemma 4 first, which avoids some complications that come from the fact that the inputs are correlated under μ. We give a protocol with communication complexity C that computes f correctly with probability at least $(1/n)\sum_i \pi(W_i|W) - O(\gamma)$. Let m denote the messages of π, and $\pi(x_i y_i m)$ denote the joint distribution of x_i, y_i, m. For fixed x_i, y_i, let $\pi(m|x_i y_i W)$ denote the conditional distribution of m.

Using standard subadditivity based arguments, one can show that for average i, $\pi(x_i y_i|W) \overset{\gamma}{\approx} \pi(x_i y_i) = \mu(x_i y_i)$, where here the approximation is in terms of the ℓ_1 distance of the distributions. Intuitively, since W has probability $2^{-\gamma^2 n}$, it cannot significantly alter all n of the inputs. We can hope to obtain a protocol that computes $f(x, y)$ by picking a random i, setting $x_i = x, y_i = y$ and simulating the execution of π conditioned on the event W. There are two challenges that need to be overcome:

The protocol must simulate $\pi(m|x_i y_i W)$ In the probability space of π conditioned on W, the messages sent by the first party can become correlated with the input of the second party, even though they were initially independent. Thus (unlike in [BBCR10]), $\pi(m|x_i y_i W)$ is no longer distributed like the messages of a communication protocol, and it is non-trivial for the parties to sample a message from this distribution.

The protocol must communicate at most $C \ll |m|$ **bits** To prove the lemma, the parties need to sample m using communication that is much smaller than the length of m.

To solve the first challenge, we use a natural protocol θ, inspired by the work of [JPY12]. The parties publicly sample a uniformly random coordinate i in $[n]$ and set $x_i = x, y_i = y$. Each message m_j sent by the first party in π is sampled according to the distribution $\pi(m_j|m_{<j} x_i W)$, and each message sent by the second party is sampled according to the distribution $\pi(m_j|m_{<j} y_i W)$. We prove that for average i, $\theta(x_i y_i m) \overset{\gamma}{\approx} \pi(x_i y_i m|W)$. [JPY12] also analyzed this protocol and showed that for average i, $\theta(x_i y_i m) \overset{\gamma t}{\approx} \pi(x_i y_i m|W)$, where here t is the number of rounds of communication in π. Our bound is independent of t, a feature that is crucial for obtaining our results.

To solve the second challenge, we need to come up with a way to *compress* the protocol θ. To use the compression methods of [BBCR10], we need to bound the *external information cost* of θ. We did not succeed in bounding this quantity, and so cannot apply the compression methods of [BBCR10] directly. Instead, we are able to bound $I_\pi(X_i Y_i; M|W)$ for average i, the corresponding quantity for the variables in the probability space of π.

This does not show that the information cost of θ is small, even though the distribution of the variables in θ is close in ℓ_1 distance to the distribution of the corresponding variables of π conditioned on W. For example, suppose θ is such that with small probability the first party sends her own input, and otherwise she sends a random string. Then θ is close to a protocol that reveals 0 information, but its information cost may be arbitrarily large.

Nevertheless, we show that any protocol that is close to having small external information cost can be simulated by a protocol that actually has small external information cost. In our example from above, the first party can simulate the protocol θ bit by bit and decide to abort it if she sees that her transmissions are significantly correlated with her input. This does not change the protocol most of the time, but does significantly reduce the amount of information that is revealed. Our general solution is very similar to this. The parties simulate θ and abort the simulation if they find that they are revealing too much information. We prove that any protocol that is close to having low information can be simulated with small communication (the term "δ-simulates" in the theorem statement is formally defined in Subsection 2.2):

Theorem 5 (Simulation for External Information)**.** *Suppose θ is a protocol with inputs x, y, public randomness r, and messages m, and q is another distribution on these variables such that $\theta(xyrm) \overset{\epsilon}{\approx} q(xyrm)$. Then, there exists a protocol τ that $O(\epsilon)$-simulates θ with $\|\tau\| \leq 2\|\theta\|$ and*

$$I_\tau(XY; M|R) \leq 2\left(\frac{I_q(XY; M|R) + 1/(e \ln 2) + 2\log(\|\theta\| + 1)}{\epsilon}\right) + \log(\|\theta\| + 1) + 2\log(1/\epsilon) + 4.$$

We give the formal proof of Theorem 5 in Section 4.2. The final protocol computing f is obtained by compressing τ using the methods of [BBCR10].

The high level outline of the proof of Lemma 3 is similar to the proof of Lemma 4. When μ is not a product distribution, we obtain a bound on the internal information cost associated with π conditioned on W, namely we bound $I_\pi(X_i; M|Y_iW) + I_\pi(Y_i; M|X_iW)$. We are unable to prove an analogue of Theorem 5 for the internal information cost (and it remains an interesting open question whether such a theorem is true or not). Instead, to prove Lemma 3, we reanalyze the compression method of [BBCR10] for internal information cost, and show that it can be used here. We prove:

Theorem 6 (Compression for Internal Information)**.** *Suppose θ is a protocol so that $\|\theta\| \geq 2$ with inputs x, y and messages m, and q is another distribution on these variables such that $\theta(xym) \overset{\epsilon}{\approx} q(xym)$. Then, there exists a protocol τ that $O(\epsilon)$-simulates θ such that*

$$\|\tau\| \leq \frac{\log\|\theta\|\sqrt{(I_q(X; M|Y) + I_q(Y; M|X) + 1 + \log\|\theta\|) \cdot \|\theta\|}}{\epsilon^{3/2}}.$$

Remark 7. *Theorem 6 can also be used to compress protocols θ that have public randomness. Indeed if the inputs are x', y', the public randomness is r and the messages are m, one can set $x = x'r, y = y'r$. Then $I_q(X; M|Y) + I_q(Y; M|X) = I_q(X'; M|Y'R) + I_q(Y'; M|X'R)$, so one can apply the theorem.*

The intuition for the proof is quite similar to the intuition for the proof of Theorem 5. We show that the compression goes well most of the time, and there is a small probability that the messages of the protocol will lead to a failure in the simulation, but this does not affect the outcome of the simulation by much. We formally prove Theorem 6 in Section 4.1.

2 Preliminaries

2.1 Notation

Unless otherwise stated, logarithms in this text are computed base two. Random variables are denoted by capital letters and values they attain are denoted by lower-case letters. For example, A may be a random variable and then a denotes a value A may attain and we may consider the event $A = a$. Given $a = a_1, a_2, \ldots, a_n$, we write $a_{\leq i}$ to denote $a_1, \ldots, a_i$. We define $a_{>i}$ and $a_{\leq i}$ similarly.

We use the notation $p(a)$ to denote both the distribution on the variable a, and the number $\Pr_p[A = a]$. The meaning will usually be clear from context, but in cases where there may be confusion we shall be more explicit about which meaning is being used. We write $p(a|b)$ to denote either the distribution of A conditioned on the event $B = b$, or the number $\Pr[A = a|B = b]$. Again, the meaning will usually be clear from context. Given a distribution $p(a, b, c, d)$, we write $p(a, b, c)$ to denote the marginal distribution on the variables a, b, c (or the corresponding probability). We often write $p(ab)$ instead of $p(a, b)$ for conciseness of notation. If W is an event, we write $p(W)$ to denote its probability according to p. We denote by $\mathbb{E}_{p(a)}\left[g(a)\right]$ the expected value of $g(a)$ with respect to a distributed according to p.

For two distributions p, q, we write $|p(a) - q(a)|$ to denote the ℓ_1 distance between the distributions p and q. We write $p \overset{\epsilon}{\approx} q$ if $|p - q| \leq \epsilon$. Given distributions $p_1, \ldots, p_n$ and $q_1, \ldots, q_n$, we sometimes say "for average i, $p_i \overset{\gamma}{\approx} q_i$" when we mean that $(1/n)\sum_{i=1}^{n} |p_i - q_i| \leq \gamma$.

The *divergence* between p, q is defined to be

$$\mathsf{D}\left(\frac{p(a)}{q(a)}\right) = \sum_{a} p(a) \log \frac{p(a)}{q(a)}.$$

For three random variables A, B, C with underlying probability distribution $p(a, b, c)$, the *mutual information* between A, B conditioned on C is defined as

$$I_p(A; B|C) = \mathop{\mathbb{E}}_{p(cb)}\left[\mathsf{D}\left(\frac{p(a|bc)}{p(a|c)}\right)\right] = \mathop{\mathbb{E}}_{p(ca)}\left[\mathsf{D}\left(\frac{p(b|ac)}{p(b|c)}\right)\right] = \sum_{a,b,c} p(abc) \log \frac{p(a|bc)}{p(a|c)}.$$

We shall often work with multiple distributions over the same space. To avoid confusion, we shall always explicitly specify the distribution being used when computing the mutual information. We shall sometimes work with an event W. In this case, we denote $I_p(A; B|CW) = I_q(A; B|C)$ where $q(abc) = p(abc|W)$.

2.2 Communication Complexity

Given a protocol π that operates on inputs x, y drawn from a distribution μ using public randomness[1] r and messages m, we write $\pi(x, y, m, r)$ to denote the joint distribution of these variables. We write $\|\pi\|$ to denote the *communication complexity* of π, namely the maximum number of bits that may be exchanged by the protocol.

Our work relies heavily on ways to measure the information complexity of a protocol, (see [BBCR10, Bra12] and references within for a more detailed overview). The *internal information cost* of π is defined to be $I_\pi(X; M|YR) + I_\pi(Y; M|XR)$. The *external information cost* is $I_\pi(XY; M|R)$. The internal information cost is always at most the external information cost, and the two measures are equal when $\pi(x, y) = \pi(x)\pi(y)$ is a product distribution. Both measures are at most the communication complexity of the protocol.

[1] In our paper we define protocols where the public randomness is sampled from a continuous (i.e. non-discrete) set. Nevertheless, we often treat the randomness as if it were supported on a discrete set, for example by taking the sum over the set rather than the integral. This simplifies notation throughout our proofs, and does not affect correctness in any way, since all of our public randomness can be approximated to arbitrary accuracy by sufficiently dense finite sets..

Let $q(x, y, a)$ be an arbitrary distribution. We say that π *δ-simulates* q, if there is a function g and a function h such that

$$\pi(x, y, g(x, r, m)) \overset{\delta}{\approx} q(x, y, a) \ \text{ and } \ \pi(g(x, r, m) \neq h(y, r, m)) \leq \delta.$$

We say that π computes[2] f with success probability $1 - \delta$, if π δ-simulates $\pi(x, y, f(x, y))$. If λ is a protocol with inputs x, y, public randomness r' and messages m', we say that π δ-simulates λ if π δ-simulates $\lambda(x, y, (r', m'))$.

2.3 Useful Protocols

The following lemma was proved by Holenstein [Hol07].

Lemma 8 (Correlated Sampling). *Suppose Alice is given a distribution p and Bob a distribution q over a common universe. Then there is a randomized sampling procedure that allows Alice and Bob to use shared randomness to jointly sample elements A, B such that A is distributed according to p, B is distributed according to q, and $\Pr[A \neq B] = |p - q|$.*

We use the following lemma of Feige et al. [FPRU94]:

Lemma 9 (Location of First Difference). *There is a randomized public coin protocol with communication complexity $O(\log(k/\epsilon))$ such given two k-bit strings x, y as input, it outputs the first index $i \in [k]$ such that $x_i \neq y_i$ with probability at least $1 - \epsilon$, if such an i exists.*

The following compression theorem from [BBCR10] will be useful:

Theorem 10. *For every protocol π, and every $\epsilon > 0$, there exists a protocol λ that ϵ-simulates π with*

$$\|\lambda\| \leq O\left(\frac{I_\pi(XY; M|R) \cdot \log(\|\pi\|/\epsilon)}{\epsilon^2}\right).$$

2.4 Basic Lemmas

The proofs of the following two lemmas can be found in [CT91]:

Lemma 11 (Divergence is Non-negative). $\mathsf{D}\left(\frac{p(a)}{q(a)}\right) \geq 0.$

Lemma 12 (Chain Rule). *If $a = a_1, \ldots, a_s$, then*

$$\mathsf{D}\left(\frac{p(a)}{q(a)}\right) = \sum_{i=1}^{s} \mathop{\mathbb{E}}_{p(a_{<i})}\left[\mathsf{D}\left(\frac{p(a_i|a_{<i})}{q(a_i|a_{<i})}\right)\right].$$

Pinsker's inequality bounds statistical distance in terms of the divergence:

[2] Our definition is different (weaker) than assuming that the messages m determine the value of f. We just assume that the parties eventually know f. Under our definition the communication complexity of a function can be significantly smaller if f maps to a large set. We believe that this definition is the right definition for what it means to compute a function.

Lemma 13 (Pinsker)**.** *If* $p(b) = q(b)$*, then*

$$p(a,b) \overset{\sqrt{\mathbb{E}_{p(b)}\left[\mathsf{D}\left(\frac{p(a|b)}{q(a|b)}\right)\right]}}{\approx} q(a,b).$$

Proof. By Pinsker's inequality [CT91] and concavity of square root,

$$|p - q| = \mathbb{E}_{p(b)}|p(a|b) - q(a|b)| \leq \mathbb{E}_{p(b)}\sqrt{D(p(a|b)||q(a|b)|)} \leq \sqrt{\mathbb{E}_{p(b)}D(p(a|b)||q(a|b)|)}.$$

□

The following lemma bounds the probability of getting a large term in the divergence:

Lemma 14 (Reverse Pinsker)**.** *Let* $S = \left\{(a,b) : \log\frac{p(a|b)}{q(a|b)} > 1\right\}$*. Then,* $p(S) < 2|p(a,b) - q(a,b)|$*.*

Proof. Let $\epsilon = |p(a,b) - q(a,b)| = 2\max\{p(S') - q(S') : S'\}$. Thus,

$$\begin{aligned}
p(S) &\leq \epsilon/2 + q(S) \\
&< \epsilon/2 + (1/2)\sum_{(a,b)\in S} q(b)\cdot p(a|b) \\
&\leq \epsilon/2 + p(S)/2 + (1/2)\sum_{(a,b)\in S} |q(b) - p(b)|\cdot p(a|b) \\
&\leq \epsilon/2 + p(S)/2 + (1/2)\sum_{b} |q(b) - p(b)| \\
&\leq \epsilon + p(S)/2.
\end{aligned}$$

□

The following bounds the contribution of the negative terms to the divergence:

Lemma 15. *Let* $S = \{a : p(a) < q(a)\}$*. Then,* $\sum_{a\in S} p(a)\log\frac{p(a)}{q(a)} \geq -1/(e\ln 2)$*.*

Proof.

$$\begin{aligned}
\sum_{a\in S} p(a)\log\frac{p(a)}{q(a)} &= -p(S)\sum_{a\in S}\frac{p(a)}{p(S)}\log\frac{q(a)}{p(a)} \\
&\geq -p(S)\log\left(\sum_{a\in S}\frac{p(a)}{p(S)}\frac{q(a)}{p(a)}\right) && \text{by concavity of log} \\
&\geq p(S)\log p(S).
\end{aligned}$$

The minimum value of the function $x\ln x$ is $-1/e$. □

2.5 Inequalities that Involve Conditioning

The following lemmas bound the change in divergence when extra conditioning is involved.

Lemma 16. *Let W be an event and A, B, M be random variables in the probability space p. Then,*

$$\mathop{\mathbb{E}}_{p(bm|W)}\left[\mathsf{D}\left(\frac{p(a|bmW)}{p(a|b)}\right)\right] \le \log\frac{1}{p(W)} + I_p(A;M|BW).$$

Proof.

$$\begin{aligned}\mathop{\mathbb{E}}_{p(bm|W)}\left[\mathsf{D}\left(\frac{p(a|bmW)}{p(a|b)}\right)\right] &= \sum_{a,b,m} p(abm|W)\log\frac{p(a|bmW)}{p(a|b)}\\ &= \sum_{a,b} p(ab|W)\log\frac{p(a|bW)}{p(a|b)} + \sum_{a,b,m} p(abm|W)\log\frac{p(a|bmW)}{p(a|bW)}\\ &= \sum_{a,b} p(ab|W)\log\frac{p(W|ab)}{p(W|b)} + I_p(A;M|BW).\end{aligned}$$

The first term can be bounded by:

$$\begin{aligned}\sum_{a,b} p(ab|W)\log\frac{p(W|ab)}{p(W|b)} &\le \sum_{a,b} p(ab|W)\log\frac{1}{p(W|b)}\\ &= \sum_{b} p(b|W)\log\frac{1}{p(W|b)}\\ &\le \log\sum_{b}\frac{p(b|W)}{p(W|b)} && \text{by concavity of log}\\ &= \log\sum_{b}\frac{p(b)}{p(W)} = \log\frac{1}{p(W)}.\end{aligned}$$

□

Lemma 17 (Conditioning does not decrease divergence)**.**

$$\mathop{\mathbb{E}}_{p(b)}\left[\mathsf{D}\left(\frac{p(a|b)}{q(a)}\right)\right] \ge \mathsf{D}\left(\frac{p(a)}{q(a)}\right).$$

Proof.

$$\begin{aligned}
\mathop{\mathbb{E}}_{p(b)}\left[\mathsf{D}\left(\frac{p(a|b)}{q(a)}\right)\right] &= \sum_b p(b)\sum_a p(a|b)\log\frac{p(a|b)}{q(a)} \\
&= -\sum_a p(a)\sum_b p(b|a)\log\frac{q(a)}{p(a|b)} \\
&\geq -\sum_a p(a)\log\left(\sum_b p(b|a)\frac{q(a)}{p(a|b)}\right) && \text{by concavity of log} \\
&= -\sum_a p(a)\log\left(\sum_b \frac{p(b)q(a)}{p(a)}\right) \\
&= \mathsf{D}\left(\frac{p(a)}{q(a)}\right).
\end{aligned}$$

□

The following lemma gives a key estimate that is used crucially in our proof. It allows us to remove the effect of conditioning on an event W on the right argument of a divergence expression. The lemma states that, on average, $\mathsf{D}\left(\frac{p(a|brW}{p(a|rW)}\right)$ cannot be larger than $\mathsf{D}\left(\frac{p(a|brW)}{p(a|r)}\right)$. Intuitively this is true because in both cases the first distribution is conditioned on W, but in the second case the second distribution is not conditioned on W. The second part of the lemma shows that conditioning on an event W of probability 2^{-s} can create a mutual information of up to s between two formerly independent random variables.

Lemma 18. *Let W be an event and A, B, R be random variables. Then,*

$$I_p(A;B|RW) \leq \mathop{\mathbb{E}}_{p(br|W)}\left[\mathsf{D}\left(\frac{p(a|brW)}{p(a|r)}\right)\right].$$

If in addition $p(abr) = p(r)p(a|r)p(b|r)$, then

$$I_p(A;B|RW) \leq \mathop{\mathbb{E}}_{p(br|W)}\left[\mathsf{D}\left(\frac{p(a|brW)}{p(a|br)}\right)\right] \leq \log\frac{1}{p(W)}.$$

Proof.

$$\begin{aligned}
I_p(A;B|RW) &= \sum_{a,b,r} p(abr|W)\log\frac{p(a|brW)}{p(a|rW)} \\
&= \sum_{a,b,r} p(abr|W)\log\frac{p(a|brW)}{p(a|r)} + \sum_{a,r} p(ar|W)\log\frac{p(a|r)}{p(a|rW)}.
\end{aligned}$$

The second term is $-\mathbb{E}_{p(r|W)}\left[\mathsf{D}\left(\frac{p(a|rW)}{p(a|r)}\right)\right] \leq 0$. This proves the first part.

To prove the second part, observe that $p(a|r) = p(a|br)$. Lemma 16 (with M being the empty variable) implies that

$$\mathop{\mathbb{E}}_{p(br|W)}\left[\mathsf{D}\left(\frac{p(a|brW)}{\overline{\overline{p(a|br)}}}\right)\right] \le \log\frac{1}{p(W)}.$$

□

2.6 Variable Truncation

We shall need to analyze protocols that are statistically close to having low information. The following lemmas show that if a variable A is statistically close to having low information, then some prefix $A_{\le K}$ of A usually has low information. By truncating the variable to $A_{\le K}$, we obtain a new variable that is statistically close to the old one, yet has low information.

Lemma 19. *Let $p(a,b,c) \overset{\epsilon}{\approx} q(a,b,c)$. Then, $q\left(\log\frac{q(a|bc)}{q(a|c)} > \beta - 2\right) > p\left(\log\frac{p(a|bc)}{p(a|c)} > \beta\right) - 9\epsilon/2$, for every real β.*

Proof.

$$\log\frac{q(a|bc)}{q(a|c)} = \log\frac{p(a|bc)}{p(a|c)} - \log\frac{p(a|bc)}{q(a|bc)} - \log\frac{q(a|c)}{p(a|c)}.$$

By Lemma 14, $p(\log\frac{p(a|bc)}{q(a|bc)} > 1) < 2\epsilon$, and $q(\log\frac{q(a|c)}{p(a|c)} > 1) < 2\epsilon$. Thus,

$$\begin{aligned}
& q\left(\log\frac{q(a|bc)}{q(a|c)} > \beta - 2\right) \\
& \ge q\left(\log\frac{p(a|bc)}{p(a|c)} > \beta \text{ and } \log\frac{p(a|bc)}{q(a|bc)} \le 1\right) - q\left(\log\frac{q(a|c)}{p(a|c)} > 1\right) \\
& > p\left(\log\frac{p(a|bc)}{p(a|c)} > \beta \text{ and } \log\frac{p(a|bc)}{q(a|bc)} \le 1\right) - 5\epsilon/2 && \text{using } q \overset{\epsilon}{\approx} p \\
& \ge p\left(\log\frac{p(a|bc)}{p(a|c)} > \beta\right) - 9\epsilon/2.
\end{aligned}$$

□

Lemma 20 (Truncation Lemma)**.** *Let $p(a,b,c) \overset{\epsilon}{\approx} q(a,b,c)$ where $a = a_1, \ldots, a_s$. For every a,b,c, define k to be the minimum number j in $[s]$ such that*

$$\log\frac{p(a_{\le j}|bc)}{p(a_{\le j}|c)} > \beta.$$

If no such index exists, set $k = s+1$. Then,

$$p(k < s+1) < \frac{I_q(A;B|C) + \log(s+1) + 1/(e\ln 2)}{\beta - 2} + 9\epsilon/2.$$

Remark 21. *One can also prove that $I_p(A_{<K}, B|C) \le \beta + \log(s+1)$, in Lemma 20. We do not need this conclusion, so we omit its proof.*

Proof of Lemma 20. Define

$$H = \begin{cases} K, A_{\leq K} & \text{if } K \leq s, \\ \perp & \text{else.} \end{cases}$$

Then

$$\begin{aligned} I_q(A;B|C) + \log(s+1) &\geq I_q(AK;B|C) \\ &\geq I_q(H;B|C) && \text{since } AK \text{ determines } H \\ &= \sum_{h,b,c} q(hbc) \log \frac{q(h|bc)}{q(h|c)}. && (1) \end{aligned}$$

By Lemma 15, we know that the negative terms contribute at least $-1/(e \ln 2)$ to (1). We shall lower bound the contribution of the positive terms using $p(k < s+1)$. By Lemma 19,

$$q\left(\log \frac{q(h|bc)}{q(h|c)} > \beta - 2\right) > p\left(\log \frac{p(h|bc)}{p(h|c)} > \beta\right) - 9\epsilon/2. \qquad (2)$$

Observe that if $h = j, a_{\leq j}$ and $p(hbc) > 0$, then $p(K = j | A_{\leq j} = a_{\leq j}, bc) = 1$, and so:

$$\begin{aligned} \frac{p(h|bc)}{p(h|c)} &= \frac{p(A_{\leq j} = a_{\leq j}|bc)}{p(A_{\leq j} = a_{\leq j}|c)} \cdot \frac{p(K = j|A_{\leq j} = a_{\leq j}, bc)}{p(K = j|A_{\leq j} = a_{\leq j}, c)} \\ &= \frac{p(A_{\leq j} = a_{\leq j}|bc)}{p(A_{\leq j} = a_{\leq j}|c)} \cdot \frac{1}{p(K = j|A_{\leq j} = a_{\leq j}, c)} \\ &\geq \frac{p(A_{\leq j} = a_{\leq j}|bc)}{p(A_{\leq j} = a_{\leq j}|c)} > 2^{\beta}. \end{aligned}$$

So,

$$p\left(\log \frac{p(h|bc)}{p(h|c)} > \beta\right) \geq p(k < s+1). \qquad (3)$$

The sentence after (1), and equations (2), (3) imply

$$\begin{aligned} &I_q(A;B|C) + \log(s+1) + 1/(e \ln 2) > (\beta - 2)(p(k < s+1) - 9\epsilon/2) \\ &\Rightarrow p(k < s+1) < \frac{I_q(A;B|C) + \log(s+1) + 1/(e \ln 2)}{\beta - 2} + 9\epsilon/2. \end{aligned}$$

□

3 Proofs of the Main Lemmas

3.1 Proof of the Main Lemma for General Distributions

In this section we prove Lemma 3. We write $M = M_1, M_2, \ldots, M_{2t}$ to denote the messages in π. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be the inputs. We write $\overline{X} = X_1, \ldots, X_n$ and $\overline{Y} = Y_1, \ldots, Y_n$. Our first goal is to show that W does not change the distribution in a typical coordinate.

Lemma 22. *For average i, $\pi(x_i y_i) \overset{\gamma}{\approx} \pi(x_i y_i | W)$.*

The proof of Lemma 22 is somewhat standard, so we defer it to Section 6. Next we eliminate a corner case:

Lemma 23. *If $\|\pi\| \le \gamma^2 n$, then for average i, $\pi(m x_i y_i | W) \overset{\sqrt{2}\gamma}{\approx} \pi(m|W) \cdot \pi(x_i y_i)$.*

The proof of Lemma 23 is also a straightforward application of subadditivity, and we defer it to Section 6. Lemma 23 implies that if $\|\pi\| \le \gamma^2 n$, then a protocol with 0 communication can approximate the messages of π conditioned on W. So

$$(1/n)\sum_{i=1}^{n} \pi(W_i|W) - \gamma/\sqrt{2} \le \mathsf{suc}(\mu, f, 0) \le \mathsf{suc}(\mu, f, C),$$

which completes the proof. The more interesting case is when $\|\pi\| \ge \gamma^2 n$, and so we assume that this holds in the rest of this section.

Define

$$R_i = \overline{X}_{<i}, \overline{Y}_{>i}.$$

The random variable R_i helps to break the dependencies between Alice and Bob.

Consider the protocol η in Figure 1. We show that η computes f with good probability, although with a lot of communication. η runs protocol θ_i given in Figure 2 as a subroutine with inputs $(x_i, r'_i), (y_i, r''_i)$. Eventually, we shall argue that for average i, $\eta((x_i, r'_i), (y_i, r''_i)) \overset{O(\gamma)}{\approx} \theta_i((x_i, r'_i), (y_i, r''_i))$ and that θ_i is statistically close to having small internal information. We shall apply Theorem 6 to compress the communication to obtain our final protocol for computing f.

Protocol η for computing $f(x, y)$ when inputs are sampled according to μ.

1. Use shared randomness to sample i uniformly from $[n]$.
2. Alice sets $x_i = x$ and Bob sets $y_i = y$.
3. Alice and Bob use Lemma 8 to sample r_i: Alice uses the distribution $\pi(r_i | x_i W)$ and Bob uses the distribution $\pi(r_i | y_i W)$. Write r'_i to denote Alice's sample and r''_i to denote Bob's sample.
4. Alice and Bob run protocol θ_i from Figure 2 with inputs (x_i, r'_i) and (y_i, r''_i).

Figure 1: Protocol for computing f.

Lemma 24. *For average i,*

$$\pi(x_i y_i)\pi(r_i | x_i W) \overset{2\gamma}{\approx} \pi(x_i y_i r_i | W) \overset{2\gamma}{\approx} \pi(x_i y_i)\pi(r_i | y_i W).$$

The proof of Lemma 24 is a standard application of subadditivity, and we defer it to Section 6.

Claim 25. *For average i, $\theta_i(x_i y_i r_i m) \overset{\gamma}{\approx} \pi(x_i y_i r_i m | W)$.*

Protocol θ_i for computing $f(x_i, y_i)$ when inputs $(x_i, r'_i), (y_i, r''_i)$ are sampled according to $\pi((x_i, r_i), (y_i, r_i)|W)$.

Alice sends each message M_j, j odd, according to the distribution $\pi(m_j|x_ir'_im_{<j}W)$. Bob sends each message M_j, j even, according to the distribution $\pi(m_j|y_ir''_im_{<j}W)$.

Figure 2: Simulation in the i'th coordinate.

Proof. Consider

$$\begin{aligned}
&\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi(x_iy_ir_i|W)} \left[\mathsf{D}\left(\frac{\pi(m|x_iy_ir_iW)}{\theta_i(m|x_iy_ir_i)} \right) \right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{2t} \mathop{\mathbb{E}}_{\pi(x_iy_ir_im_{<j}|W)} \left[\mathsf{D}\left(\frac{\pi(m_j|m_{<j}x_iy_ir_iW)}{\theta_i(m_j|m_{<j}x_iy_ir_i)} \right) \right]. && \text{by the chain rule} \quad (4)
\end{aligned}$$

The odd j's correspond to the cases when Alice speaks. These terms contribute:

$$\begin{aligned}
&\sum_{i,\text{odd j}} \mathop{\mathbb{E}}_{\pi(x_iy_ir_im_{<j}|W)} \left[\mathsf{D}\left(\frac{\pi(m_j|m_{<j}x_iy_ir_iW)}{\pi(m_j|m_{<j}x_ir_iW)} \right) \right] \\
&= \sum_{i,\text{odd j}} I_\pi(M_j; Y_i|X_iR_iM_{<j}W) && (5) \\
&\leq \sum_{i,\text{odd j}} \mathop{\mathbb{E}}_{\pi(x_iy_ir_im_{<j}|W)} \left[\mathsf{D}\left(\frac{\pi(m_j|m_{<j}x_iy_ir_iW)}{\pi(m_j|m_{<j}x_iy_ir_i)} \right) \right], && \text{by Lemma 18}
\end{aligned}$$

where here we used that when Alice speaks, $\pi(m_j|m_{<j}x_iy_ir_i) = \pi(m_j|m_{<j}x_ir_i)$. Repeating the same argument for Bob's messages gives

$$\begin{aligned}
(4) &\leq \sum_{i=1}^{n} \sum_{j=1}^{2t} \mathop{\mathbb{E}}_{\pi(x_iy_ir_im_{<j}|W)} \left[\mathsf{D}\left(\frac{\pi(m_j|m_{<j}x_iy_ir_iW)}{\pi(m_j|m_{<j}x_iy_ir_i)} \right) \right] \\
&= \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi(x_iy_ir_i|W)} \left[\mathsf{D}\left(\frac{\pi(m|x_iy_ir_iW)}{\pi(m|x_iy_ir_i)} \right) \right] && \text{by the chain rule} \\
&\leq \gamma^2 n. && \text{by Lemma 16}
\end{aligned}$$

We apply Lemma 13 to conclude the proof. □

Claim 26. *The expression for the internal information cost according to π conditioned on W can be bounded:*

$$\begin{aligned}
\sum_{i=1}^{n} I_\pi(X_i; M|Y_iR_iW) + I_\pi(Y_i; M|X_iR_iW) &\leq 2\gamma^2 n + I_\pi(\overline{X}; M|\overline{Y}W) + I_\pi(\overline{Y}; M|\overline{X}W) \\
&\leq 2\gamma^2 n + \|\pi\|.
\end{aligned}$$

Proof.

$$\begin{aligned}
\sum_{i=1}^{n} I_\pi(X_i;M|Y_iR_iW) &= \sum_{i=1}^{n} I_\pi(X_i;M|\overline{X}_{<i}\overline{Y}_{\geq i}W) \\
&\leq \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi(m\overline{x}_{<i}\overline{y}_{\geq i}|W)}\left[\mathsf{D}\left(\begin{array}{c}\pi(x_i|m\overline{x}_{<i}\overline{y}_{\geq i}W)\\ \hline\hline \pi(x_i|\overline{x}_{<i}\overline{y}_{\geq i})\end{array}\right)\right] && \text{by Lemma 18} \\
&\leq \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi(m\overline{x}_{<i}\overline{y}|W)}\left[\mathsf{D}\left(\begin{array}{c}\pi(x_i|m\overline{x}_{<i}\overline{y}W)\\ \hline\hline \pi(x_i|\overline{x}_{<i}\overline{y}_{\geq i})\end{array}\right)\right] && \text{by Lemma 17} \\
&= \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi(m\overline{x}_{<i}\overline{y}|W)}\left[\mathsf{D}\left(\begin{array}{c}\pi(x_i|m\overline{x}_{<i}\overline{y}W)\\ \hline\hline \pi(x_i|\overline{x}_{<i}\overline{y})\end{array}\right)\right] \\
&= \mathop{\mathbb{E}}_{\pi(m\overline{y}|W)}\left[\mathsf{D}\left(\begin{array}{c}\pi(\overline{x}|m\overline{y}W)\\ \hline\hline \pi(\overline{x}|\overline{y})\end{array}\right)\right] && \text{by the chain rule} \\
&\leq \log(1/\pi(W)) + I_\pi(\overline{X};M|\overline{Y}W) && \text{by Lemma 16} \\
&\leq \gamma^2 n + I_\pi(\overline{X};M|\overline{Y}W).
\end{aligned}$$

Repeating the argument gives

$$\begin{aligned}
&\sum_{i=1}^{n} I_\pi(X_i;M|Y_iR_iW) + I_\pi(Y_i;M|X_iR_iW) \\
&\qquad \leq 2\gamma^2 n + I_\pi(\overline{X};M|\overline{Y}W) + I_\pi(\overline{Y};M|\overline{X}W) \leq 2\gamma^2 n + \|\pi\|,
\end{aligned}$$

where the last inequality is using the fact that $I_\pi(\overline{X};M|\overline{Y}W) + I_\pi(\overline{Y};M|\overline{X}W)$ can be viewed as the internal information cost of the deterministic protocol π operating on inputs drawn from the distribution $\pi(\overline{XY}|W)$. □

In the probability space of π, let i be a uniformly random coordinate in $[n]$, independent of all other variables. Let $x' = (i, x_i, r_i)$ and $y' = (i, y_i, r_i)$. Define the protocol θ that gets inputs (i, x_i, r'_i) and (i, y_i, r''_i), where the inputs are distributed according to $\pi((i, x_i, r_i), (i, y_i, r_i)|W)$, and executes $\theta_i((x_i, r_i), (y_i, r_i))$.

By Lemma 8, $\Pr_\eta[R'_i \neq R''_i] \leq 4\gamma$. Thus for average i, $\eta((x_i, r'_i), (y_i, r''_i)) \overset{8\gamma}{\approx} \eta((x_i, r'_i), (y_i, r'_i))$, where here $\eta((x_i, r'_i), (y_i, r'_i))$ denotes the distribution where Bob's sample for r is set to be the same as Alice's sample. By Lemma 24, $\eta(ixyr'_i) \overset{2\gamma}{\approx} \pi(ix_iy_ir_i|W)$. Therefore the protocol η can be viewed as executing θ as a subroutine with inputs that are $O(\gamma)$-close to $\theta(x', y')$.

Claim 25 implies that $\theta(x'y'm) \overset{\gamma}{\approx} \pi(x'y'm|W)$. Claim 26 implies that

$$\begin{aligned}
&I_\pi(X';M|Y'W) + I_\pi(Y';M|X'W) \\
&\qquad = \mathop{\mathbb{E}}_{i}\left[I_\pi(X_i;M|Y_iR_iW) + I_\pi(Y_i;M|X_iR_iW)\right] \\
&\qquad \leq 2\gamma^2 + \|\pi\|/n \leq 3\|\pi\|/n. && \text{since } \|\pi\| \geq \gamma^2 n
\end{aligned}$$

To prove Lemma 3, we apply Theorem 6 to conclude that there exists a protocol that $O(\gamma)$-simulates θ with communication at most

$$\frac{\log\|\pi\|\sqrt{3\|\pi\|/n + 1 + \log\|\pi\|)\|\pi\|}}{\gamma^{3/2}} < O\left(\frac{\|\pi\| \cdot \log^{3/2}\|\pi\|}{\sqrt{n}\gamma^{5/2}}\right) < C,$$

where the first inequality appealed to the fact that $\|\pi\|/n > \gamma^2$ and the second is by our choice of α in the statement of Lemma 3. The proof of Lemma 3 is complete, since η computes f with probability of success at least $(1/n)\sum_{i=1}^{n} \pi(W_i|W) - O(\gamma)$. □

3.2 Proof of the Main Lemma for Product Distributions

Here we prove Lemma 4. We assume the same setup as in the proof of Lemma 3. We shall describe a protocol for computing f that has small external information cost. We shall then appeal to the compression result of [BBCR10] to obtain our final protocol. As in the proof of Lemma 3, we assume that $\|\pi\| \geq \gamma^2 n$, since if not, the lemma holds using a trivial reduction. The protocol for the reduction is given in Figure 3.

Protocol θ for computing $f(x, y)$ when inputs are sampled according to product μ.

1. Use shared randomness to sample i uniformly from $[n]$.
2. Alice sets $x_i = x$ and Bob sets $y_i = y$.
3. Run protocol θ_i from Figure 4.

Figure 3: Protocol for computing f.

Protocol θ_i for computing $f(x_i, y_i)$ when inputs are sampled according to $\pi(x_i y_i|W)$, where μ is product.

Alice samples M_j, j odd, according to $\pi(m_j|x_i m_{<j} W)$. Bob samples M_j, j even, according to $\pi(m_j|y_i m_{<j} W)$.

Figure 4: Protocol for computing $f(x_i, y_i)$.

3.2.1 Analysis

As in the proof of Lemma 3, we show that the sampled messages are close to the intended distribution. The proof is similar to the proof of Claim 25.

Claim 27. *For average i, $\theta_i(x_i y_i m) \overset{\gamma}{\approx} \pi(x_i y_i m|W)$.*

Proof. Consider

$$\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi(x_i y_i|W)} \left[\mathsf{D}\left(\frac{\pi(m|x_i y_i W)}{\theta_i(m|x_i y_i)} \right) \right]$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{2t} \mathop{\mathbb{E}}_{\pi(x_i y_i m_{<j}|W)} \left[\mathsf{D}\left(\frac{\pi(m_j|m_{<j} x_i y_i W)}{\theta_i(m_j|m_{<j} x_i y_i)} \right) \right]. \qquad \text{by the chain rule} \qquad (6)$$

The odd j correspond to bits sent by Alice. These terms contribute

$$\begin{aligned}
&\sum_{i,\text{odd j}} \mathop{\mathbb{E}}_{\pi(x_iy_im_{<j}|W)} \left[\mathsf{D}\left(\frac{\pi(m_j|m_{<j}x_iy_iW)}{\pi(m_j|m_{<j}x_iW)}\right)\right] \\
&= \sum_{i,\text{odd j}} I_\pi(M_j;Y_i|X_iM_{<j}W) && (7) \\
&\le \sum_{i,\text{odd j}} \mathop{\mathbb{E}}_{\pi(x_iy_im_{<j}|W)} \left[\mathsf{D}\left(\frac{\pi(m_j|m_{<j}x_iy_iW)}{\pi(m_j|m_{<j}x_iy_i)}\right)\right], && \text{by Lemma 18}
\end{aligned}$$

where we used the fact that when Alice speaks, $\pi(m_j|m_{<j}x_iy_i) = \pi(m_j|m_{<j}x_i)$. Repeating the same argument for Bob's messages gives

$$\begin{aligned}
(6) &\le \sum_{i=1}^{n}\sum_{j=1}^{2t} \mathop{\mathbb{E}}_{\pi(x_iy_im_{<j}|W)} \left[\mathsf{D}\left(\frac{\pi(m_j|m_{<j}x_iy_iW)}{\pi(m_j|m_{<j}x_iy_i)}\right)\right] \\
&= \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi(x_iy_i|W)} \left[\mathsf{D}\left(\frac{\pi(m|x_iy_iW)}{\pi(m|x_iy_i)}\right)\right] && \text{by the chain rule} \\
&\le \gamma^2 n. && \text{by Lemma 16}
\end{aligned}$$

Lemma 13 complete the proof. □

Next observe that

$$\begin{aligned}
\sum_{i=1}^{n} I_\pi(X_iY_i;M|W) &\le \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi(m|W)} \left[\mathsf{D}\left(\frac{\pi(x_iy_i|mW)}{\pi(x_iy_i)}\right)\right] && \text{by Lemma 18} \\
&\le \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi(\overline{x}_{<i}\overline{y}_{<i}m|W)} \left[\mathsf{D}\left(\frac{\pi(x_iy_i|m\overline{x}_{<i}\overline{y}_{<i}W)}{\pi(x_iy_i|\overline{x}_{<i}\overline{y}_{<i})}\right)\right] && \text{by Lemma 17} \\
&= \mathop{\mathbb{E}}_{\pi(m|W)} \left[\mathsf{D}\left(\frac{\pi(\overline{xy}|mW)}{\pi(\overline{xy})}\right)\right] && \text{by the chain rule} \\
&\le \log(1/\pi(W)) + I(\overline{XY};M|W) \le \gamma^2 n + \|\pi\|. && \text{by Lemma 16}
\end{aligned}$$

Recall that the public randomness of θ is a uniformly random coordinate i. In the probability space of π, let i be a uniformly random coordinate independent of all the other variables. Set $X = X_i$ and $Y = Y_i$. By Lemma 22 and Claim 27, $\theta(ixym) \overset{O(\gamma)}{\approx} \pi(ixym|W)$. We can bound the information under π as follows:

$$I_\pi(XY;M|iW) = \mathop{\mathbb{E}}_{i}\left[I_\pi(X_iY_i;M|W)\right] \le \gamma^2 + \|\pi\|/n \le 2\|\pi\|/n,$$

where the final inequality follows from $\|\pi\| \ge \gamma^2 n$. We apply Theorem 5 to obtain a protocol τ simulating θ with error $O(\gamma)$, whose external information cost is $O\left(\frac{\|\pi\|/n+\log\|\pi\|}{\gamma}\right) = O\left(\frac{\|\pi\|\log\|\pi\|}{\gamma^3 n}\right)$, where here we used $\|\pi\| \ge 2$ and $\|\pi\| \ge \gamma^2 n$. Finally, we apply Theorem 10 with error parameter γ, to obtain a protocol that computes f with probability $(1/n)\sum_i \pi(W_i|W) - O(\gamma)$, with communication

$$O\left(\frac{\|\pi\| \cdot \log\|\pi\| \cdot \log(\|\pi\|/\gamma)}{\gamma^5 n}\right) \leq O\left(\frac{\|\pi\| \cdot \log^2\|\pi\|}{\gamma^6 n}\right) \leq C,$$

by our choice of α in Lemma 4. □

4 Proofs of the Compression/Simulation Theorems

4.1 Compressing Protocols that are Close to Low Internal Information

Here we prove Theorem 6, showing how to compress protocols that are close to having low internal information. For the rest of this proof, denote

$$I = I_q(X;M|Y) + I_q(Y;M|X).$$

The simulating protocol τ is given in Figure 5.

Protocol τ for simulating θ

Phase 1: For every binary string m' of length at most $\|\theta\|$, the parties use shared randomness to sample a uniformly random number $\rho_{m'} \in [0,1]$. Alice uses this number to compute a bit

$$a_{m'} = \begin{cases} 0 & \text{if } \theta(M_j = 0|xm') > \rho_{m'}, \\ 1 & \text{else.} \end{cases}$$

Similarly, Bob computes

$$b_{m'} = \begin{cases} 0 & \text{if } \theta(M_j = 0|ym') > \rho_{m'}, \\ 1 & \text{else.} \end{cases}$$

Phase 2: The parties repeat the following steps as long as at most $C = \frac{\log\|\theta\|\sqrt{(I+1+\log\|\theta\|)\|\theta\|}}{\epsilon^{3/2}}$ bits are communicated:

1. Alice computes the messages $a \in \{0,1\}^{\|\theta\|}$ defined inductively by $a_j = a_{a_{<j}}$ for each j. Similarly, Bob computes the messages $b \in \{0,1\}^{\|\theta\|}$ defined inductively by $b_j = b_{b_{<j}}$ for each j.
2. Alice and Bob use the protocol of Lemma 9 with error parameter 1/10 to find the smallest location j such that $a_j \neq b_j$. If j is odd Bob resets $b_{b_{<j}} = a_j$. If j is even, Alice resets $a_{a_{<j}} = b_j$. If no such j is found, the parties do nothing.

Alice (resp. Bob) considers the final a (resp. b) the simulated outcome of the protocol.

Figure 5: Compression according to internal information cost.

4.1.1 Analysis

The communication complexity of τ is bounded by C by definition. Define $m \in \{0,1\}^{\|\theta\|}$ inductively by:

$$m_j = \begin{cases} a_{m_{<j}} & \text{if } j \text{ is odd,} \\ b_{m_{<j}} & \text{if } j \text{ is even.} \end{cases}$$

The string m is the intended simulation that Alice and Bob should converge to at the end of τ. The first observation is that m is correctly distributed, i.e., as the messages of θ.

Claim 28. *$\tau(xym) = \theta(xym)$.*

Proof. By the definition of m, for odd j, $\theta(m_j|xym_{<j}) = \theta(m_j|xm_{<j}) = \tau(m_j|xym_{<j})$, and similarly for even j, $\theta(m_j|xym_{<j}) = \tau(m_j|xym_{<j})$. □

We shall argue that the probability $\tau(a = m = b)$ is very close to 1. Say that there is a *mistake at coordinate* j if $a_{m_{<j}} \neq b_{m_{<j}}$. The location of the first (uncorrected) mistake is exactly the same as the location of the first disagreement between a, b in Phase 2 of τ. As long as the number of successful executions of the algorithm from Lemma 9 in Phase 2 exceeds the number of mistakes in Phase 1, we will eventually have $a = m = b$. Let $\ell \geq \Omega(C/\log\|\theta\|)$ denote the number of times that the application of Lemma 9 is run in Phase 2. By the Chernoff bound, at least $\ell/2$ executions of the algorithm from Lemma 9 find the correct coordinate, except with probability $\exp(-\Omega(\ell))$.To complete the proof of the theorem, we shall argue that the number of mistakes is at most $\ell/2$ with high probability.

Let

$$\beta = \frac{I + 1/(e\ln 2) + \log(\|\theta\| + 1)}{\epsilon} + 2.$$

For any x, y, r, m, let k denote the smallest index j such that either

$$\log\frac{\theta(m_{\leq j}|xy)}{\theta(m_{\leq j}|x)} > \beta \text{ or } \log\frac{\theta(m_{\leq j}|xy)}{\theta(m_{\leq j}|y)} > \beta. \tag{8}$$

If no such index exists, define $k = \|\theta\| + 1$. The random variable[3] K is a function of X, Y, M.

Claim 29. *The expected number of mistakes up to the k'th coordinate is small:*

$$\mathop{\mathbb{E}}_{\theta}\left[|\{j < k : a_{m_{<j}} \neq b_{m_{<j}}\}|\right] \leq \sqrt{\frac{\beta \cdot \|\theta\|}{2}}.$$

Proof. Suppose j is odd. There is a mistake in the j'th step only when $\rho_{m_{<j}}$ lies in between $\theta(m_j|xym_{<j}) = \theta(m_j|xm_{<j})$ and $\theta(m_j|ym_{<j})$. The probability of a mistake in the j'th message contributing to the expectation is at most

$$(1/2)\sum_{x,y,m_{<j},k} \theta(xym_{<j}k) \cdot \mathbf{1}_{j<k} \cdot |\theta(m_j|xym_{<j}) - \theta(m_j|ym_{<j})|,$$

[3] Since it can be ambiguous whether the expression $p(m_k)$ refers to $p(M_K = m_k)$ or $p(M_k = m_k)$, we shall be more explicit with the notation in the rest of this section. However, observe that $p(m_k, k)$ has only one interpretation, so in such cases we use the more concise notation.

where $\mathbf{1}_{j<k}$ is the indicator variable for whether or not $j < k$. We bound this by

$$\begin{aligned}
&(1/2) \sum_{x,y,m_{<j},k} \theta(xym_{<j}k) \cdot \mathbf{1}_{j<k} \cdot \sqrt{\mathsf{D}\left(\frac{\theta(m_j|xym_{<j})}{\theta(m_j|ym_{<j})} \right)} && \text{by Lemma 13} \\
&\leq (1/2) \sqrt{\sum_{x,y,m_{<j},k} \theta(xym_{<j}k) \cdot \mathbf{1}_{j<k} \cdot \mathsf{D}\left(\frac{\theta(m_j|xym_{<j})}{\theta(m_j|ym_{<j})} \right)} && \text{by concavity} \\
&= (1/2) \sqrt{\sum_{x,y,m,k} \mathbf{1}_{j<k} \cdot \theta(xymk) \log \frac{\theta(m_j|xym_{<j})}{\theta(m_j|ym_{<j})}}.
\end{aligned}$$

A similar bound applies for even j, and the expected number of mistakes in the j'th step for all j is at most

$$(1/2) \sqrt{\sum_{x,y,m,k} \mathbf{1}_{j<k} \cdot \theta(xymk) \log \frac{\theta(m_j|xym_{<j})^2}{\theta(m_j|xm_{<j})\theta(m_j|ym_{<j})}}.$$

The expected number of mistakes before the k'th message is therefore at most

$$\begin{aligned}
&(1/2) \sum_j \sqrt{\sum_{x,y,m,k} \mathbf{1}_{j<k} \cdot \theta(xymk) \log \frac{\theta(m_j|xym_{<j})^2}{\theta(m_j|xm_{<j})\theta(m_j|ym_{<j})}} \\
&\leq (1/2) \sqrt{\|\theta\| \cdot \sum_{x,y,m,j,k} \mathbf{1}_{j<k} \cdot \theta(xymk) \log \frac{\theta(m_j|xym_{<j})^2}{\theta(m_j|xm_{<j})\theta(m_j|ym_{<j})}} && \text{by Cauchy-Schwartz} \\
&= (1/2) \sqrt{\|\theta\| \cdot \sum_{x,y,m,k} \theta(xymk) \log \frac{\theta(m_{<k}|xy)^2}{\theta(m_{<k}|x)\theta(m_{<k}|y)}} \leq \sqrt{\frac{\beta \cdot \|\theta\|}{2}}. && \text{by the definition of } k
\end{aligned}$$

□

Next we show that, with high probability, $k = \|\theta\| + 1$.

Claim 30. *$\theta(k \leq \|\theta\|) < 11\epsilon$.*

Proof. Define k_1 and k_2 to be the minimum indices so that

$$\log \frac{\theta(m_{\leq k_1}|xy)}{\theta(m_{\leq k_1}|x)} > \beta \quad \text{and} \quad \log \frac{\theta(m_{\leq k_2}|xy)}{\theta(m_{\leq k_2}|y)} > \beta,$$

respectively (if no such index exists, set the value to be $\|\theta\|+1$). Then $k = \min\{k_1, k_2\}$. By Lemma 20 we have

$$\theta(k_1 \leq \|\theta\|) < \frac{I_q(M;Y|X) + \log(\|\theta\| + 1) + 1/(e \ln 2)}{\beta - 2} + 9\epsilon/2 \leq 11\epsilon/2.$$

Similarly, $\theta(k_2 \leq \|\theta\|) < 11\epsilon/2$, and the claim is proved by the union bound. □

By Claim 28, Claim 29, Claim 30 and Markov's inequality, the probability that the number of mistakes in τ exceeds $\ell/2$ is at most $\frac{\sqrt{2\beta\cdot\|\theta\|}}{\ell}+11\epsilon$. The simulation, therefore, computes m except with probability

$$\frac{\sqrt{2\beta\cdot\|\theta\|}}{\ell}+11\epsilon+\exp(-\Omega(\ell))=O\left(\frac{\log\|\theta\|\sqrt{(I+1+\log\|\theta\|)\cdot\|\theta\|}}{\sqrt{\epsilon}C}+\epsilon\right)=O(\epsilon),$$

where here we used the fact that $C\neq 0$. □

4.2 Simulating Protocols that are Close to Low External Information

In this section we prove Theorem 5, showing that protocols that are statistically close to having low external information cost can be modified so that they actually have low external information cost.

4.2.1 The Simulating Protocol τ

The protocol τ is defined as follows. Let

$$\beta=\frac{I_q(XY;M|R)+1/(e\ln 2)+\log(\|\theta\|+1)}{\epsilon}+\log(1/\epsilon)+2.$$

The parties simulate θ, but before each message m_j sent by Alice, she checks whether the sequence of messages $m_{<j}$ sent by her so far, including the message m_j that will result from her transmission, satisfies

$$\sum_{d\leq j,d\text{ odd}}\log\frac{\theta(m_d|xrm_{<d})}{\theta(m_d|rm_{<d})}\leq\beta.$$

If this is not the case, she sends a bit e_j to Bob indicating that the protocol must be aborted. If the condition is met, she sends a bit e_j indicating that the protocol will continue, and then transmits the sampled bit m_j.

Similarly, before each message m_j sent by Bob, he checks whether the sequence of messages $m_{<j}$ sent by him so far, including the message m_j that will result from his transmission, satisfies

$$\sum_{d\leq j,d\text{ even}}\log\frac{\theta(m_d|yrm_{<d})}{\theta(m_d|rm_{<d})}\leq\beta.$$

If this is not the case, he sends a bit e_j to Alice indicating that the protocol must be aborted. If the condition is met, he sends a bit e_j indicating that the protocol will continue, and transmits the sampled bit m_j .

For clarity of notation, we accomplish the aborts by having Alice and Bob transmit 0's for the rest of the protocol, so that all full transcripts are of the same length. This gives $\|\tau\|\leq 2\|\theta\|$. The full transcript of the parties in τ is denoted by the random variables E,M, where E is the concatenation of all the abort bits E_j, and M is the protocol transcript of θ.

4.2.2 Analysis

For any x, y, r, m, let k denote the smallest index j such that either

$$\sum_{d \leq j, d \text{ odd}} \log \frac{\theta(m_d|xrm_{<d})}{\theta(m_d|rm_{<d})} > \beta \quad \text{or} \quad \sum_{d \leq j, d \text{ even}} \log \frac{\theta(m_d|yrm_{<d})}{\theta(m_d|rm_{<d})} > \beta. \tag{9}$$

If no such index, define $k = \|\theta\| + 1$. The random variable[4] K is a function of X, Y, R, M.

Claim 31. *For each $x, y, r, m_{<k}, k$, $\theta(xyrkm_{<k}) = \tau(xyrkm_{<k})$.*

Proof. Fix any $x, y, r, m_{<k}, k$. If there exists a $j < k$ such that the messages $m_{\leq j}$ cause an abort in the j'th step, then we must have that $\theta(xyrm_{<k}k) = 0 = \tau(xyrm_{<k}k)$. So, we can assume that there is no such j.

By definition, $\tau(xyr) = \theta(xyr)$. We prove by induction on j that for all $j < k$, $\tau(M_j = m_j|xyrm_{<j}) = \theta(m_j|xyrm_{<j})$. In τ the sender of the j'th message samples $M_j = m_j$ with probability $\theta(m_j|xyrm_{<j})$. Since we have assumed that the sender does not abort, this m_j is transmitted, and we have proved the inductive step. This shows that for every fixed k,

$$\tau(M_{<k} = m_{<k}, xyr) = \prod_{j<k} \tau(m_j|xyrm_{<j}) = \prod_{j<k} \theta(m_j|xyrm_{<j}) = \theta(M_{<k} = m_{<k}, xyr).$$

We have $\tau(K = k|M_{<k} = m_{<k}, xyr) = \theta(K = k|M_{<k} = m_{<k}, xyr)$, since both numbers are the probability that the k'th message leads to an abort. □

Claim 32. *$I_\tau(XY; ME|R) \leq 2\beta + \log(\|\theta\| + 1)$.*

Proof. The random variables $K, M_{<K}$ determine M, E in τ. Thus,

$$\begin{aligned}
&I_\tau(XY; ME|R) \\
&\leq I_\tau(XY; KM_{<K}|R) \\
&= I_\theta(XY; KM_{<K}|R) && \text{Claim 31} \\
&= \sum_{x,y,r,k,m_{<k}} \theta(xyrkm_{<k}) \log \frac{\theta(km_{<k}|xyr)}{\theta(km_{<k}|r)} \\
&= \sum_{x,y,r,k,m_{<k}} \theta(xyrkm_{<k}) \left(\log \frac{\theta(M_{<k} = m_{<k}|xyr)}{\theta(M_{<k} = m_{<k}|r)} + \log \frac{\theta(K = k|M_{<k} = m_{<k}, xyr)}{\theta(K = k|M_{<k} = m_{<k}, r)} \right). && (10)
\end{aligned}$$

[4] Since it can be ambiguous whether the expression $p(m_k)$ refers to $p(M_K = m_k)$ or $p(M_k = m_k)$, we shall be more explicit with the notation in the rest of this section. However, observe that $p(m_k, k)$ has only one interpretation, so in such cases we use the more concise notation.

The second term can be bounded as follows:

$$\begin{aligned}
&\sum_{x,y,r,k,m_{<k}} \theta(xyrkm_{<k}) \log \frac{\theta(K=k|M_{<k}=m_{<k},xyr)}{\theta(K=k|M_{<k}=m_{<k},r)} \\
&\qquad \le \sum_{r,k,m_{<k}} \theta(rkm_{<k}) \log \frac{1}{\theta(K=k|M_{<k}=m_{<k},r)} \\
&\qquad \le \log \sum_{r,k,m_{<k}} \frac{\theta(rkm_{<k})}{\theta(K=k|M_{<k}=m_{<k},r)} && \text{by concavity of log} \\
&\qquad = \log \sum_{r,k,m_{<k}} \theta(M_{<k}=m_{<k},r) = \log(\|\theta\|+1). && (11)
\end{aligned}$$

Next we bound the first term in (10):

$$\log\left(\frac{\theta(M_{<k}=m_{<k}|xyr)}{\theta(M_{<k}=m_{<k}|r)}\right) = \sum_{j<k,j \text{ odd}} \log \frac{\theta(m_j|xrm_{<j})}{\theta(m_j|rm_{<j})} + \sum_{j<k,j \text{ even}} \log \frac{\theta(m_j|yrm_{<j})}{\theta(m_j|rm_{<j})},$$

where here we used the fact that since θ is a protocol, each (odd) message m_j sent by Alice satisfies $\theta(m_j|xyrm_{<j}) = \theta(m_j|xrm_{<j})$, and that a similar statement holds for Bob's messages. Thus by the definition of K,

$$\sum_{x,y,r,k,m_{<k}} \theta(xyrkm_{<k}) \log \frac{\theta(M_{<k}=m_{<k}|xyr)}{\theta(M_{<k}=m_{<k}|r)} \le 2\beta. \tag{12}$$

Combining (10), (11) and (12), we conclude that $I_\tau(X;ME|R) \le 2\beta + \log(\|\theta\|+1)$. □

Next, we argue that the probability that the protocol aborts is small.

Claim 33. $\tau(k \le \|\theta\|) = \theta(k \le \|\theta\|) < 15\epsilon/2$.

Proof. For any x, y, r, m, let k' denote the smallest index such that

$$\log \frac{\theta(m_{\le k'}|xyr)}{\theta(m_{\le k'}|r)} = \sum_{j\le k',j \text{ odd}} \log \frac{\theta(m_j|xrm_{<j})}{\theta(m_j|rm_{<j})} + \sum_{j\le k',j \text{ even}} \log \frac{\theta(m_j|yrm_{<j})}{\theta(m_j|rm_{<j})} > \beta - \log(1/\epsilon).$$

If no such index, define $k' = \|\theta\|+1$. By Lemma 20, we have

$$\theta(k' \le \|\theta\|) < \frac{I_q(XY;M|R) + 1/(e\ln 2) + \log(\|\theta\|+1)}{\beta - 2 - \log(1/\epsilon)} + 9\epsilon/2 \le 11\epsilon/2. \tag{13}$$

We shall show that $\theta(k < k') < 2\epsilon$, which will complete the proof. Define

$$S_1 = \left\{(x,y,r,m) : k(x,y,r,m) \le \|\theta\| \text{ and } \sum_{d\le k,d \text{ odd}} \log \frac{\theta(m_d|xrm_{<d})}{\theta(m_d|rm_{<d})} \le -\log(1/\epsilon)\right\},$$

$$S_2 = \left\{(x,y,r,m) : k(x,y,r,m) \le \|\theta\| \text{ and } \sum_{d\le k,d \text{ even}} \log \frac{\theta(m_d|yrm_{<d})}{\theta(m_d|rm_{<d})} \le -\log(1/\epsilon)\right\}.$$

Observe that $k < k'$ implies that $(x, y, r, m) \in S_1 \cup S_2$. We shall prove that $\theta(S_1) \leq \epsilon$ and $\theta(S_2) \leq \epsilon$. Consider the distribution

$$\theta'(xyrm) = \theta(xyr) \cdot \prod_{d \text{ odd}} \theta(m_d | rm_{<d}) \cdot \prod_{d \text{ even}} \theta(m_d | yrm_{<d}).$$

Fix any $(x, y, r, m) \in S_1$, and let $k = k(x, y, r, m)$ be defined as above. We have:

$$\begin{aligned}
&\log \frac{\theta(km_{\leq k}|xyr)}{\theta'(km_{\leq k}|xyr)} \\
&= \sum_{d \leq k, d \text{ odd}} \log \frac{\theta(m_d|xrm_{<d})}{\theta(m_d|rm_{<d})} + \sum_{d \leq k, d \text{ even}} \log \frac{\theta(m_d|yrm_{<d})}{\theta(m_d|yrm_{<d})} + \log \frac{\theta(K = k|M_{\leq k} = m_{\leq k}, xyr)}{\theta'(K = k|M_{\leq k} = m_{\leq k}, xyr)} \\
&= \sum_{d \leq k, d \text{ odd}} \log \frac{\theta(m_d|xrm_{<d})}{\theta(m_d|rm_{<d})} \leq -\log(1/\epsilon).
\end{aligned}$$

Thus $\theta(xyrkm_{\leq k}) \leq \epsilon \cdot \theta'(xyrkm_{\leq k})$. So (here we set $k = k(x, y, r, m)$ in the sum):

$$\begin{aligned}
\theta(S_1) &= \sum_{(x,y,r,m) \in S_1} \theta(xyrm) \\
&= \sum_{(x,y,r,m) \in S_1} \theta(xyrkm_{\leq k}) \cdot \theta(m|xyrkm_{\leq k}) \\
&\leq \epsilon \sum_{(x,y,r,m) \in S_1} \theta'(xyrkm_{\leq k}) \cdot \theta(m|xyrkm_{\leq k}) \leq \epsilon.
\end{aligned}$$

A similar argument proves $\theta(S_2) \leq \epsilon$. Thus, by (13), we have that $\theta(k \leq \|\theta\|) \leq \theta(k' \leq \|\theta\|) + \theta(k < k') < 11\epsilon/2 + 2\epsilon = 15\epsilon/2$ as required. □

5 Open Problem: Direct Products for Information Complexity

Both the direct sum result of [BBCR10] and our direct product result rely on methods to compress protocols. So it is natural to ask whether our ability to prove direct product results is limited only by our ability to compress protocols with low information cost. In fact, information cost can be made into a meaningful complexity measure. The *information complexity* of a function f with respect to a distribution μ is the lowest internal information cost attainable by a protocol computing f with respect to μ and error 1/3 [BR11, Bra12]. It turns out that the amortized communication complexity of f is exactly equal to its information complexity [BR11]. [BW11, KLL$^+$12] showed that many communication lower bound techniques actually give lower bounds on the information complexity.

Given this new complexity measure, we might have hoped that direct sum and direct product theorems holds with respect to it. Indeed [BBCR10] show that an optimal direct sum theorem holds for information complexity. However, a direct product theorem (with small success probability) cannot hold, because of the following counterexample. Let f be a function with information complexity I. Consider the protocol that computes f^n as follows. Let $\epsilon > 0$ be an arbitrary parameter. With probability ϵ, the protocol executes n copies of the optimal protocol for computing f. With probability $1 - \epsilon$ the protocol transmits nothing and fails. This protocol computes f^n

with probability ϵ, yet its information complexity is at most ϵIn. For example, setting $\epsilon = 1/n$ shows that even without increasing the information complexity, one can compute f^n with success probability $1/n$.

The following question is still interesting, and may be easier than proving new direct product results for communication complexity:

Open Problem 34. *Is there a universal constant α such that if the information complexity of f with respect to the distribution μ is I, $T \geq 2$, and $T < \alpha In$, then $\mathsf{suc}(\mu^n, f^n, T) \leq \exp\left(-\alpha\gamma^2 n\right)$?*

A potential avenue of attack on Problem 34 would be to prove an analogue of Theorem 5 for general distributions μ, showing that any protocol that is close to having low internal information cost can be simulated by a protocol with low internal information cost. Thus, one can hope to solve Problem 34 without giving an improved compression scheme for internal information cost.

6 Proofs of the Standard Lemmas

Proof of Lemma 22.

$$\begin{aligned}
\gamma^2 n &\geq \log(1/\pi(W)) \\
&\geq \mathsf{D}\left(\frac{\pi(\overline{xy}|W)}{\pi(\overline{xy})}\right) && \text{by Lemma 16} \\
&= \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi(\overline{x}_{<i}\overline{y}_{<i}|W)}\left[\mathsf{D}\left(\frac{\pi(x_i y_i|\overline{x}_{<i}\overline{y}_{<i}W)}{\pi(x_i y_i|\overline{x}_{<i}\overline{y}_{<i})}\right)\right] && \text{by the chain rule} \\
&\geq \sum_{i=1}^{n} \mathsf{D}\left(\frac{\pi(x_i y_i|W)}{\pi(x_i y_i)}\right). && \text{by Lemma 17}
\end{aligned}$$

The lemma follows by Lemma 13. □

Proof of Lemma 23.

$$\begin{aligned}
2\gamma^2 n &\geq \log(1/\pi(W)) + I(XY; M|W) \\
&\geq \mathop{\mathbb{E}}_{\pi(m|W)}\left[\mathsf{D}\left(\frac{\pi(\overline{xy}|mW)}{\pi(\overline{xy})}\right)\right] && \text{by Lemma 16} \\
&= \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi(m\overline{x}_{<i}\overline{y}_{<i}|W)}\left[\mathsf{D}\left(\frac{\pi(x_i y_i|\overline{x}_{<i}\overline{y}_{<i}mW)}{\pi(x_i y_i|\overline{x}_{<i}\overline{y}_{<i})}\right)\right] && \text{by the chain rule} \\
&\geq \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi(m|W)}\left[\mathsf{D}\left(\frac{\pi(x_i y_i|mW)}{\pi(x_i y_i)}\right)\right]. && \text{by Lemma 17}
\end{aligned}$$

Thus by Lemma 13, for average i, $\pi(x_i y_i m|W) \overset{\sqrt{2}\gamma}{\approx} \pi(x_i y_i) \cdot \pi(m|W)$. □

Proof of Lemma 24.

$$
\begin{aligned}
\gamma^2 n &\geq \log(1/\pi(W)) \\
&\geq \mathop{\mathbb{E}}_{\pi(y|W)}\left[\mathsf{D}\left(\frac{\pi(\overline{x}|\overline{y}W)}{\pi(\overline{x}|\overline{y})}\right)\right] && \text{by Lemma 16} \\
&= \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi(\overline{y}|W)}\left[\mathsf{D}\left(\frac{\pi(x_i|\overline{x}_{<i}\overline{y}W)}{\pi(x_i|\overline{x}_{<i}\overline{y})}\right)\right] && \text{by the chain rule} \\
&= \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi(\overline{y}|W)}\left[\mathsf{D}\left(\frac{\pi(x_i|\overline{x}_{<i}yW)}{\pi(x_i|y_i)}\right)\right] \\
&\geq \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi(\overline{y}|W)}\left[\mathsf{D}\left(\frac{\pi(x_i|\overline{x}_{<i}\overline{y}_{\geq i}W)}{\pi(x_i|y_i)}\right)\right] && \text{by Lemma 17} \\
&= \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi(r_iy_i|W)}\left[\mathsf{D}\left(\frac{\pi(x_i|r_iy_iW)}{\pi(x_i|y_i)}\right)\right].
\end{aligned}
$$

Lemma 22 and Lemma 13 imply that for average i,

$$
\begin{aligned}
\pi(x_iy_ir_i|W) &= \pi(y_i|W) \cdot \pi(r_i|y_iW) \cdot \pi(x_i|r_iy_iW) \\
&\overset{\gamma}{\approx} \pi(y_i|W) \cdot \pi(x_i|y_i) \cdot \pi(r_i|y_iW) \\
&\overset{\gamma}{\approx} \pi(x_iy_i) \cdot \pi(r_i|y_iW).
\end{aligned}
$$

□

Acknowledgements

We thank Paul Beame for useful comments.

References

[Abl96] F. Ablayev. Lower bounds for one-way probabilistic communication complexity and their application to space complexity. *Theoretical Computer Science*, 157(2):139–159, 1996.

[BBCR10] Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. How to compress interactive communication. In *Proceedings of the 2010 ACM International Symposium on Theory of Computing*, pages 67–76, 2010.

[BR11] Mark Braverman and Anup Rao. Information equals amortized communication. In Rafail Ostrovsky, editor, *FOCS*, pages 748–757. IEEE, 2011.

[Bra12] Mark 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.

[BW11] Mark Braverman and Omri Weinstein. A discrepancy lower bound for information complexity. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:164, 2011.

[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.

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

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

[FPRU94] Uriel Feige, David Peleg, Prabhakar Raghavan, and Eli Upfal. Computing with noisy information. *SIAM Journal on Computing*, 23(5):1001–1018, 1994.

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

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

[JY12] Rahul Jain and Penghui Yao. A strong direct product theorem in terms of the smooth rectangle bound. *CoRR*, abs/1209.0263, 2012.

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

[KLL$^+$12] Iordanis Kerenidis, Sophie Laplante, Virginie Lerays, Jérémie Roland, and David Xiao. Lower bounds on information complexity via zero-communication protocols and applications. *Electronic Colloquium on Computational Complexity (ECCC)*, 19:38, 2012.

[LSS08] Troy Lee, Adi Shraibman, and Robert Spalek. A direct product theorem for discrepancy. In *CCC*, pages 71–80, 2008.

[MWY13] Marco Molinaro, David Woodruff, and Grigory Yaroslavtsev. Beating the direct sum theorem in communication complexity with implications for sketching. In *SODA*, page to appear, 2013.

[NW93] Noam Nisan and Avi Wigderson. Rounds in communication complexity revisited. *SIAM Journal on Computing*, 22(1):211–219, February 1993.

[PRW97] Itzhak Parnafes, Ran Raz, and Avi Wigderson. Direct product results and the GCD problem, in old and new communication models. In *Proceedings of the 29th Annual ACM Symposium on the Theory of Computing (STOC '97)*, pages 363–372, New York, May 1997. Association for Computing Machinery.

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

[Raz98] Ran Raz. A parallel repetition theorem. *SIAM Journal on Computing*, 27(3):763–803, June 1998. Prelim version in STOC '95.

[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. Strong direct product theorems for quantum communication and query complexity. In *STOC*, pages 41–50, 2011.

[Yao79] Andrew Chi-Chih Yao. Some complexity questions related to distributive computing. In *STOC*, pages 209–213, 1979.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de