Interactive Information Complexity and Applications

Omri Weinstein

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Computer Science
Adviser: Mark Braverman

March 2015

Abstract

With applications in nearly every field of computer science, Communication Complexity constitutes one of the most useful methods for proving unconditional lower bounds - the holy grail of complexity theory. Developing tools in communication complexity is a promising approach for making progress in other computational models such as streaming, property testing, distributed computing, circuit complexity and data structures. One striking example of such tool is Shannon's information theory, introduced in the late 1940's in the context of the one way data transmission problem. While revealing the intimate connection between information and communication, Shannon's work and its classical extensions do not readily convert to interactive setups such as the communication complexity model, where two parties must engage in a multi-round conversation to accomplish some desirable interactive task. The research presented in this monograph aspires to extend Shannon's theory, develop the right tools, and understand how information behaves in interactive setups. The main measure if interest is the *Interactive Information Complexity* of a function, which informally captures the least amount of information the parties need to disclose each other about their inputs in order to solve the underlying task. We develop information-theoretic tools with applications to some of the most fundamental questions in communication complexity, including the limits of parallel computation, interactive compression, and the KRW conjecture. We then demonstrate the power of information complexity beyond communication complexity, with applications to various theoretical models, including data streaming, circuit lower bounds, privacy and economics. Lurking beneath these results is the fascinating question about the role of interaction and information in obtaining efficient outcomes, where efficiency may be measured in terms of social welfare, space, privacy or communication, depending on the model and context.

Acknowledgements

I am thankful every day for the privilege I had in being mentored by my advisor Mark Braverman. His utter commitment for my professional development, the freedom on one hand and the countless hours of devotion to our work on the other, and most of all his creativity have had a huge impact on me. Working with him was a truly humbling experience (sometimes to the point of frustration) and I could have never asked for a better inspiration.

I am particularly grateful to Anup Rao for his guidance, never-ending commitment and for paving the way to what became the core subject of this thesis. A special thanks to Oded Regev for his support and leap-of-faith in the past four years, and to Michael Saks, Sanjeev Arora and Zeev Dvir who agreed devote their time and serve on my committee.

I would like to deeply thank my collaborators over the past few year, whom I have had the true privilege of working with and learning from: To Noga Alon, Boaz Barak, Shahar Dobzinski, Michal Feldman Moran Feldman, Ankit Garg, Dmitry Gavinsky, Avinatan Hassidim, Young Kun-Ko, Edo Liberty, Shachar Lovett, Or Meir, Dana Moshkovitz, Rajsekar Manokaran, Yishay Mansour, Noam Nisan, Denis Pankratov, Ran Raz, Dana Ron, Ronitt Rubinfeld, Muli Safra, Alex Samorodnitsky, Inbal Talgam-Cohen , Moshe Tennenholtz, Avi Wigderson, David P. Woodruff and last but not least to Amir Yehudayoff.

Thanks to the CS department at Princeton, especially to Mitra Kelly, Melissa Lawson and Nicki Gotsis for bearing with such an astronaut like me. Thanks also to all my Princeton friends at the theory group and outside it, for the insightful hallway conversations, endless cups of coffee and humility that I admired so much. A special thanks to my dear friend Thiago Pereira for being there for me from day one. Thanks to Yuval Peres and Moshe Tennenholtz at Microsoft Research laboratories for hosting me in the vibrant and intellectual hubs of MSR Redmond and MSR Israel. A heartfelt thanks the Simons Foundation for supporting me through my graduate studies, I believe the impact of this foundation has the potential to revolutionize the entire field. I am also grateful to the

Siebel foundation for their generous scholarship. Lastly, a huge thanks to my dear parents Udi and Orna, my brother Asaf (for being my compass and showing me that there is more than meets the eye in everything in this world) and the rest of my family (Hagar and Hai), and to my friends for the constant encouragement and comic breaks. Last but not least, to my wife Lihi , my ray of light, for her love and understanding of the twisted minds of mathematicians.

To my parents, Orna and Udi.

Contents

Chapter 1

Introduction

The holy grail of complexity theory is proving lower bounds on different computational models, thereby delimiting computational problems according to the resources they require for solving. One of the most useful abstractions for proving such lower bounds is communication complexity. Since its introduction [154], this model has had a profound impact on nearly every field of theoretical computer science, including streaming algorithms, VLSI chip design, data structures, mechanism design and property testing [149, 130, 61, 25] to mention a few, and constitutes one of the few known tools for proving *unconditional* lower bounds. As such, developing new tools in communication complexity is a promising approaches for making progress within computational complexity, and in particular, for proving strong circuit lower bounds that appear viable (such as Karchmer-Wigderson games and ACC lower bounds [100, 20]).

One striking example for such a tool is information theory, introduced by Shannon in the late 1940s in the context of one-way communication problems [Sha48]. Shannon's noiseless coding theorem revealed the tight connection between communication and information, namely, that the amortized description length of a random one-way message (M) is equivalent to the amount of information it contains (its Entropy $H(M)$). In the 65 years that have passed since then, information theory has been widely applied and

developed, and has become the primary mathematical tool for analyzing communication problems.

Although classical information theory provides a complete understanding of the one-way transmission setup (where only one party speaks), it does not readily convert to *interactive setups*, such as the communication complexity model. Our research goal is to extend Shannons theory, develop appropriate machinery and understand how information behaves in interactive setups, where two (or more) parties must engage in a two-way conversation in order to accomplish some desirable task (e.g. compute a joint function $f(x,y)$ of their respective inputs). Our main measure of interest is the *Information Complexity* of a function $\mathsf{IC}_\mu(f,\varepsilon)$, which informally measures the average amount of information the players need to reveal each other about their inputs in order to solve f with some prescribed error under the input distribution μ.

From this perspective, communication complexity can be viewed as the extension of transmission problems to general tasks performed by two (or more) parties over a channel. Surprisingly, it turns out that an analogue of Shannons noiseless coding theorem does in fact hold for interactive computation, asserting that the amortized communication cost of computing many independent instances of any function f scales as its information complexity:

Theorem 1.0.1 ([35]). *For any $\varepsilon > 0$ and any two-party communication function $f(x,y)$, it holds that*

$$\lim_{n\longrightarrow\infty} \frac{\mathsf{D}_{\mu^n}(f^n,\varepsilon)}{n} = \mathsf{IC}_\mu(f,\varepsilon).$$

This theorem, which plays a central role in this thesis, assigns an *operational* meaning to $\mathsf{IC}_\mu(f,\varepsilon)$ (namely, one which is grounded in reality) and insinuates that information theory is the "right" tool for studying communication problems. The results we present below provide further evidence for this intuition (a notable one is asserts that Theorem 1.0.1 is in fact a "sharp-threshold" characterization of amortized computation, see Chapter 3).

Broadly speaking, Shannon's information theory has two general benefits in addressing communication problems. Firstly, it gives us a set of simple yet powerful tools for reasoning about transmission problems and more broadly about relationships between interdependent random variables. Tools that include mutual information, the chain rule, and the data processing inequality [58]. Indeed, a remarkable feature of information complexity, which stems from these simple tools, is that it is a fully additive measure over composition[1] of tasks:

$$\mathsf{IC}_{\mu_1 \times \mu_2}(T_1 \otimes T_2) = \mathsf{IC}_{\mu_1}(T_1) + \mathsf{IC}_{\mu_2}(T_2). \tag{1.1}$$

It is this benefit that has been primarily used in prior works involving information-theoretic tools in communication complexity [1, 52, 116, 49, 14, 96, 16]. Secondly, in the context of transmission problems – starting with Shannon's noiseless coding theorem – information theory is known to give tight precise bounds on rates and capacities. In fact, unlike computational complexity, where we often ignore linear, polylogarithmic, and sometimes even polynomial factors, a large fraction of results in information theory provide us with precise answers up to additive lower-order terms. For instance, we know that a sequence of random digits would take exactly $\log_2 10 \approx 3.322$ bits per digit, and that the capacity of a binary symmetric channel with substitution probability 0.2 is exactly $1 - H(0.2) \approx 0.278$ bits per symbol. A program which has emerged in the field over the past few years is to understand whether such fundamental results translate into the interactive setup. While this program is only at its preliminary stage, we provide encouraging results in this direction – For example, the tools we developed enabled us to determine the *exact* communication complexity of the Set-Disjointness function, which turns out to be surprisingly low: $C_{DISJ} \cdot n \approx 0.48n$ (see Section 2.3). Such precise results were beyond the reach of analytical techniques before the emergence of information

[1] $T1 \otimes T2$ denotes the task composed of successfully performing both T_1 and T_2 on the respective inputs $(X1, Y1) \sim \mu_1$ and $(X2, Y2) \sim \mu_2$.

complexity.

One caveat is that mathematically striking characterizations such as Shannon's noiseless coding theorem only become possible in the limit, where the size of the message we are trying to transmit over the channel – i.e. the block-length – grows to infinity. One exception is Huffman coding [86], where it was shown that the expected number of bits $C(M)$ needed to transmit a *single sample* from M, is very close (yet not equal!) to the optimal rate ($H(M) \leq C(M) \leq H(M)+1$). What happens for small block lengths is important for obvious practical and theoretical reasons, and even more so in the interactive regime (see e.g., [16, 40]). Indeed, this distinction between amortized and so called "one-shot" results is one of the main distinguishing features of information complexity from classic information theory. Another distinctive aspect is in that communication complexity often studies functions whose output is only a single bit or a small number of bits, thus counting style direct lower bound proofs rarely apply.

Organization

The results of this dissertation are divided into two main categories. The first set of results (Part 1) concerns direct applications of information complexity to communication complexity: New techniques for obtaining strong information and communication lower bounds, with applications to some of the most well studied problems in the literature (the Gap-Hamming,Inner-Product, Greater-Than,Set-Disjointness and Set Intersection problems); New advances on the *Direct Sum* and *Direct Product* conjectures and the closely related interactive compression problem (a quest for an interactive analogue of Huffman coding). A by-product of these results is the development of many new tools, frameworks and understanding of how information behaves in interactive setups.

The second set of results (Part 2) describes applications of information complexity to various fields and computational models, including circuit complexity, data streaming,

security and economics. The common feature of these (seemingly disparate) models is that they all involve an interactive system in which information is distributed among multiple agents who are required to solve or optimize some objective function. These agents may be honest, strategic or even adversarial (malicious). We explore the role of information and interaction in obtaining efficient solutions for those various interactive systems, where efficiency may be measured in terms of communication, social welfare, revenue, space or privacy, depending on the context.

1.1 Preliminaries and Background

The following technical background contains basic definitions, notations and facts used throughout this monograph. Additional notations with a specific scope are defined locally in each respective section. For a more thorough and detailed treatment of communication complexity and information theory, we refer the reader to the excellent texts by Cover and Thomas [59] and by Kushilevits and Nisan [111].

1.1.1 Information Theory

Definition 1.1.1 (Entropy)**.** *The* Shannon entropy *of a random variable X is*

$$H(X) := \sum_x \Pr[X = x] \log(1/\Pr[X = x]).$$

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

Fact 1.1.2 (Entropy Chain Rule)**.** $H(AB) = H(A) + H(B|A)$.

Definition 1.1.3 (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 1.1.4 (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 1.1.5 (Kullback-Leiber Divergence)**.** *The informational divergence between two distributions is*

$$\mathbb{D}(p\|q) := \sum_x p(x)\log(p(x)/q(x)).$$

By a slight abuse of notation, when A and B are two random variables, we write $\mathbb{D}(A\|B) := \sum_x A(x)\log(A(x)/B(x))$ to mean the divergence between the corresponding distribution of A and B.

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

Proposition 1.1.6. *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 $I(A;B|C) = \mathbb{E}_{a,c\in_R A,C}[\mathbb{D}(B_{ac}\|B_c)]$.*

1.1.2 Communication Complexity

Let $\mathcal{X}, \mathcal{Y}$ denote the set of possible inputs to the two players, who we call Alice and Bob. A *private-coin protocol* for computing a function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}_K$ is a rooted tree with the following structure:

- Each non-leaf node is *owned* by Alice or by Bob.

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

On input x, y, the protocol π is executed as in Figure 1.1.

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

Figure 1.1: A communication protocol.

A *public coin* protocol is a distribution on private coins protocols, run by first using shared randomness to sample a random string R and then running the corresponding private coin protocol π_R. Every private coin protocol is thus a public coin protocol (however, the distinction between private and public coin protocols will be crucial when dealing with information complexity, as elaborated in Subsection 1.1.6). The protocol is called deterministic if all distributions labeling the nodes have support size 1.

Definition 1.1.7 (Communication Complexity notation)**.** *For a function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, a distribution $\mu \in \Delta\mathcal{X} \times \mathcal{Y}$ supported on $\mathcal{X} \times \mathcal{Y}$, and a parameter $\varepsilon > 0$, $\mathsf{D}_\mu(f, \varepsilon)$ denotes the communication complexity of the cheapest deterministic protocol computing f on inputs sampled*

according to μ with error (at most) ε. We call this the distributional communication complexity of *f with respect to μ. $\mathsf{R}(f, \varepsilon)$ denotes the cost of the best* randomized *public coin protocol for computing f with error at most ε over* all *possible inputs. We call this the* randomized communication complexity of *f. When measuring the communication cost of a particular protocol π, we sometimes use the notation $\|\pi\|$ for brevity, to denote the* maximum *length of a path in the protocol tree of π. When clear from context, we use the notation $\mathsf{D}_\mu(f)$ for the* deterministic *communication complexity of f.*

The following theorem due to Yao, relates the randomized communication complexity of a function to its distributional communication complexity:

Theorem 1.1.8 (Yao's Min-Max)**.** $R_\rho(f) = \max_\mu D^\mu_\rho(f)$.

Given a communication protocol π, $\pi(x, y)$ denotes the concatenation of the public randomness with all the messages that are sent during the execution of π (for information purposes, this is without loss of generality, since the public string R conveys no information about the inputs). We call this the *transcript* of the protocol. When referring to the random variable denoting the transcript, rather than a specific transcript, we will use the notation $\Pi(x, y)$ — or simply Π when x and y are clear from the context.

1.1.3 Two-Party Information Complexity

This subsection introduces some of the central concepts used throughout this monograph. For a more detailed overview of information complexity, we refer the reader to [28].

We begin by defining the information cost of a protocol, which informally captures the (average) amount of additional information that Alice and Bob learn about each others inputs from the protocol π:

Definition 1.1.9. *The (internal) information cost* of a protocol ([14, 17]) over inputs drawn from a distribution μ on $\mathcal{X} \times \mathcal{Y}$, is given by:

$$\mathsf{IC}_\mu(\pi) := I(\Pi; X|Y) + I(\Pi; Y|X). \tag{1.2}$$

Intuitively, the definition in (1.2) captures how much the two parties learn about each other's inputs from the execution transcript of the protocol π. The first term captures what the second player learns about X from Π – the mutual information between the input X and the transcript Π given the input Y. Another information measure which makes sense at certain contexts is the *external* information cost of a protocol,

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

This definition captures what the first player learns about Y from Π. The second definition captures what an *external* observer learns on average from the transcript of π, about the inputs of both players. The latter quantity will be of minor interest in this writeup.

Note that the information of a protocol π depends on the prior distribution μ, as the mutual information between the transcript Π and the inputs depends on the prior distribution on the inputs. To give an extreme example, if μ is a singleton distribution, i.e. one with $\mu(\{(x, y)\}) = 1$ for some $(x, y) \in \mathcal{X} \times \mathcal{Y}$, then $\mathsf{IC}_\mu(\pi) = 0$ for all possible π, as no protocol can reveal anything to the players about the inputs that they do not already know *a-priori*. Similarly, $\mathsf{IC}_\mu(\pi) = 0$ if $\mathcal{X} = \mathcal{Y}$ and μ is supported on the diagonal $\{(x, x) : x \in \mathcal{X}\}$. Since one bit of information can never reveal more than one bit of communication, the communication cost $\|\pi\|$ of π is always an upper bound on its information cost over *any* distribution μ:

Lemma 1.1.10 ([35])**.** *For any distribution* μ, $\mathsf{IC}_\mu(\pi) \leq \|\pi\|$.

One can now define the *information complexity* of a function f with respect to μ and error ε as the least amount of information the players need to disclose each other in order to compute f with error at most ε:

Definition 1.1.11. *The Information Complexity* of f with respect to μ (and error ε) is

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

1.1.4 General Useful Facts

We denote by $|p-q|$ the *total variation* distance between the distributions p and q. Pinsker's inequality bounds statistical distance in terms of the divergence:

Lemma 1.1.12 (Pinsker's inequality)**.** $|p-q|^2 \leq \mathbb{D}\left(p\|q\right)$.

Lemma 1.1.13 (Mutual information in terms of Divergence)**.**

$$I(A;B|C) = \mathbb{E}_{b,c}\left[\mathbb{D}\left((A|bc)\|(A|c)\right)\right] = \mathbb{E}_{a,c}\left[\mathbb{D}\left((B|ac)\|(B|c)\right)\right].$$

Lemma 1.1.14 (Conditioning on independent variables increases information)**.** *Let A, B, C, D be four random variables in the same probability space. If A and D are conditionally independent given C, then it holds that $I(A;B|C) \leq I(A;B|CD)$.*

Proof. We apply the chain rule twice. On one hand, we have

$$I(A;BD|C) = I(A;B|C) + I(A;D|CB) \geq I(A;B|C)$$

since mutual information is nonnegative. On the other hand,

$$I(A;BD|C) = I(A;D|C) + I(A;B|CD) = I(A;B|CD)$$

since $I(A; D|C) = 0$ by the independence assumption on A and D. Combining both equations completes the proof. □

Fact 1.1.15 (Divergence is Non-negative)**.** $\mathbb{D}(p(a)\|q(a)) \geq 0$.

Fact 1.1.16 (Chain Rule)**.** *If $a = a_1, \ldots, a_s$, then*

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

Fact 1.1.17 (Convexity of divergence)**.** *Let $Q = \mathbb{E}_x[Q_x]$, P be two distributions. Then*

$$\mathbb{E}_x\left[\mathbb{D}(P\|Q_x)\right] \geq \left[\mathbb{D}(P\|Q)\right].$$

Fact 1.1.18 (Data Processing Inequality)**.** *Let $X \to Y \to Z$ be a Markov Chain in the same probability space (i.e., $Z \perp X|Y$). Then $I(X;Y) \geq I(X;Z)$.*

The following lemma asserts that if a random variable $Y = g(X)$ allows one to reconstruct X with high probability, then Y must "consume" most of the entropy of X:

Lemma 1.1.19 (Fano's Inequality)**.** *Let X be a random variable chosen from domain $\mathbf{X}$ according to distribution μ_X, and Y be a random variable chosen from domain $\mathcal{Y}$ according to distribution μ_Y . Then for any reconstruction function $g : Y \longrightarrow X$ with error ε_g, it holds that $H(X|Y) \leq H(\varepsilon_g) + \varepsilon_g \log(|\mathbf{X}| - 1)$.*

1.1.5 Additivity of Information Cost

Perhaps the single most remarkable property of information complexity is that it is a fully additive measure over composition of tasks. This property is what makes information complexity such a natural "relaxation" for addressing direct sum and product conjectures in communication complexity. The main ingredient of the following lemma appeared first in the works of [136, 134] and more explicitly in [17, 35, 28].

Lemma 1.1.20 (Additivity of IC)**.** $\mathsf{IC}_{\mu^n}(f^n, \varepsilon) = n \cdot \mathsf{IC}_{\mu}(f, \varepsilon)$.

Proof. The ($\leq$) direction of the lemma is easy, and follows from a simple argument that applies the single-copy optimal protocol independently to each copy of f^n, with independent randomness. We leave the simple analysis of this protocol as an exercise to the reader.

The ($\geq$) direction is the main challenge. Will will prove it in a contra-positive fashion: Let Π be an n-fold protocol for f^n, such that $\mathsf{IC}_{\mu^n}(f^n, \varepsilon) = I$. We shall use Π to produce a *single-copy* protocol for f whose information cost is $\leq I/n$, which would complete the proof. The guiding intuition for this is that Π should reveal I/n bits of information about an average coordinate.

To formalize this intuition, let $(x, y) \sim \mu$, and denote $\mathbf{X} := X_1 \dots X_n$, $X_{\leq i} := X_1 \dots X_i$, and similarly for $\mathbf{Y}, Y_{\leq i}$. Alice and Bob will "embed" their respective inputs (x, y) to a (publicly chosen) random coordinate $i \in [n]$ of Π. However, Π is defined over n input copies, so in order to execute it, the players need to "fill in" the rest $(n-1)$ coordinates, each according to μ. How should this step be done? The first attempt is for Alice and Bob to try and complete X_{-i}, Y_{-i} privately. This approach fails if μ is a non-product distribution, since there's no way the players can sample X and Y privately, such that $(X, Y) \sim \mu$ if μ correlates the inputs. The other extreme – sampling X_{-i}, Y_{-i} using public randomness only – would resolve the aforementioned correctness issue, but would leak too much information: An instructive example is where the first message of Π is the XOR of Alice's n-bit uniform input $M = X_1 \oplus X_2 \oplus \dots \oplus X_n$. Conditioned on X_{-i}, Y_{-i}, M reveals 1 bit of information about X_i to Bob, while we way to argue that in this case, only $1/n$ bits are revealed about X_i. It turns out that the "right" way of breaking the dependence across the coordinates is by sampling publicly the random variable

$$R := X_{<i}, Y_{>i}.$$

Note that given R, Alice can complete all her missing inputs $X_{>i}$ *privately* according to μ, and Bob can do the same for $Y_{<i}$. Let us denote by $\theta(x,y)$ the protocol produced by running $\Pi(X_1,...,X_{i-1},x,X_{i+1},...,X_n\ ,\ Y_1,...,Y_{i-1},y,Y_{i+1},...,Y_n)$ and outputting its answer on the i'th coordinate.

By definition, Π computes f^n with a *per-copy* error of ε, and thus in particular $\theta(x,y) = f(x,y)$ with probability $\geq 1-\varepsilon$. To analyze the information cost of θ, we write:

$$\begin{aligned} I(\theta;x|y) &= \mathbb{E}_R[I(\theta;x|y,R)] = \sum_{i=1}^{n} \frac{1}{n} \cdot I(\Pi;X_i \mid Y_i,R) \\ &= \frac{1}{n}\sum_{i=1}^{n} I(\Pi;X_i \mid Y_i,X_{<i}Y_{>i}) = \frac{1}{n}\sum_{i=1}^{n} I(\Pi;X_i \mid X_{<i}Y_{\geq i}) \\ &\leq \frac{1}{n}\sum_{i=1}^{n} I(\Pi;X_i \mid X_{<i}\mathbf{Y}) = \frac{1}{n}\cdot I(\Pi;\mathbf{X} \mid \mathbf{Y}), \end{aligned}$$

where the inequality follows from Lemma 1.1.14, since $I(Y_{<i};X_i|X_{<i}) = 0$ by construction, and the last transition is by the chain rule for mutual information. By symmetry of construction, an analogues argument shows that $I(\theta;y|x) \leq I(\Pi;\mathbf{Y} \mid \mathbf{X})/n$, and combining these facts gives

$$\mathsf{IC}_\mu(\theta) \leq \frac{1}{n}\left(I(\Pi;\mathbf{X} \mid \mathbf{Y}) + I(\Pi;\mathbf{Y} \mid \mathbf{X})\right) = \frac{I}{n}. \tag{1.3}$$

□

1.1.6 The importance of private randomness

A subtle but vital issue when dealing with information complexity, is understanding the role of private vs. public coins. In randomized communication complexity, one often ignores the usage of private coins in a protocol. This is justified by the fact that private coins can be always simulated by public coins (on a "separate" part of the shared public tape), and [Newman]. Moreover, in the distributional model, when inputs arrive from some prior distribution μ, the averaging principle asserts that the communication cost can

be minimized by some fixed choice of the public randomness (which is why we consider only deterministic protocols in this model).

When dealing with *information complexity*, the situation is somewhat the opposite: Public coins are essentially a redundant resource (as it can be easily show by the chain rule that $\mathsf{IC}_\mu(\pi) = \mathbb{E}_R[\mathsf{IC}_\mu(\pi_R)]$) while the usage of private coins is crucial for minimizing the information cost of the protocol, and fixing these coins is prohibitive even when the information is measured with respect to a specific distribution. Consider the simple example where in the protocol π, Alice sends Bob her 1-bit input X, XORed with some random bit Z. If Z is private, Alice's message clearly reveals 0 bits of information to Bob about X. However, for any fixing of Z, this message would reveal X! The general intuition is that a protocol with minimum information cost requires the parties to reveal information about their inputs "carefully", and the usage of private coins serves to "conceal" parts of their inputs. This is not a theorem (and in fact, quantifying the role of private coins in minimizing information cost would essentially resolve the interactive compression problem , see e.g., [25]), but many known examples (e.g., [31]) support this intuition.

Thus, for the remainder of this article, communication protocols π are assumed to use private coins (and thus in particular, they are randomized even conditioned on the inputs x, y), while $\mathsf{IC}_\mu(\pi) = I(\Pi; X|YR) + I(\Pi; Y|XR)$ is the information conditioned on the *public* randomness R, but never on the private coins of π.

Part I

Applications to Communication Complexity

Chapter 2

Information Lower Bounds: New Techniques and Applications

The "Information=Amortized Communication" theorem (Theorem 1.0.1) asserts that proving lower bounds on the information complexity of f is equivalent to proving a lower bound on the amortized communication complexity of f. In particular, if f satisfies $IC(f) = \Omega(CC(f))$, i.e. that its information cost is asymptotically equal to its communication complexity, then a strong direct sum theorem holds for f. In addition to the intrinsic interest of understanding the amount of information exchange that needs to be involved in computing f, direct sum and product theorems (see Chapter 3) motivate the development of techniques for proving lower bounds on the information complexity of functions.

Another important motivation for proving information lower bounds stems from privacy aspects and understanding the limits of security in two-party computation (more on this in Chapter 7). In a celebrated result Ben-Or et al. [21] (see also [10]) showed how a multi-party computation (with three or more parties) may be carried out in a way that reveals no information to the participants except for the computation's output. The protocol relies heavily on the use of random bits that are shared between some, but not all,

parties. Such a resource can clearly not exist in the two-party setting. While it can be shown that a perfect information security is unattainable by two-party protocols [53, 12], quantitatively it is not clear just how much information must the parties "leak" to each other to compute f. The quantitative answer depends on the model in which the leakage occurs, and whether quantum computation is allowed [107]. Information complexity answers this question in the strongest possible sense for classical protocols: the parties are allowed to use private randomness to help them "hide" their information, and the information revealed is measured on average. Thus an information complexity lower bound of I on a problem implies that the *average* (as opposed to worst-case) amount of information revealed to the parties is at least I.

As mentioned above, the information complexity is always upper bounded by the communication complexity of f. The converse is not known to be true. Moreover, lower bound techniques for communication complexity do not readily translate into lower bound techniques for information complexity. The key difference is that a low-information protocol is not limited in the amount of communication it uses (an extreme example of this feature follows in Section 2.3), and thus rectangle-based communication bounds do not readily convert into information lower bound. No general technique has been known to yield sharp information complexity lower bounds. A linear lower bound on the communication complexity of the disjointness function has been shown in [137]. An information-theoretic proof of this lower bound [14] can be adapted to prove a linear *information* lower bound on disjointness [27]. One general technique for obtaining (weak) information complexity lower bounds was introduced in [27], where it has been shown that any function that has I bits of information complexity, has communication complexity bounded by $2^{O(I)}$. This immediately implies that the information complexity of a function f is at least the log of its communication complexity ($IC(f) \geq \Omega(\log(CC(f)))$). In fact, this result easily follows from the stronger result we prove below (Theorem 2.1.6).

In this section, we present general and specific techniques for proving information lower bounds on functions, and use then to prove strong communication lower bounds on several important functions: Gap-Hamming, Inner-Product, Greater-Than, Set-Disjointness and Set Intersection.

2.1 A Discrepancy Lower Bound for Information Complexity

Our first result is a general technique for proving information lower bounds on two-party unbounded-rounds communication problems. We show that the discrepancy lower bound, which applies to randomized communication complexity, also applies to information complexity. More precisely, if the discrepancy of a two-party function f with respect to a distribution μ is $Disc_\mu f$, then any two party randomized protocol computing f must reveal at least $\Omega(\log(1/Disc_\mu f))$ bits of information to the participants.

The proof we shall see establishes a general relationship between "weak" interactive compression results and information lower bounds, which has played a central role in the recent breakthrough work of Kerenidis et al. [102], who showed that almost all known lower bound techniques for communication complexity (and not just discrepancy) apply to information complexity.

By proving that the discrepancy of the Greater-Than function is $\Omega(1/\sqrt{n})$, we will use the above tool as a corollary to reprove Viola's [148] $\Omega(\log n)$ lower bound on the communication (and information) complexity of this well-studied function.

The Formal Result

The discrepancy of f with respect to a distribution μ on inputs, denoted $Disc_\mu(f)$, measures how "unbalanced" can f get on any rectangle, where the balancedness is measured

with respect to μ:

$$Disc_\mu(f) \stackrel{\Delta}{=} \max_{\text{rectangles } R} \left| \Pr_\mu[f(x,y) = 0 \wedge (x,y) \in R] - \Pr_\mu[f(x,y) = 1 \wedge (x,y) \in R] \right|.$$

A well-known lower bound (see e.g [112]) asserts that the distributional communication complexity of f, denoted $D^\mu_{1/2-\varepsilon}(f)$, when required to predict f with advantage ε over a random guess (with respect to μ), is bounded from below by $\Omega(\log 1/Disc_\mu(f))$:

$$D^\mu_{1/2-\varepsilon}(f) \geq \log(2\varepsilon/Disc_\mu(f)).$$

Note that the lower bound holds even if we are merely trying to get an advantage of $\varepsilon = \sqrt{Disc_\mu(f)}$ over random guessing in computing f. We prove that the information complexity of computing f with probability $9/10$ with respect to μ is also bounded from below by $\Omega(\log(1/Disc_\mu(f)))$.

Theorem 2.1.1. *[[39]] Let $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ be a Boolean function and let μ be any probability distribution on $\mathcal{X} \times \mathcal{Y}$. Then*

$$\mathsf{IC}_\mu(f, 1/10) \geq \Omega(\log(1/Disc_\mu(f))).$$

Remark 2.1.2. *The choice of $9/10$ is somewhat arbitrary. For randomized worst-case protocols, we may replace the success probability with $1/2 + \delta$ for a constant δ, since repeating the protocol constantly many times would yield the aforementioned success rate, while the information cost of the repeated protocol differs only by a constant factor from the original one. In particular, using prior-free information cost [28] this implies $\mathsf{IC} f 1/2 - \delta \geq \Omega_\delta(\log(1/Disc_\mu(f)))$.*

In particular, Theorem 2.1.1 implies a linear lower bound on the information complexity of the inner product function $IP(x,y) = \sum_{i=1}^n x_i y_i \mod 2$, and on a random boolean function $f_r : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$, expanding the (limited) list of functions for which nontrivial information-complexity lower bounds are known:

Corollary 2.1.3. *The information complexity* $\mathsf{IC}_{uniform}(IP, 1/10)$ *of* $IP(x, y)$ *is* $\Omega(n)$*. The information complexity* $\mathsf{IC}_{uniform}(f_r, 1/10)$ *of a random function* f_r *is* $\Omega(n)$*, except with probability* $2^{-\Omega(n)}$*.*

We study the communication and information complexity of the Greater-Than function (GT_n) on numbers of length n. This is a very well-studied problem [145, 124, 112]. Only very recently the tight lower bound of $\Omega(\log n)$ in the public-coins probabilistic model was given by Viola [148]. We show that the discrepancy of the GT_n function is $\Omega(1/\sqrt{n})$:

Lemma 2.1.4. *There exist a distribution* μ_n *on* $\mathcal{X} \times \mathcal{Y}$ *such that the discrepancy of* GT_n *with respect to* μ_n *satisfies*

$$Disc_{\mu_n}(GT_n) < \frac{20}{\sqrt{n}}.$$

Due to space constrains, we omit the proof of this lemma and refer the reader to the full version of the paper [39]. Lemma 2.1.4 provides an alternative (simpler) proof of Viola's [148] lower bound on the *communication complexity* of GT_n. By Theorem 2.1.1, Lemma 2.1.4 immediately implies a lower bound on the *information complexity* of GT_n:

Corollary 2.1.5. $\mathsf{IC}_{\mu_n}(GT_n, 1/10) = \Omega(\log n)$

This settles the information complexity of the GT function, since this problem can be solved by a randomized protocol with $O(\log n)$ communication (see [112]). This lower bound is particularly interesting since it demonstrates the first tight information-complexity lower bound that is not linear.

The key technical idea in the proof of Theorem 2.1.1 is a new simulation procedure that allows us to convert any protocol that has information cost I into a (two-round) protocol that has communication complexity $O(I)$ and succeeds with probability $> 1/2 + 2^{-O(I)}$, yielding a $2^{-O(I)}$ advantage over random guessing. Combined with the discrepancy lower bound for communication complexity, this proves Theorem 2.1.1.

Comparison and connections to prior results

The most relevant prior work is an article by Lee, Shraibman, and Špalek [114]. Improving on an earlier work of Shaltiel [140], Lee et al. show a direct product theorem for discrepancy, proving that the discrepancy of $f^{\otimes k}$ — the k-wise XOR of a function f with itself — behaves as $Disc(f)^{\Omega(k)}$. This implies in particular that the communication complexity of $f^{\otimes k}$ scales at least as $\Omega(k \cdot \log Disc(f))$. Using the fact that the limit as $k \to \infty$ of the amortized communication complexity of f is equal to the information cost of f [34], the result of Lee et al. "almost" implies the bound of Theorem 2.1.1. Unfortunately, the amortized communication complexity in the sense of [34] is the amortized cost of k copies of f, where *each* copy is allowed to err with some probability (say $1/10$). Generally speaking, this task is much easier than computing the XOR (which requires *all* copies to be evaluated correctly with high probability. Thus the lower bound that follows from Lee et al. applies to a more difficult problem, and does not imply the information complexity lower bound.

Another generic approach one may try to take is to use compression results such as [17] to lower bound the information cost from communication complexity lower bounds. The logic of such a proof would go as follows: "Suppose there was a information-complexity-I protocol π for f, then if one can compress it into a low-communication protocol one may get a contradiction to the communication complexity lower bound f". Unfortunately, all known compression results compress π into a protocol π' whose communication complexity depends on I but also on $CC(\pi)$. Even for external information complexity (which is always greater than the internal information compelxity, the bound obtained in [17] is of the form $I_{ext}(\pi) \cdot polylog(CC(\pi))$. Thus compression results of this type cannot rule out protocols that have low information complexity but a very high (e.g. exponential) communication complexity.

Our result can be viewed as a weak compression result for protocols, where a protocol for computing f that conveys I bits of information is converted into a protocol that

uses $O(I)$ bits of *communication* and giving an advantage of $2^{-O(I)}$ in computing f. This strengthens the result in [27] where a compression to $2^{O(I)}$ bits of communication has been shown. We still do not know whether compression to a protocol that uses $O(I)$ bits of communication and succeeds with high probability (as opposed to getting a small advantage over random) is possible.

Proof of Theorem 2.1.1

To establish the correctness of Theorem 2.1.1, we prove the following "weak simulation" theorem, which is the central result of [39]:

Theorem 2.1.6. *Suppose that* $\mathsf{IC}_\mu(f, 1/10) = I_\mu$. *Then there exist a protocol* π' *such that*

- $\mathsf{CC}(\pi') = O(I_\mu)$.
- $\Pr_{(x,y)\sim\mu}[\pi'(x,y) = f(x,y)] \geq 1/2 + 2^{-O(I_\mu)}$

We first show how Theorem 2.1.1 follows from Theorem 2.1.6:

Proof of Theorem 2.1.1. Let f, μ be as in theorem 2.1.1, and let $\mathsf{IC}_\mu(f, 1/10) = I_\mu$. By theorem 2.1.6, there exists a protocol π' computing f with error probability $1/2 - 2^{-O(I_\mu)}$ using $O(I_\mu)$ bits of communication. Applying the discrepancy lower bound for communication complexity we obtain

$$O(I_\mu) \geq D^\mu_{1/2-2^{-O(I_\mu)}}(f) \geq \log(2 \cdot 2^{-O(I_\mu)}/Disc_\mu(f)) \tag{2.1}$$

which after rearranging gives $I_\mu \geq \Omega(\log(1/Disc_\mu(f)))$, as desired.

We now turn to prove Theorem 2.1.6. The main step is the following sampling lemma.

Lemma 2.1.7. *Let μ be any distribution over a universe $\mathcal{U}$ and let $I \geq 0$ be a parameter that is known to both A and B. Further, let ν_A and ν_B be two distributions over $\mathcal{U}$ such that $\mathbb{D}(\mu\|\nu_A) \leq I$ and $\mathbb{D}(\mu\|\nu_B) \leq I$. The players are each given a pair of real functions (p_A, q_A), (p_B, q_B), $p_A, q_A, p_B, q_B : \mathcal{U} \to [0,1]$ such that for all $x \in \mathcal{U}$, $\mu(x) = p_A(x) \cdot p_B(x)$, $\nu_A(x) = p_A(x) \cdot q_A(x)$, and $\nu_B(x) = p_B(x) \cdot q_B(x)$. Then there is a (two round) sampling protocol $\Pi_1 = \Pi_1(p_A, p_B, q_A, q_B, I)$ which has the following properties:*

1. *at the end of the protocol, the players either declare that the protocol "fails", or output $x_A \in \mathcal{U}$ and $x_B \in \mathcal{U}$, respectively ("success").*

2. *let* suc *be the event that the players output "success". Then* suc $\Rightarrow x_A = x_B$, *and* $0.9 \cdot 2^{-50(I+1)} \leq \Pr[\mathsf{suc}] \leq 2^{-50(I+1)}$.

3. *if μ_1 is the distribution of x_A conditioned on* suc, *then $|\mu - \mu_1| < 2/9$.*

Furthermore, Π_1 can be "compressed" to a protocol Π_2 such that $\mathsf{CC}(\Pi_2) = 211I+1$, whereas $|\Pi_1 - \Pi_2| \leq 2^{-59I}$ (by an abuse of notation, here we identify Π_i with the random variable representing its output).

We will use the following technical fact about the information divergence of distributions.

Claim 2.1.8 (3). *[Claim 5.1 in [27]] Suppose that $\mathbb{D}(\mu\|\nu) \leq I$. Let ε be any parameter. Then*

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

For completeness, we repeat the short proof in the appendix.

Proof of Lemma 2.1.7 . Throughout the execution of Π_1, Alice and Bob interpret their shared random tape as a source of points (x_i, α_i, β_i) uniformly distributed in $\mathcal{U} \times [0, 2^{50(I+1)}] \times [0, 2^{50(I+1)}]$. Alice and Bob consider the first $T = |\mathcal{U}| \cdot 2^{100(I+1)} \cdot 60I$

such points. Their goal will be to discover the first index τ such that $\alpha_\tau \leq p_A(x_\tau)$ and $\beta_\tau \leq p_B(x_\tau)$ (where they wish to find it using a minimal amount of communication, even if they are most likely to fail). First, we note that the probability that an index t satisfies $\alpha_t \leq p_A(x_t)$ and $\beta_t \leq p_B(x_t)$ is exactly $1/|\mathcal{U}|2^{50(I+1)}2^{50(I+1)} = 1/|\mathcal{U}|2^{100(I+1)}$. Hence the probability that $\tau > T$ (i.e. that x_τ is not among the T points considered) is bounded by

$$\left(1 - 1/|\mathcal{U}|2^{100(I+1)}\right)^T < e^{-T/|\mathcal{U}|2^{100(I+1)}} = e^{-60I} < 2^{-60I} \tag{2.2}$$

Denote by $\mathcal{A}$ the following set of indices $\mathcal{A} := \{i \leq T : \ \alpha_i \leq p_A(x_i)$ and $\beta_i \leq 2^{50(I+1)} \cdot q_A(x_i)\}$, the set of potential candidates for τ from A's viewpoint. Similarly, denote $\mathcal{B} := \{i \leq T : \ \alpha_i \leq 2^{50(I+1)} \cdot q_B(x_i)$ and $\beta_i \leq p_B(x_i)\}$.

The protocol Π_1 is very simple. Alice takes her bet on the first element $a \in \mathcal{A}$ and sends it to Bob. Bob outputs a only if (it just so happens that) $\beta_\tau \leq p_B(a)$. The details are given in Figure 1.1 in the appendix.

We turn to analyze Π_1. Denote the set of "Good" elements by

$$\mathcal{G} \stackrel{\Delta}{=} \{x : \ 2^{50(I+1)} \cdot \nu_A(x) \geq \mu(x) \ \text{ and } \ 2^{50(I+1)} \cdot \nu_B(x) \geq \mu(x)\}\}.$$

Then by Claim 2.1.8, $\mu(\mathcal{G}) \geq 48/50 = 24/25$. The following claim asserts that if it succeeds, the output of Π_1 has the "correct" distribution on elements in $\mathcal{G}$.

Claim 2.1.9. *Assume $\mathcal{A}$ is nonempty. Then for any $x_i \in \mathcal{U}$, the probability that Π_1 outputs x_i is at most $\mu(x_i) \cdot 2^{-50(I+1)}$. If $x_i \in \mathcal{G}$, then this probability is* exactly $\mu(x_i) \cdot 2^{-50(I+1)}$.

Proof. Note that if $\mathcal{A}$ is nonempty, then for any $x_i \in \mathcal{U}$, the probability that x_i is the first element in $\mathcal{A}$ (i.e, $a = x_i$) is $p_A(x_i)q_A(x_i) = \nu_A(x_i)$. By construction, the probability that

$\beta_i \leq p_B(a)$ is $\min\{p_B(x_i)/(2^{50(I+1)}q_A(x_i)), 1\}$, and thus

$$\Pr[\Pi_1 \text{ outputs } x_i] \leq p_A(x_i)q_A(x_i) \cdot \frac{p_B(x_i)}{2^{50(I+1)}q_A(x_i)} = \mu(x_i) \cdot 2^{-50(I+1)}.$$

On the other hand, if $x_i \in \mathcal{G}$, then we know that $p_B(x_i)/q_A(x_i) = \mu(x_i)/\nu_A(x_i) \leq 2^{50(I+1)}$, and so the probability that $\beta_i \leq p_B(a)$ is *exactly* $p_B(x_i)/(2^{50(I+1)}q_A(x_i))$. Since $\Pr[\Pi_1 \text{ outputs } x_i] = \Pr[a = x_i]\Pr[\beta_i \leq p_B(a)]$ (assuming $\mathcal{A}$ is nonempty), we conclude that:

$$x_i \in \mathcal{G} \Longrightarrow \Pr[\Pi_1 \text{ outputs } x_i] = p_A(x_i)q_A(x_i) \cdot \frac{p_B(x_i)}{2^{50(I+1)}q_A(x_i)} = \mu(x_i) \cdot 2^{-50(I+1)}.$$

□

We are now ready to estimate the success probability of the protocol.

Proposition 2.1.10. *Let* suc *denote the event that* $\mathcal{A} \neq 0$ *and* $a \in \mathcal{B}$ *(i.e, that the protocol succeeds). Then*

$$0.9 \cdot 2^{-50(I+1)} \leq \Pr[\mathsf{suc}] \leq 2^{-50(I+1)}.$$

Proof. Using Claim 2.1.9, we have that

$$\begin{aligned} \Pr[\mathsf{suc}] &\leq \Pr[a \in \mathcal{B} \mid \mathcal{A} \neq \emptyset] = \sum_{i \in \mathcal{U}} \Pr[a = x_i]\Pr[\beta_i \leq p_B(a)] \leq \\ &\leq \sum_{i \in \mathcal{U}} \mu(x_i) \cdot 2^{-50(I+1)} = 2^{-50(I+1)} \end{aligned} \tag{2.3}$$

For the lower bound, we have

$$
\begin{aligned}
\Pr[\mathsf{suc}] &\geq \Pr[\beta_i \leq p_B(a) \mid \mathcal{A} \neq \emptyset] \cdot \Pr[\mathcal{A} \neq \emptyset] \geq \\
&\geq (1 - 2^{-60I})\left(\sum_{i \in \mathcal{U}} \Pr[a = x_i] \Pr[\beta_i \leq p_B(a)] \right) \geq \\
&\geq (1 - 2^{-60I})\left(\sum_{i \in \mathcal{G}} \Pr[a = x_i] \Pr[\beta_i \leq p_B(a)] \right) = \\
&= (1 - 2^{-60I})\left(2^{-50(I+1)} \sum_{i \in \mathcal{G}} \mu(x_i) \right) = (1 - 2^{-60I})\left(2^{-50(I+1)} \mu(\mathcal{G}) \right) \geq \\
&\geq \frac{24}{25}(1 - 2^{-60I}) 2^{-50(I+1)} \geq 0.9 \cdot 2^{-50(I+1)} \qquad (2.4)
\end{aligned}
$$

where the equality follows again from claim 2.1.9. This proves the second claim of Lemma 2.1.7. □

The following claim asserts that if suc occurs, then the distribution of a is indeed close to μ.

Claim 2.1.11 (4). *Let μ_1 be the distribution of a|suc. Then $|\mu_1 - \mu| \leq 2/9$.*

Proof. The claim follows directly from proposition 2.1.10. We defer the proof to the appendix.

We turn to the "Furthermore" part of of Lemma 2.1.7. The protocol Π_1 satisfies the premises of the lemma, except it has a high communication cost. This is due to the fact that Alice explicitly sends a to Bob. To reduce the communication, Alice will instead send $O(I)$ random hash values of a, and Bob will add corresponding consistency constraints to his set of candidates. The final protocol Π_2 is given in Figure 2.1.

Let $\mathcal{E}$ denote the event that in step 4 of the protocol, Bob finds an element $x_i \neq a$ (that is, the probability that the protocol outputs "success" but $x_A \neq x_B$). We upper bound the probability of $\mathcal{E}$. Given $a \in \mathcal{A}$ and $x_i \in \mathcal{B}$ such that $a \neq x_i$, the probability (over possible choices of the hash functions) that $h_j(a) = h_j(x_i)$ for $j = 1..d$ is $2^{-d} \leq 2^{-211I}$. For any t,

Information-cost sampling protocol Π_2

1. Alice computes the set $\mathcal{A}$. Bob computes the set $\mathcal{B}$.
2. If $\mathcal{A} = \emptyset$, the protocol fails. Otherwise, Alice finds the first element $a \in \mathcal{A}$ and sets $x_A = a$. She then computes $d = \lceil 211I \rceil$ random hash values $h_1(a), \ldots, h_d(a)$, where the hash functions are evaluated using public randomness.
3. Alice sends the values $\{h_j(a)\}_{1 \le j \le d}$ to Bob.
4. Bob finds the first index τ such that there is a $b \in \mathcal{B}$ for which $h_j(b) = h_j(a)$ for $j = 1..d$ (if such an τ exists). Bob outputs $x_B = x_\tau$. If there is no such index, the protocol fails.
5. Bob outputs x_B ("success").
6. Alice outputs x_A.

Figure 2.1: The sampling protocol Π_2 from Lemma 2.1.7

$\Pr[t \in \mathcal{B}] \le \frac{1}{|\mathcal{U}|} \sum_{x_i \in \mathcal{U}} p_B(x_i) q_B(x_i) \cdot 2^{50(I+1)} = \frac{1}{|\mathcal{U}|} \sum_{x_i \in \mathcal{U}} \nu_B(x_i) \cdot 2^{50(I+1)} = 2^{50(I+1)}/|\mathcal{U}|$. Thus, by a union bound we have

$$\begin{aligned} \Pr[\mathcal{E}] &\le \Pr[\exists x_i \in \mathcal{B}\ s.t\ x_i \neq a \ \wedge\ h_j(a) = h_j(x_i)\ \forall\ j = 1, \ldots, d] \le \\ &\le T \cdot 2^{50(I+1)} \cdot 2^{-d}/|\mathcal{U}| = 2^{150(I+1)} \cdot 60I \cdot 2^{-211I} \ll 2^{-60I}. \end{aligned} \tag{2.5}$$

By a slight abuse of notation, let Π_2 be the distribution of Π_2's output. Similarly, denote by Π_1 the distribution of the output of protocol Π_1. Note that if $\mathcal{E}$ does not occur, then the outcome of the execution of Π_2 is identical to the outcome of Π_1. Since $\Pr[\mathcal{E}] \le 2^{-60I}$, we have

$$|\Pi_2 - \Pi_1| = \Pr[\mathcal{E}] \cdot |[\Pi_2|\mathcal{E}] - [\Pi_1|\mathcal{E}]| \le 2 \cdot 2^{-60I} \ll 2^{-59I}$$

which finishes the proof of the lemma.

□

Using the above lemma, we are now ready to prove our main theorem.

Proof of Theorem 2.1.6 . Let π be a protocol that realizes the value $I_\mu := \mathsf{IC}_\mu(f, 1/10)$. In other words, π has an error rate of at most $1/10$ and information cost of at most I_μ with respect to μ. Denote by π_{xy} the random variable that represents that transcript π given the inputs (x, y), and by π_x (resp. π_y) the protocol conditioned on only the input x (resp. y). We denote by π_{XY} the transcripts where (X, Y) are also a pair of random variables. By Claim 2.1.8, we know that

$$I_\mu = I(X; \pi_{XY}|Y) + I(Y; \pi_{XY}|X) = \mathbb{E}_{(x,y)\sim\mu}[\mathbb{D}\left(\pi_{xy}\|\pi_x\right) + \mathbb{D}\left(\pi_{xy}\|\pi_y\right)]. \tag{2.6}$$

Let us now run the sampling algorithm Π_1 from Lemma 2.1.7, with the distribution μ taken to be π_{xy}, the distributions ν_A and ν_B taken to be π_x and π_y respectively, and I taken to be $20I_\mu$.

At each node v of the protocol tree that is owned by player X let $p_0(v)$ and $p_1(v) = 1 - p_0(v)$ denote the probabilities that the next bit sent by X is 0 and 1, respectively. For nodes owned by player Y, let $q_0(v)$ and $q_1(v) = 1 - q_0(v)$ denote the probabilities that the next bit sent by Y is 0 and 1, respectively, *as estimated by player X given the input x*. For each leaf ℓ let $p_X(\ell)$ be the product of all the values of $p_b(v)$ from the nodes that are owned by X along the path from the root to ℓ; let $q_X(\ell)$ be the product of all the values of $q_b(v)$ from the nodes that are owned by Y along the path from the root to ℓ. The values $p_Y(\ell)$ and $q_Y(\ell)$ are defined similarly. For each ℓ we have $\Pr[\pi_{xy} = \ell] = p_X(\ell) \cdot p_Y(\ell)$, $\Pr[\pi_x = \ell] = p_X(\ell) \cdot q_X(\ell)$, and $\Pr[\pi_y = \ell] = p_Y(\ell) \cdot q_Y(\ell)$. Thus we can apply Lemma 2.1.7 so as to obtain the following protocol π' for computing f:

- If Π_1 fails, we return a random unbiased coin flip.
- If Π_1 succeeds, we return the final bit of the transcript sample T. Denote this bit by T_{out}.

To prove the correctness of the protocol, let $\mathcal{Z}$ denote the event that both $\mathbb{D}\left(\pi_{xy}\|\pi_x\right) \leq 20I_\mu$ and $\mathbb{D}\left(\pi_{xy}\|\pi_y\right) \leq 20I_\mu$. By (2.6) and Markov inequality, $\Pr[\mathcal{Z}] \geq 19/20$ (where the probability is taken with respect to μ). Denote by δ the probability that Π_1 succeeds. By the assertions of Lemma 2.1.7, $\delta \geq 0.9 \cdot 2^{-50(I+1)}$. Furthermore, if Π_1 succeeds, then we have $|T - \pi_{xy}| \leq 2/9$, which in particular implies that $\Pr[T_{out} = \pi_{out}] \geq 7/9$. Finally, $\Pr[\pi_{out} = f(x,y)] \geq 9/10$, since π has error at most $1/10$ with respect to μ. Now, let $\mathcal{W}$ denote the indicator variable whose value is 1 iff $\pi'(x,y) = f(x,y)$. Putting together the above,

$$\mathbb{E}[\mathcal{W} \mid \mathcal{Z}] = (1-\delta)\cdot\frac{1}{2} + \delta\cdot\left(\frac{7}{9} - \frac{1}{10}\right) > \frac{1}{2} + \delta\cdot\frac{1}{6} > \frac{1}{2} + \frac{1}{8}\cdot 2^{-50(I+1)}. \tag{2.7}$$

On the other hand, note that by lemma 2.1.7 the probability that Π_1 succeeds is at most $2^{-50(I+1)}$ (no matter how large $\mathbb{D}\left(\pi_{xy}\|\pi_x\right)$ and $\mathbb{D}\left(\pi_{xy}\|\pi_y\right)$ are!), and so $\mathbb{E}[\mathcal{W} \mid \neg\mathcal{Z}] \geq (1 - 2^{-50(I+1)})/2$.

Hence we conclude that

$$\begin{aligned}\mathbb{E}[\mathcal{W}] &= \mathbb{E}[\mathcal{W} \mid \mathcal{Z}]\cdot\Pr[\mathcal{Z}] + \mathbb{E}[\mathcal{W} \mid \neg\mathcal{Z}]\cdot\Pr[\neg\mathcal{Z}] \geq \left(\frac{1}{2} + \frac{1}{8}\cdot 2^{-50(I+1)}\right)\cdot\frac{19}{20} + \\ &+ \left(1 - 2^{-50(I+1)}\right)\cdot\frac{1}{2}\cdot\frac{1}{20} \geq \frac{1}{2} + \frac{1}{12}\cdot 2^{-50(I+1)} > \frac{1}{2} + \frac{1}{12}\cdot 2^{-1000(I_\mu+1)}.\end{aligned}$$

Finally, Lemma 2.1.7 asserts that $|\Pi_1 - \Pi_2| < 2^{-59I}$. Thus if we replace Π_1 by Π_2 in the execution of protocol π', the success probability decreases by at most $2^{-59I} \ll \frac{1}{12}\cdot 2^{-50(I+1)}$. Furthermore, the amount of communication used by π' is now

$$211I = 4220I_\mu = O(I_\mu).$$

Hence we conclude that with this modification, π' has the following properties:

- $\mathsf{CC}(\pi') = 4220 \cdot I_\mu$;

- $\Pr(x,y) \sim \mu[\pi'(x,y) = f(x,y)] \geq 1/2 + 2^{-1000(I_\mu+1)-4}$;

which completes the proof. □

Remark 2.1.12. *Using similar techniques, it was recently shown in [27] that any function f whose information complexity is I has communication cost at most $2^{O(I)}$ [1], thus implying that $IC(f) \geq \Omega(\log(CC(f)))$. We note that this result can be easily derived (up to constant factors) from Theorem 2.1.6. Indeed, applying the "compressed" protocol $2^{O(I)}\log(1/\varepsilon)$ independent times and taking a majority vote guarantees an error of at most ε (by a standard Chernoff bound[2]), with communication $O(I) \cdot 2^{O(I)} = 2^{O(I)}$. Thus, our result is strictly stronger than the former one.*

2.2 Information Lower Bounds via Self Reducibility

In this short section we illustrate a more specific technique for deriving strong information complexity lower bounds for "self-reducible" functions. This approach is somewhat opposite to the standard one, in that the information lower bounds are obtained from postulated (known) *communication complexity* lower bounds. We note that such information lower bounds are valuable even if a communication bound is already known, since they imply strong direct sum (and product) theorems for such functions (See Chapter 3), and also an inherent limit on the privacy required to solve such functions, as noted in the introduction of this chapter. Due to space constraints, we only outline our main results and techniques, and refer the reader to the full version of this paper [32] for the complete proofs.

The technique we present works for functions that exhibit a "self-reducible structure". Informally speaking f has a self-reducible structure, if for large enough inputs, solving f_{nk} essentially amounts to solving f_n^k (f_{nk} denotes the function f under inputs of length nk, while f_n^k denotes k independent copies of f under inputs of size n). Our departing point is a communication complexity lower bound for f_{nk} (that may be obtained by

[1] More precisely, it shows that for any distribution μ, $D^{\mu}_{\varepsilon+\delta}(f) = 2^{O(1+\mathsf{IC}_\mu(f,\varepsilon)/\delta^2)}$.

[2] See N.Alon and J. Spencer, "The Probabilistic Method" (Third Edition) ,Corollary A.1.14, p.312.

any means). Assuming self-reducibility, the same bound applies to f_n^k, which through the "Information= Amortized Communication" theorem [35], implies a lower bound on the information complexity of f_n. In the following work we develop tools to make this reasoning go through.

We use the self-reducibility technique to prove results about the information complexity of Gap Hamming Distance and Inner Product. We prove that the information complexity of the Gap Hamming Distance problem with respect to the uniform distribution is linear. This was explicitly stated as an open problem by Chakrabarti et al. [50]. Formally, let $\mathsf{IC}_\mu(GHD_{n,t,g}, \varepsilon)$ denote the information cost of the Gap Hamming promise problem, where inputs x, y are n-bit strings distributed according to μ, and the players need to determine whether the Hamming distance between x and y is at least $t+g$, or at most $t-g$, with error at most ε under μ. We prove

Theorem 2.2.1 ([32]). *There exists an absolute constant $\varepsilon > 0$ for which*

$$\mathsf{IC}_{\mathcal{U}}(GHD_{n,n/2,\sqrt{n}}, \varepsilon) = \Omega(n)$$

where $\mathcal{U}$ is the uniform distribution.

For the Inner Product problem, where the players need to compute $\sum_{i=1}^{n} x_i y_i (mod\ 2)$, we prove a stronger bound on its information complexity. Formally

Theorem 2.2.2 ([32]). *For every constant $\delta > 0$, there exists a constant $\epsilon > 0$, and n_0 such that $\forall\ n \geq n_0$, $IC_{\mathcal{U}_n}(IP_n, \epsilon) \geq (1-\delta)n$. Here $\mathcal{U}_n$ is the uniform distribution over $\{0,1\}^n \times \{0,1\}^n$.*

Note that $IC_{\mathcal{U}_n}(IP_n, \epsilon) \leq (1-2\epsilon)(n+1)$, since the parties can always output a random value $\in \{0,1\}$ with probability 2ϵ, and have one of the parties send its entire input with probability $1-2\epsilon$ (Indeed, this protocol has error $(1/2)\cdot 2\epsilon + (1-2\epsilon)\cdot 1 = 1-\epsilon$, and information cost $(1/2)\cdot 0 + (1-2\epsilon)\cdot(n+1) = (1-2\epsilon)\cdot(n+1)$). Also it is known that $IC_{\mathcal{U}_n}(IP_n, \epsilon) \geq \Omega(n)$, for all $\epsilon \in [0, 1/2)$ [38]. We prove that the information complexity of

IP_n can be arbitrarily close to the trivial upper bound n as we keep decreasing the error (though keeping it a constant).

We refer the reader to [32] for a broader discussion and the full proofs of Theorems 2.2.1 and 2.2.2.

2.3 From Information to Exact Communication

Traditional communication complexity lower bound techniques were combinatorial in nature and most of them rely on studying the combinatorial and analytic properties of the communication matrix M_f corresponding to the function f (notable examples are the rank, Discrepancy, corruption and "smooth rectangle" lower bounds). While most existing state-of-the-art communication lower bounds were proved this way (including recently ones such as the lower bound for Gap Hamming Distance [48, 143]), such techniques often lose constants by design, and are too crude to give exact communication bounds.

In contrast, information theory is known to give *precise* bounds on rates and capacities. This important feature essentially stems from the additivity property of information complexity (see equation 1.1), which can be viewed as a generalization of Shannon's noiseless coding theorem to the interactive setup. For example, we know that a sequence of random digits would take exactly $\log_2 10 \approx 3.322$ bits per digit, and that the capacity of a binary symmetric channel with substitution probability 0.2 is exactly $1 - H(0.2) \approx 0.278$ bits per symbol. Generally speaking, prior to this work, this benefit has not been fully realized in an interactive communication complexity scenario.

In this section, we present a framework for proving *exact* (zero-error) information lower bounds, thereby bringing tight bounds into the realm of communication complexity: We develop a new local characterization of the zero-error information complexity function for two-party communication problems, and use it to compute the exact internal

and external information complexity of the $2-$bit AND function: $\mathsf{IC}(\text{AND}, 0) = C_{\wedge} \approx 1.4923$ bits, and $\mathsf{IC}^{\text{ext}}(\text{AND}, 0) = \log_2 3 \approx 1.5839$ bits. We shall see that this leads to a tight (upper and lower bound) characterization of the communication complexity of the *set intersection* problem (where players are required to compute the intersection of their sets), whose randomized communication complexity tends to $C_{\wedge} \cdot n \pm o(n)$ as the error tends to zero.

The information-optimal protocol we present has an *infinite* number of rounds. We shall see that this tradeoff is necessary by proving that the rate of convergence of the $r-$round information cost of AND to $\mathsf{IC}(\text{AND}, 0) = C_{\wedge}$ behaves like $\Theta(1/r^2)$, i.e. that the r-round information complexity of AND is $C_{\wedge} + \Theta(1/r^2)$.

We will then leverage the tight analysis obtained for the information complexity of AND to calculate and prove the exact communication complexity of the *set disjointness* function $Disj_n(X, Y) = \neg \vee_{i=1}^{n} AND(x_i, y_i)$ with error tending to 0, which turns out to be $= C_{DISJ} \cdot n \pm o(n)$, where $C_{DISJ} \approx 0.4827$. Our rate of convergence results imply that an asymptotically optimal protocol for set disjointness will have to use $\omega(1)$ rounds of communication, since every r-round protocol will be sub-optimal by at least $\Omega(n/r^2)$ bits of communication.

In a similar spirit, we obtain a tight bound of $\frac{2}{\ln 2}k \pm o(k)$ on the communication complexity of disjointness of sets of size $\leq k$, sharpening the asymptotic bound of $\Theta(k)$ previously shown by Håstad and Wigderson.

2.3.1 Main Results

Let π be a communication protocol attempting to solve some two-party function $f(x, y)$ with zero error where inputs are sampled according to a joint distribution μ. Our first contribution is a characterization of the zero-error information cost function $IC_{\mu}(f, 0)$ in terms of certain local concavity constrains. A related – but more abstract – characteriza-

tion was given in the information theory literature by Ma and Ishwar [120]. Let $\Delta(\mathcal{X} \times \mathcal{Y})$ denote the set of distributions over $\mathcal{X} \times \mathcal{Y}$.

Lemma 2.3.1. *For any function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ there exists a family $\mathfrak{C}(f)$ of functions $C : \Delta(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}^+$ satisfying certain local concavity constraints, such that for any distribution μ, and any protocol π solving f with zero error under μ, it holds that*

$$\forall \ C \in \mathfrak{C}(f) \quad C(\mu) \leq \mathsf{IC}_\mu(\pi).$$

Furthermore, $IC_\mu(f, 0)$ is the point-wise maximum of $\mathfrak{C}(f)$.

This lemma gives a very general technique for proving information-complexity lower bounds, and plays a central role in one of our main results: the exact information complexity of the 2-bit AND function $f(x, y) = x \wedge y$. Since the inputs of the parties consist of only 2 bits, the information complexity of this function is trivially bounded by 2. By fixing $x = 1$, it is also easy to see that 1 is a lower bound on the information complexity. We present a zero-error "clocked" protocol which has an infinite number of rounds and computes the AND function, under any input distribution μ, with information cost at most $C_\wedge \approx 1.4923$. The maximum external information cost of our protocol is $\log_2 3 \approx 1.58496$. While the analysis itself is nontrivial, the main bulk of effort is proving this protocol is in fact optimal, both in the internal and external sense:

Theorem 2.3.2.

$$\mathsf{IC}(AND, 0) = C_\wedge \approx 1.4923$$

Theorem 2.3.3.

$$\mathsf{IC}^{ext}(AND, 0) = \log_2 3 \approx 1.58496$$

We also analyze the rate of convergence to the optimal information cost, as the number r of permitted rounds increases. We view this result as a step towards proving that information complexity of functions is computable.

Theorem 2.3.4. *For all $\mu \in \Delta(\{0,1\} \times \{0,1\})$ with full support we have*

$$\mathsf{IC}^r_\mu(AND, 0) = \mathsf{IC}_\mu(AND, 0) + \Theta_\mu\left(\frac{1}{r^2}\right).$$

In the second part of our work we show how tight information bounds may lead to exact communication bounds.

We leverage our in-depth information analysis of AND to prove the *exact* randomized communication complexity of the $Disj_n$ function, with error tending to zero. For the general disjointness function we get:

Theorem 2.3.5. *For all $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that $\delta \to 0$ as $\varepsilon \to 0$ and*

$$(C_{DISJ} - \delta) \cdot n \leq \mathsf{R}(DISJ_n, \varepsilon) \leq C_{DISJ} \cdot n + o(n).$$

where $C_{DISJ} \approx 0.4827$ bits.

For the case of disjointness $DISJ^k_n$ of sets of size $\leq k$ we get

Theorem 2.3.6. *Let n, k be such that $k = \omega(1)$ and $n/k = \omega(1)$. Then for all constant $\varepsilon > 0$,*

$$\left(\frac{2}{\ln 2} - O(\sqrt{\varepsilon})\right) \cdot k - o(k) \leq \mathsf{R}(DISJ^k_n, \varepsilon) \leq \frac{2}{\ln 2} \cdot k + o(k).$$

We also observe that Theorem 2.3.2 leads to the exact (randomized) communication complexity of the Set Intersection problem, which turns out to be $C_\wedge \cdot n \approx 1.492 \cdot n$.

Our results rely on new insights for understanding communication protocols from an informational point of view, as functionals on the space of distributions. This requires further development of new properties of the information cost function. One such property is the continuity of the information complexity function at $\varepsilon = 0$:

Theorem 2.3.7. *For all $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ and $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$ we have*

$$\lim_{\epsilon \to 0} \mathsf{IC}_\mu(f, \epsilon) = \mathsf{IC}_\mu(f, 0), \tag{2.8}$$

$$\lim_{\epsilon \to 0} \mathsf{IC}^{ext}_{\mu}(f, \epsilon) = \mathsf{IC}^{ext}_{\mu}(f, 0). \tag{2.9}$$

Preliminaries

Notation

For random variables A and B_i ($i \in [n]$) and elements $b_i \in \operatorname{range} B_i$ ($i \in [n]$) we write $A_{b_1 b_2 \cdots b_n}$ to denote the random variable A conditioned on the event "$B_1 = b_1, B_2 = b_2, \ldots, B_n = b_n$".

For notational convenience, in this section we shall sometimes view a probability distribution μ on a sample space $\mathcal{X} \times \mathcal{Y}$ as a $|\mathcal{X}| \times |\mathcal{Y}|$ matrix, where the rows are indexed by elements of $\mathcal{X}$ and columns are indexed by elements of $\mathcal{Y}$ in some standard order (e. g., lexicographic order when $\mathcal{X}$ and $\mathcal{Y}$ are sets of binary strings). For example, we shall often write distribution μ on $\{0,1\} \times \{0,1\}$ as $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & \delta \\ \hline \end{array}$ meaning that $\mu(0,0) = \alpha, \mu(0,1) = \beta, \mu(0,1) = \gamma$, and $\mu(1,1) = \delta$.

For a particular distribution μ on $\mathcal{X} \times \mathcal{Y}$ we use μ^T to denote the probability distribution on $\mathcal{Y} \times \mathcal{X}$ that is given by the transpose of the matrix representation of μ.

In this section we will be interested in the information revealed by protocols under the"worst distribution". To capture this, we use the notion of *prior-free* information complexity (or simply, the *information cost* of f) with error ϵ is defined as

$$\mathsf{IC}(f, \epsilon) := \inf_{\pi} \max_{\mu \in \Delta(\mathcal{X} \times \mathcal{Y})} \mathsf{IC}_{\mu}(\pi).$$

where the infimum is over protocols that work correctly for each input, except with probability ε. The *external prior-free* information cost is defined analogously.

The special case $\mathsf{IC}(f, 0)$ is referred to as the *zero error* information complexity of f, and will be of primary interest in this paper. It turns out that for this special case ($\epsilon = 0$), we may reverse the order of quantifiers:

Theorem 2.3.8. *[29]*

$$IC(f,0) = \max_{\mu} \inf_{\pi \text{ correct on support of } \mu} \mathsf{IC}_{\mu}(\pi),$$

i.e, we can choose the protocol dependent on the distribution and yet the information cost doesn't decrease.

For $r \in \mathbb{N}$, the *r-round information complexity* of a function f is defined as

$$\mathsf{IC}^{r}_{\mu}(f,\epsilon) := \inf_{\pi} \mathsf{IC}_{\mu}(\pi),$$

where the infimum ranges over all r-round protocols π solving f with error at most ϵ when inputs are sampled according to μ. The *r-round external information cost* is defined analogously.

2.3.4 Optimal Information-Theoretic Protocol for AND

The information complexity of a function is the infimum over protocols of the information cost of the protocol. Therefore the information complexity may not be achieved by any single protocol. This is indeed the case for the AND function, as we will see in Section 2.3.7. Nevertheless if we allow slightly more powerful protocols we can find a *single optimal protocol* for the AND function. In this section we present a "protocol with a clock" (see Protocol 1) whose information cost is *exactly equal to* the information cost of the AND function. The inputs (X, Y) to AND are distributed according to a prior $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & \delta \\ \hline \end{array}$.

Protocol 1 consists of two parts. In the first part (steps 1 and 2), Alice and Bob check to see if their prior is symmetric, and if it is not they communicate "a bit" to make it symmetric. During this communication one of the players may reveal that his or her input is 0, in which case the protocol terminates, as the answer to AND can be deduced by both players. In the second part (steps $3-6$), Alice and Bob privately generate random numbers $N^A \in [0,1]$ and $N^B \in [0,1]$ and observe the clock as it increases from 0 to 1. When

1. If $\beta < \gamma$ then Bob sends bit B as follows

$$B = \begin{cases} 1 & \text{if } y = 1 \\ 0 & \text{with probability } 1 - \beta/\gamma \text{ if } y = 0 \\ 1 & \text{with probability } \beta/\gamma \text{ if } y = 0 \end{cases}$$

   If $B = 0$ the protocol terminates and players output 0.

2. If $\beta > \gamma$ then Alice sends bit B as follows

$$B = \begin{cases} 1 & \text{if } x = 1 \\ 0 & \text{with probability } 1 - \gamma/\beta \text{ if } x = 0 \\ 1 & \text{with probability } \gamma/\beta \text{ if } x = 0 \end{cases}$$

   If $B = 0$ the protocol terminates and players output 0.

3. If $x = 0$ then Alice samples $N^A \in_R [0, 1)$ uniformly at random. If $x = 1$ then Alice sets $N^A = 1$.

4. If $y = 0$ then Bob samples $N^B \in_R [0, 1)$ uniformly at random. If $y = 1$ then Bob sets $N^B = 1$.

5. Alice and Bob monitor the clock C, which starts at value 0.

6. The clock continuously increases to 1. If $\min(N^A, N^B) < 1$, when the clock reaches $\min(N^A, N^B)$ the corresponding player sends 0 to the other player, the protocol ends, the players output 0. If $\min(N^A, N^B) = 1$, once the clock reaches 1, Alice sends 1 to Bob, the protocol ends, and the players output 1.

Protocol 1: Protocol π for the AND-function

some player's private number is reached by the clock, the player immediately notifies the other player. The rules for picking a private number ensure that the number is less than 1 if and only if the owner of the number has 0 as input. Therefore once one of the players speaks in the second part, both players can deduce the answer to AND, so the protocol terminates.

From the description of Protocol 1, it is clear that it correctly solves AND on all inputs. The proof of the optimality of the information cost of this protocol proceeds in two steps. The first step is to analyze the information cost of Protocol 1. The result of this analysis is a precise and simple formula for $I(\mu) := \mathsf{IC}_\mu(\pi)$ in terms of $\alpha, \beta, \gamma, \delta$. In addition, we

conclude that $I(\mu) \geq \mathsf{IC}_\mu(\text{AND}, 0)$. For the second step, we need a new technique to prove *exact* information lower bounds. This technique relies on the new characterization of the information cost presented in Section 2.3.5. In that section we show that any function satisfying certain local concavity constraints is a lower bound on the information cost. To complete the proof that $I(\mu) = \mathsf{IC}_\mu(\text{AND}, 0)$ we simply check that $I(\mu)$ satisfies those local concavity constraints, and indeed it does.

We attempt to demystify the steps of this protocol by presenting the intuition behind optimality of its information cost. To this end, we may view any protocol as a random walk on the space of distributions on $\mathcal{X} \times \mathcal{Y}$. We observe that for the AND function the space of distributions μ on $\{0,1\}^2$ may be divided into three regions:

Alice's region consists of all distributions μ with $\beta > \gamma$, i. e., those distributions μ, for which Alice has greater probability of having 0 than Bob.

Bob's region consists of all distributions μ with $\beta < \gamma$, i. e., those distributions μ, for which Bob has greater probability of having 0 than Alice.

Diagonal region consists of symmetric distributions μ, i. e., $\beta = \gamma$ and both players are equally likely to have 0 as input.

We note that a protocol in which Alice talks in Bob's region and then the players play optimally, reveals more information about the inputs than a protocol in which Bob talks in Bob's region and then players play optimally (and Similarly for Alice's region). A formalization of this argument appears in the full version of this paper. Therefore in an optimal protocol, each player should speak *only in his own region*. The interesting scenario is when the protocol finds itself in the diagonal region. Suppose that players want to convince each other that they are more likely to have 1 as input. If Bob makes a random step, he will step into Alice's region with some probability revealing suboptimal amount of information. The same goes for Alice. What we'd like them to do is to walk "along the diagonal region". This can be accomplished without revealing suboptimal amount of

information only if we allow the players to take infinitesimal steps. This is precisely what the clock from our protocol achieves. As the clock increases from 0 to 1, the distribution stays symmetric, but gets modified *simultaneously* by increasing its mass on $(1,1)$-entry.

Remark 2.3.9. *It turns out that Protocol 1 achieves* both internal and external *information costs. The analysis reveals that the internal and external information costs are different for the AND function.*

We refer an interested reader to the full version of the paper for the details on how to make the above intuition precise, and for a careful analysis of the information cost of Protocol 1. In the rest of this section we present a summary of results (omitting the proofs) that we were able to achieve using the above techniques.

Observe that since AND is a symmetric function $\mathsf{IC}_{\mu}(\text{AND},0) = \mathsf{IC}_{\mu^T}(\text{AND},0)$, therefore it suffices to compute the information cost for the AND function only for distributions with $\beta \leq \gamma$.

Theorem 2.3.10 ([31])**.** *For a symmetric distribution* $\nu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \beta & \delta \\ \hline \end{array}$ *we have*

$$\begin{aligned}\mathsf{IC}_{\nu}(AND,0) &= \frac{\beta}{\ln 2} + 2\delta \log \frac{\beta+\delta}{\delta} + \\ &+ 2\beta \log \frac{\beta+\delta}{\beta} + \frac{\beta^2}{\alpha} \log \frac{\beta}{\beta+\alpha} + \alpha \log \frac{\alpha+\beta}{\alpha}. \end{aligned} \tag{2.10}$$

For a distribution $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & \delta \\ \hline \end{array}$*, where* $\beta < \gamma$*, we have*

$$IC_{\mu}(AND,0) = I(Y;B|X) + t\mathsf{IC}_{\widetilde{\nu}}(\pi)$$

where $t = \delta + 2\beta + \frac{\alpha\beta}{\gamma}$, $\widetilde{\nu} = \begin{array}{|c|c|}\hline \frac{\beta\alpha}{\gamma t} & \frac{\beta}{t} \\ \hline \frac{\beta}{t} & \frac{\delta}{t} \\ \hline \end{array}$ *and*

$$\begin{aligned} & I(Y;B|X) \\ & = (\alpha+\beta)H\left(\frac{\beta}{\gamma}\cdot\frac{\alpha+\gamma}{\alpha+\beta}\right) + (\gamma+\delta)H\left(\frac{\delta+\beta}{\gamma+\delta}\right) - \\ & - (\alpha+\gamma)H\left(\frac{\beta}{\gamma}\right). \end{aligned}$$

Theorem 2.3.11 ([31])**.** *(Theorem 2.3.2 restated)*

$$\mathsf{IC}(AND, 0) = C_{\wedge} = 1.49238\ldots$$

The distribution that achieves this maximum is

$$\mu = \begin{array}{|c|c|}\hline 0.0808931\ldots & 0.264381\ldots \\ \hline 0.264381\ldots & 0.390346\ldots \\ \hline \end{array}.$$

Remark 2.3.12. *Observe that the maximum of* $\mathsf{IC}(AND, 0)$ *is achieved for a symmetric distribution. This is not a coincidence. Let* f *be a symmetric function and* μ *be an arbitrary distribution on the inputs of* f*. Then* $\mathsf{IC}_{\mu}(f, 0) = \mathsf{IC}_{\mu^T}(f, 0)$ *and it is easy to see that the information complexity is a concave function in* μ*. Thus for* $\mu' = \mu/2 + \mu^T/2$*, which is symmetric, we have* $\mathsf{IC}_{\mu'}(f, 0) \geq \mathsf{IC}_{\mu}(f, 0)/2 + \mathsf{IC}_{\mu^T}(f, 0)/2 = \mathsf{IC}_{\mu}(f, 0)$*.*

Theorem 2.3.13 ([31])**.** *(Theorem 2.3.3 restated)*

$$\mathsf{IC}^{ext}(AND, 0) = \log 3 = 1.58396\ldots$$

The distribution that achieves this maximum is

$$\mu = \begin{array}{|c|c|}\hline 0 & 1/3 \\ \hline 1/3 & 1/3 \\ \hline \end{array}.$$

In Section 2.3.5 on communication complexity results, distributions μ that place 0 mass on $(1,1)$ entry play a crucial role. Note that for such distributions we still insist that the protocol solving AND has 0 error on *all* inputs.

Theorem 2.3.14 ([31]). *For symmetric distributions* $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \beta & 0 \\ \hline \end{array}$ *we have*

$$\bar{\mathsf{IC}}_\mu(AND, 0) = \frac{\beta}{\ln 2} + \frac{\beta^2}{\alpha}\log\frac{\beta}{\alpha+\beta} + \alpha\log\frac{\alpha+\beta}{\alpha}.$$

Theorem 2.3.15 ([31]). *For distributions* $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & 0 \\ \hline \end{array}$ *we have*

$$\bar{\mathsf{IC}}_\mu(AND, 0) = (\alpha+\beta)H\left(\frac{\beta}{\gamma}\frac{\alpha+\gamma}{\alpha+\beta}\right) - \alpha H\left(\frac{\beta}{\gamma}\right) +$$
$$+ t\bar{\mathsf{IC}}_\nu(AND, 0),$$

where $t = 2\beta + \frac{\alpha\beta}{\gamma}$ *and* $\nu = \begin{array}{|c|c|}\hline \frac{\beta\alpha}{\gamma t} & \frac{\beta}{t} \\ \hline \frac{\beta}{t} & 0 \\ \hline \end{array}$.

Theorem 2.3.16 ([31]).

$$\max_{\mu:\mu(1,1)=0} \bar{\mathsf{IC}}_\mu(AND, 0) = 0.482702\ldots.$$

The distribution that achieves this maximum is

$$\mu = \begin{array}{|c|c|}\hline 0.36532\ldots & 0.31734\ldots \\ \hline 0.31734\ldots & 0 \\ \hline \end{array}.$$

2.3.5 A Local Characterization of the Information Cost Function

In this section we prove Lemma 2.3.1, a local characterization of the zero-error information cost function. More precisely, for an arbitrary function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ we shall define a family $\mathfrak{C}(f)$ of functions $\Delta(\mathcal{X} \times \mathcal{Y}) \to \mathcal{Z}$ satisfying certain local concavity constraints. Then we show that each member of $\mathfrak{C}(f)$ is a lower bound on the zero-error information cost function $I(\mu) := \mathsf{IC}_\mu(f, 0)$ of f. It will be evident that $I(\mu)$ itself satisfies the local concavity constraints, i. e., $I(\mu) \in \mathfrak{C}(f)$. Thus the zero-error information cost of a function f is a point-wise maximum over all functions in the family $\mathfrak{C}(f)$. This technique is used to prove that the information cost of Protocol 1 is *exactly* $\mathsf{IC}_\mu(\mathrm{AND}, 0)$.

It turns out that the number of local concavity constraints that are used to define $\mathfrak{C}(f)$ can be greatly reduced if we assume that every bit sent in a protocol π, nearly achieving the information cost of f, is uniformly distributed from an external point of view. In other words, for each node u in a protocol π we have

$$P(\text{owner of } u \text{ sends } 0 | \Pi \text{ reaches } u) = 1/2.$$

We say that such a protocol is in *normal form*. The proof that the normal form assumption can be made without loss of generality is straightforward and appears in the full version of the paper. Now we proceed to define the family $\mathfrak{C}(f)$.

Definition 2.3.17. *Let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a given function. Define a family $\mathfrak{C}(f)$ of all functions $C : \Delta(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}^+$ satisfying the following constraints:*

- $(\forall \mu \in \Delta(\mathcal{X} \times \mathcal{Y}))(f|_{\mathrm{supp}(\mu)}$ *is constant* $\Rightarrow C(\mu) = 0)$,
- $\forall \mu, \mu_0^A, \mu_1^A \in \Delta(\mathcal{X} \times \mathcal{Y})$ *if Alice can send bit B (that is a randomized function of Alice's input x) from μ s. t. $P(B = 0) = P(B = 1) = 1/2$ and $\mu_i^A(x, y) = P(X = x, Y = y | B = i)$ for $i \in \{0, 1\}$ then*

$$C(\mu) \leq C(\mu_0^A)/2 + C(\mu_1^A)/2 + I(X; B|Y),$$

Here $(X, Y) \sim \mu$.

- $\forall \mu, \mu_0^B, \mu_1^B \in \Delta(\mathcal{X} \times \mathcal{Y})$ *if Bob can send bit* B *(that is a randomized function of Bob's input* y*) from* μ *s. t.* $P(B = 0) = P(B = 1) = 1/2$ *and* $\mu_i^B(x, y) = P(X = x, Y = y | B = i)$ *for* $i \in \{0, 1\}$ *then*

$$C(\mu) \leq C(\mu_0^B)/2 + C(\mu_1^B)/2 + I(Y; B|X),$$

Remark 2.3.18. *The notation* $f|_{\text{supp}(\mu)} \equiv Constant$ *means that both parties can determine the function's output under* μ *by looking at their own input - We do not consider the player's output as part of the protocol transcript, so the latter condition need not imply that the function is determined under* μ *from an* external *point of view. The example* $f(0, 0) = 0$, $f(1, 1) = 1$, $\mu(0, 0) = \mu(1, 1) = 1/2$ *illustrates this point.*

Lemma 2.3.19. *Let* $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ *be a given function. Let* π *be a protocol that solves* f *correctly on all inputs. Then for all* $C \in \mathfrak{C}(f)$ *and all* $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$ *we have* $C(\mu) \leq \mathsf{IC}_\mu(\pi)$.

Proof by induction on $c := \mathsf{CC}(\pi)$. When $c = 0$ the claim is clearly true, since then $f|_{\text{supp}(\mu)}$ is constant and hence $C(\mu) = 0$. Also $\mathsf{IC}_\mu(\pi) = 0$.

Assume the claim holds for all c-bit protocols where $c \geq 0$. Consider a $c+1$-bit protocol π. As discussed prior to the proof, we may assume that π is in normal form. Assume that Alice sends the first bit B. If this bit is 0 then Alice and Bob end up with a new distribution on the inputs μ_0^A, otherwise they end up with distribution μ_1^A. After the first bit, the protocol π reduces to a c-bit protocol π^0 if 0 was sent and π^1 if 1 was sent. Since Alice's bit is uniformly distributed we have

$$\begin{aligned} I(\pi; X|Y) &= I(\pi^1; X|Y) + I(\pi^{\geq 2}; X|Y\pi_1) \\ &= I(B; X|Y) + I(\pi^0; X|Y)/2 + I(\pi^1; Y|X)/2. \end{aligned}$$

Similarly for $I(\pi; Y|X)$. Thus we obtain

$$\begin{aligned}
\mathsf{IC}_\mu(\pi) &= \mathsf{IC}_{\mu_0^A}(\pi^0)/2 + \mathsf{IC}_{\mu_1^A}(\pi^1)/2 + I(X; B|Y) \\
&\geq C(\mu_0^A)/2 + C(\mu_1^A)/2 + I(X; B|Y) \text{ (by induction)} \\
&\geq C(\mu) \text{ (by properties of } C)
\end{aligned}$$

□

Corollary 2.3.20. *For all $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ we have*

1. $\mathsf{IC}_\mu(f, 0) \in \mathfrak{C}(f)$,
2. *for all $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$ and for all $C \in \mathfrak{C}(f)$ we have $\mathsf{IC}_\mu(f, 0) \geq C(\mu)$.*
3. *for all $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$ we have $\mathsf{IC}_\mu(f, 0) = \max_{C \in \mathfrak{C}(f)} C(\mu)$.*

2.3.6 The Exact Communication Complexity of Set Disjointness

In this section we leverage our in-depth analysis of the information complexity of the AND function to compute the *exact* randomized communication complexity of three well-studied problems in the communication complexity literature: *Set-Intersection* ($Int_n(X, Y) = \{i \ : \ X_i \wedge Y_i = 1\}$), *Disjointness* ($Disj_n(X, Y) = \neg \bigvee_{i=1}^n (X_i \wedge Y_i)$) and *k-Disjointness* ($Disj_n^k(X, Y) = \neg \bigvee_{i=1}^n (X_i \wedge Y_i)$ where $|X| = |Y| = k$).

While the AND function "embeds" to all three communication problems, they differ in their difficulty. It turns out that solving each of the three problems above is equivalent to solving n independent copies of the AND function, albeit under a different subset of distributions on $\{0, 1\}^2$.

The Set-Intersection problem corresponds to solving n independent copies of AND under the "worst" possible distribution $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & \delta \\ \hline \end{array}$.

because of "information equals amortized communication" ([29]), Thus Theorem 2.3.2 (along with continuity of information cost at error $= 0$) implies that

Corollary 2.3.21. *For all $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that $\delta \to 0$ as $\varepsilon \to 0$ and*

$$(C_\wedge - \delta) \cdot n \leq \mathsf{R}(Int_n, \varepsilon) \leq C_\wedge \cdot n + o(n),$$

where $C_\wedge \approx 1.492$.

For the Set Disjointness problem, we show that solving $Disj_n(X, Y)$ is equivalent to solving n independent copies of AND under the "worst" distribution μ on $\{0,1\}^2$ satisfying $\mu(1,1) = 0$. This distribution therfore has the form:

$$\mu = \begin{array}{|c|c|} \hline \alpha & \beta \\ \hline \gamma & 0 \\ \hline \end{array}.$$

The intuition as to why the above quantity captures the communication required to solve $Disj_n$ is as follows: Since solving Disjointness is equivalent to solving $\bigvee_{i=1}^n (X_i \wedge Y_i)$, then if the (marginal) distribution of a coordinate $\mu_i(X_i, Y_i)$ satisfies $\mu_i(1,1) \geq \omega(1/n)$, the parties can simply exchange a small (sublinear) number of random coordinates, and finish the job with very small communication (since with very high probability they will find an overlapping coordinate). Thus, the above set of distributions captures the hardness of this task. In fact, our result for the Set Disjointness problem follows from a more general theorem we prove, which characterizes the exact randomized communication complexity of "$\bigvee$"-type functions with error tending to zero, in terms of the informational quantity $\mathsf{IC}^0(f, 0)$, which informally measures the information complexity of f under the "worst" distribution supported on $f^{-1}(0)$[3]:

Theorem 2.3.22 ([31])**.** *For any Boolean function $f : \{0,1\}^k \times \{0,1\}^k \to \{0,1\}$, let $g(\bar{x}, \bar{y}) := \vee_{i=1}^n f(x_i, y_i)$, where $\bar{x} = \{x_i\}_{i=1}^n, \bar{y} = \{y_i\}_{i=1}^n$ and $x_i, y_i \in \{0,1\}^k$. Then for all $\varepsilon > 0$, there exists*

[3] An analogous result holds for "$\bigwedge$"-type functions.

$\delta = \delta(f, \varepsilon) > 0$ *such that* $\delta \to 0$ *as* $\varepsilon \to 0$ *and*

$$(\mathsf{IC}^0(f, 0) - \delta) \cdot n \leq \mathsf{R}(g_n, \varepsilon) \leq \mathsf{IC}^0(f, 0) \cdot n + o(n \cdot k),$$

where $\mathsf{IC}^0(f, 0) := \max_{\mu:\mu(1,1)=0} \bar{\mathsf{IC}}_\mu(f, 0)$.[4]

The formal proof is given in the full version of this paper. Here we only present the main ideas. The high-level idea for the upper bound is to produce a low *information* protocol for computing g_n and then use the fact that information = amortized communication to obtain a *low communication* protocol. To this end, we exploit the self-reducible structure of $\bigvee$-type functions. For the lower bound, we show that a low-error protocol for g_n which uses $< \mathsf{IC}^0(f, 0) \cdot n$ communication, can be used to produce a low-error protocol for a single copy of f, whose information under any distribution supported on $f^{-1}(0)$ is $< \mathsf{IC}^0(f, 0)$. Now by using continuity of information cost at error $= 0$ (Theorem 2.3.7), we get a contradiction.

Theorem 2.3.5 now follows from Theorem 2.3.22. For convenience, we restate it below

Corollary 2.3.23 (Theorem 2.3.5 restated)**.** *For all* $\varepsilon > 0$*, there exists* $\delta = \delta(\varepsilon) > 0$ *such that* $\delta \to 0$ *as* $\varepsilon \to 0$ *and*

$$(C_{DISJ} - \delta) \cdot n \leq \mathsf{R}(Disj_n, \varepsilon) \leq C_{DISJ} \cdot n + o(n).$$

where $C_{DISJ} \approx 0.4827$ *bits.*

Proof. Since randomized communication complexity is closed under complementation, $\mathsf{R}(Disj_n, \varepsilon) = \mathsf{R}(\bigvee_{i=1}^n (X_i \wedge Y_i), \varepsilon)$, and thus Theorem 2.3.22 (with $f = AND$ and $k = 1$) implies that

$$(\mathsf{IC}^0(AND, 0) - \delta) \cdot n \leq \mathsf{R}(Disj_n, \varepsilon) \leq \mathsf{IC}^0(AND, 0) \cdot n + o(n).$$

[4]Note that this quantity is not zero, since our definition of $\bar{\mathsf{IC}}_\mu(\mathrm{f}, 0)$ ranges only over protocols which solve f for *all* inputs.

But Theorem 2.3.16 asserts that $\max_{\mu:\mu(1,1)=0} \bar{\mathsf{IC}}_\mu(\text{AND}, 0) = 0.4827\ldots$, which completes the proof. □

The communication complexity of the k-Disjointness problem is known to be $\Theta(k)$ [80]. We are able to determine the exact constant in this regime as well.

Theorem 2.3.24 ([31]). *(Theorem 2.3.6 restated) Let n, k be such that $k = \omega(1)$ and $n/k = \omega(1)$. Then for all constant $\varepsilon > 0$,*

$$\left(\frac{2}{\ln 2} - O(\sqrt{\varepsilon})\right) \cdot k - o(k) \leq R_\varepsilon(DISJ_n^k) \leq \frac{2}{\ln 2} \cdot k + o(k).$$

To this end, we consider the set of distributions taking the form:

$$\mu_k = \begin{array}{|c|c|} \hline 1 - 2k/n - o(k/n) & k/n \\ \hline k/n & o(k/n) \\ \hline \end{array}.$$

The formula in Theorem 2.3.10 implies that $IC_{\mu_k}(AND, 0) = \frac{2}{\ln 2}\frac{k}{n} \pm o(\frac{k}{n})$. The proof of Theorem 2.3.24 follows the ideas of Theorem 2.3.22, but is considerably more complicated, mainly due to the fact that $IC_{\mu_k}(AND, 0)$ is now tiny. We need to use a more nuanced approach to get similar bounds, and in particular strengthen the rate of convergence of continuity of the information complexity of AND at $\varepsilon = 0$, using a recursive application of our optimal protocol from section 2.3.4. For a formal proof see the full version of this paper.

2.3.7 Rate of Convergence

In this section we prove that for most distributions μ the rate at which $\mathsf{IC}_\mu^r(\text{AND}, 0)$ converges to $\mathsf{IC}_\mu(\text{AND}, 0)$ is $\Theta(1/r^2)$. We also present implications that this result has in communication complexity. The empirical evidence that the rate of convergence is $\Theta(1/r^2)$ has appeared in the information theory literature prior to our work. In [121], Ishwar and

Ma consider the task f of computing AND when only Bob is required to learn the answer. They derive an explicit formula for $\mathsf{IC}_\mu(f)$ for product distributions μ and design an algorithm that computes $\mathsf{IC}^r_\mu(f)$ to within a desired accuracy. Ishwar and Ma generously provided their scripts, which we used to generate Figure 2.2 (it is a variant of Figure 4(a) from [121]). Figure 2.2 demonstrates that $\max_{\mu \text{ - product}} \mathsf{IC}^r_\mu(f) - \mathsf{IC}_\mu(f)$ asymptotically behaves like $\Theta(1/r^2)$.

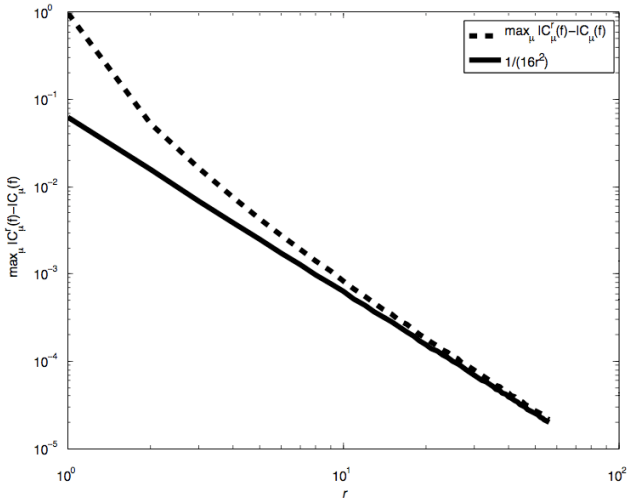

Figure 2.2: Empirical evidence that rate of convergence is $\Theta(1/r^2)$. The log-log scale figure shows the graph of $\max_{\mu \text{ - product}} \mathsf{IC}^r_\mu(f) - \mathsf{IC}_\mu(f)$ for a range of values r together with the line $1/(16r^2)$. The x-axis is the number of rounds r. The y-axis is the change in the information cost $\max_{\mu \text{ - product}} \mathsf{IC}^r_\mu(f) - \mathsf{IC}_\mu(f)$.

Our proof of the rate of convergence consists of two parts: (1) the lower bound $\Omega(1/r^2)$ on the rate of convergence and (2) a matching upper bound $O(1/r^2)$.

Theorem 2.3.25 ([31]). *For all* $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & \delta \\ \hline \end{array}$ *with* $\{\alpha, \beta, \gamma\} \subseteq \operatorname{supp}(\mu)$ *we have*

$$\mathsf{IC}^r_\mu(AND, 0) = \mathsf{IC}_\mu(AND, 0) + \Omega_\mu\left(\frac{1}{r^2}\right).$$

We present the high-level idea of the proof of Theorem 2.3.25. Let π be an r-round protocol that solves AND with 0-error on all inputs. We may view π as a random walk on $\Delta(\{0,1\}^2)$. Each round is a random step made by either Alice or Bob. Suppose that

the statistical distance traveled by a player in the wrong region during ith message (see Section 2.3.4 for the definition of Alice's, Bob's and diagonal regions) is ϵ_i. Then the first observation is that such a step wastes $\Omega(\epsilon_i^3)$ information as compared to an optimal protocol. The second observation is that any feasible protocol must travel a total distance of $\Omega(1)$ in the wrong region, thus $\sum_{i=1}^{r} \epsilon_i = \Omega(1)$. Then the overall wastage $\Omega(\sum_{i=1}^{r} \epsilon_i^3)$ is minimized for $\epsilon_i = 1/r$, and hence the total extra information leaked is $\Omega(1/r^2)$.

Theorem 2.3.26 ([31])**.**

$$\mathsf{IC}_{\mu}^{r}(AND, 0) = \mathsf{IC}_{\mu}(AND, 0) + O_{\mu}\left(\frac{1}{r^2}\right).$$

The upper bound on the rate of convergence is obtained by analyzing a family $(\pi_r)_{r=1}^{\infty}$ of $2r$-round protocols. Protocol π_r is a discretization of our infeasible Protocol 1, where Alice and Bob are allowed to generate their random numbers N^A and N^B only from a finite set $\left\{\frac{0}{r}, \frac{1}{r}, \cdots, \frac{r-1}{r}\right\}$. The most natural way to discretize our continuous AND protocol would be to sample numbers N^A and N^B uniformly at random from the set $\{\frac{0}{r}, \ldots, \frac{r-1}{r}\}$ when the corresponding player(s) have 0 as input. While analyzing this option, we discovered that this discretization wastes increasing amounts of information in later rounds as the counter C approaches r. This leads to a total information wasted $\approx \frac{1}{r^2} \sum_{i=1}^{r} \frac{1}{i} = \Theta\left(\frac{\log r}{r^2}\right)$. A natural remedy is to select numbers N^A and N^B non-uniformly, assigning less probability mass to the later rounds. Indeed, our discretized protocol π_r assigns probability $\frac{2r-2i-1}{r^2}$ to the ith value of N^A and N^B leading to the correct $O(\frac{1}{r^2})$ bound on the total information wasted. Theorem 2.3.26 follows from a careful analysis and calculation of round-by-round information cost difference between the discretized and continuous protocols.

The full proofs of the above theorems appear in the full version of the paper.

From the $\Theta(1/r^2)$-bound on the rate of convergence of r-round information cost of AND function together with results from Section 2.3.5 we can derive conclusions about

the utility of rounds in the communication complexity problems discussed earlier. The rate of convergence result implies that both for set intersection and for set disjointness an r-round protocol will we suboptimal by $\Theta(n/r^2)$ bits. Thus for both problems a protocol that is optimal up to lower-order terms will need to use $\omega(1)$ rounds of communication. It is quite possible that a similar statement holds for size-k disjointness, but our rate of convergence results do not imply this.

Chapter 3

Direct Sums and Products and the Interactive Compression Problem

3.1 The Direct Sum and Product Conjectures

Direct sum and direct product theorems assert a lower bound on the complexity of solving n copies of a problem f in parallel, in terms of the cost of a single copy. Let f^n denote the function which maps the tuple $((x_1, \ldots, x_n), (y_1, \ldots, y_n))$ to $(f(x_1, y_1), \ldots, f(x_n, y_n)))$, and $C(f)$ denote the cost of solving f (in some arbitrary computational model). The obvious solution to f^n is to apply the single-copy optimal solution n times sequentially and independently to each coordinate, yielding a linear scaling of the resources, so clearly $C(f^n) \leq n \cdot C(f)$. The *strong direct sum* conjecture postulates that this naive solution is essentially tight. When the computational model is randomized (as in randomized communication complexity), a direct sum theorem aims to give a lower bound (ideally, linear in n) on the resources for computing f^n with some fixed *overall* error $\varepsilon > 0$ in terms of the cost of computing a single copy of f with the same (or comparable) error ε. In the context of communication complexity, the strong direct sum conjecture informally asks whether

$$\mathsf{D}_{\mu^n}(f^n, \varepsilon) = \Omega(n) \cdot \mathsf{D}_{\mu}(f, \varepsilon) \ ? \tag{3.1}$$

A *direct product* theorem further asserts that unless sufficient resources are provided, the probability of successfully computing all n copies of f will be exponentially small, potentially as low as $(1-\varepsilon)^{\Omega(n)}$. This is intuitively plausible, since the naive solution which applies the best protocol for one copy of f independently to each of the n coordinates, would succeed in solving f^n probability $(1-\varepsilon)^n$. Can one do better?

To make this more precise, let us denote by $\mathsf{suc}(\mu, f, C)$ the maximum success probability of a protocol with communication complexity $\leq C$ in computing f under input distribution μ. A direct product theorem asserts that any protocol attempting to solve f^n (under μ^n) using some fixed number T of communication bits (ideally $T = \Omega(n \cdot C)$), will succeed only with exponentially small probability: $\mathsf{suc}(\mu^n, f^n, T) \lesssim (1-\varepsilon)^{\Omega(n)}$. Informally, the strong direct product question asks whether

$$\mathsf{suc}(\mu^n, f^n, o(n \cdot C)) \lesssim (\mathsf{suc}(\mu, f, C))^{\Omega(n)} \quad ? \tag{3.2}$$

The difference between a direct sum theorem and the (stronger) direct product theorem can be put as follows: A direct sum result fixes the success probability (of both the single-copy and parallel computation), and focuses on the increase in resources; A direct product result fixes the resources T for the parallel computation, and focuses on the decay of success probability (hence the terms "sum" and "product").

Classic direct product results in complexity theory are Raz's Parallel Repetition Theorem [134, 133] and Yao's XOR Lemma [155]. The value of such results to complexity theory is clear: direct sum and product theorems, together with a lower bound on the (easier-to-reason-about) sub-problem, yield a lower bound on the composite problem in a "black-box" fashion (a method also known as *hardness amplification*). For example, the Karchmer-Raz-Wigderson program for separating $\mathbf{P}$ from $\mathbf{NC}^1$ can be completed via a (currently open) direct sum conjecture for Boolean formulas [99] (after more than a decade, some progress on this conjecture was recently made using information-complexity machinery [72]). Other fields in which direct sums and products have played a central role in proving tight lower bounds are streaming [14, 138, 125, 77] and distributed computing [84].

Can we always hope for such strong lower bounds to hold? It turns out that the validity of these conjectures highly depends on the underlying computational model, and the short answer is no[1]. What about the communication complexity model? This question has had a long history and was answered positively for several restricted models of communication [108, 141, 115, 144, 94, 125, 128] (For a broader overview of direct sums

[1]In the context of circuit complexity, for example, this conjecture fails (at least in its strongest form): Multiplying an $n \times n$ matrix by a (worst case) n-dimensional vector requires n^2 operations, while (deterministic) multiplication of n different vectors by the same matrix amounts to matrix-multiplication of two $n \times n$ matrices, which can be done in $n^{2.37} \ll n^3$ operations [151].

and products and their importance in communication complexity we refer the reader to [94, 36] and references therein). In the *determistic* communication complexity model, Feder et al. [64] showed that $\mathsf{D}(f^n) \geq n \cdot \Omega\left(\sqrt{\mathsf{D}(f)}\right)$. In the (unbounded-round) randomized communication model, however, there is a tight connection between the direct sum question for the (2-party) function f and its information complexity. By now, this should come as no surprise: The "Information=Amortized Communication" theorem (Theorem 1.0.1) asserts that, for large enough n, it holds that $\mathsf{D}_{\mu^n}(f^n, \varepsilon) \gtrsim n \cdot \mathsf{IC}_\mu(f, \varepsilon)$, and hence the direct sum question (3.1) boils down to understanding the relationship between $\mathsf{D}_\mu(f, \varepsilon)$ and $\mathsf{IC}_\mu(f, \varepsilon)$. Note this question is in fact a question about the ability to compress interactive communication protocols in the "one-shot" regime:

Problem 3.1.1 (Interactive compression problem [2], [16])**.** *Given a protocol π over inputs $x, y \sim \mu$, with $\|\pi\| = C, \mathsf{IC}_\mu(\pi) = I$ ($I \ll C$), what is the smallest amount of communication of a protocol τ which (approximately) simulate π (i.e., $|\tau(x, y) - \pi(x, y)|_1 \leq \delta$ for a small constant δ)?*

In particular, if one could compress any protocol into $O(I)$ bits, Corollary 3.1.2 would have implied the strong direct sum conjecture. In fact, the additivity of information cost implies the following general quantitative relationship between (possibly weaker) interactive compression results and direct sum theorems in communication complexity:

Claim 3.1.2 (One-Shot Compression implies Direct Sum)**.** *Suppose that for any given protocol π for which $\mathsf{IC}_\mu(\pi) = I$, $\|\pi\| = C$, there is a compression scheme that δ-simulates[3] π using $g_\delta(I, C)$ bits of communication. Then*

$$g_\delta\left(\frac{\mathsf{D}_{\mu^n}(f^n, \varepsilon)}{n}, \mathsf{D}_{\mu^n}(f^n, \varepsilon)\right) \geq \mathsf{D}_\mu(f, \varepsilon + \delta).$$

[2]One could argue that the requirement in Problem 3.1.1 is too harsh as it requires a simulation of the entire transcript of (an arbitrary) protocol, while in the direct sum context we are merely interested in the output of f. This is a valid point, but all known compression schemes satisfy the stronger condition, and therefore this became the standard problem formulation. (a more formal equivalence argument between compression and direct sum theorems appears in [35])

[3]The simulation here is in an internal sense.

Proof. Let Π be an optimal n-fold protocol for f^n under μ^n, i.e., $\|\Pi\| = \mathsf{D}_{\mu^n}(f^n, \varepsilon) := C_n$. By Lemma 1.1.20 (equation (1.3)), there is a single-copy protocol θ whose information cost is at most $\mathsf{IC}_{\mu^n}(\Pi)/n \leq C_n/n$ (since communication always upper bounds information). But the premise of the claim guarantees that θ can now be δ-simulated using $g_\delta(C_n/n, C_n)$ communication, so as to produce a single-copy protocol with error $\leq \varepsilon + \delta$ for f, and therefore $\mathsf{D}_\mu(f, \varepsilon + \delta) \leq g_\delta(C_n/n\ ,\ C_n)$. □

The first general interactive compression result was proposed in the seminal work of Barak, Braverman, Chen and Rao [16], who showed that any protocol π can be δ-simulated using $g_\delta(I, C) = \tilde{O}_\delta(\sqrt{C \cdot I})$ communication. Plugging this compression result into Claim 3.1.2, this yields the following weaker direct sum theorem:

Theorem 3.1.3 (Weak Direct Sum, [16])**.** *For every Boolean function f, distribution μ, and any positive constant $\delta > 0$,*

$$\mathsf{D}_{\mu^n}(f^n, \varepsilon) \geq \tilde{\Omega}(\sqrt{n} \cdot \mathsf{D}_\mu(f, \varepsilon + \delta)).$$

Later, Braverman [28] showed that it is always possible to simulate π using $2^{O_\delta(I)}$ bits of communication, thereby exhibiting the first interactive compression scheme which depends solely on the information cost of π. Notice that the last two compression results are indeed incomparable, since the communication of π could be arbitrarily larger than its information complexity (e.g., $C \geq 2^{2^{2^I}}$). The current state of the art for the *general* problem can be therefore summarized as follows: Any protocol with communication C and information cost I can be compressed to

$$g_\delta(I, C) \leq \min\left\{2^{O_\delta(I)}\ ,\ \tilde{O}_\delta(\sqrt{I \cdot C})\right\}. \tag{3.3}$$

Can we hope to compress all the way down to $O(I)$? Unfortunately, this task turns out to be too ambitious: In a recent breakthrough, Ganor, Kol and Raz [71] showed that

$$g_\delta(I, C) \geq \max\left\{2^{\Omega(I)}\ ,\ \tilde{\Omega}(I \cdot \log C)\right\}. \tag{3.4}$$

More specifically, they exhibit a Boolean function f which can be solved using a protocol with information cost I, but cannot be simulated by a protocol π' with communication cost $< 2^{O(I)}$. Since the *communication* of their low information protocol is 2^{2^I}, this also rules out a compression to $I \cdot o(\log C)$, or else such compression would have produced a too good to be true ($2^{o(I)}$ communication) protocol. Though the margin of this text is too narrow to contain their proof, we remark that this result was particularly challenging in light of another line of work which showed that essentially all previously known techniques for proving communication lower bounds apply to information complexity as well [39, 103], and hence could not be used to separate information complexity and communication complexity. Using Claim 3.1.2, the compression lower bound in (3.4) refutes the strongest possible direct sum, but leaves open the following gap

$$O\left(\frac{n}{\log n}\right) \;\geq\; \min_{f} \frac{\mathsf{D}_{\mu^n}(f^n, \varepsilon)}{\mathsf{D}_{\mu}(f, \varepsilon + \delta)} \;\geq\; \tilde{\Omega}\left(\sqrt{n}\right). \tag{3.5}$$

Notice that this still leave a huge gap in the direct sum question, which has yet to be resolved. It is still conceivable that improved compression to $g_\delta(I, C) = I \cdot C^{o(1)}$ is in fact possible, and the quest to beat the compression scheme of [16] remains open (we discuss a potential approach in Section 3.4).

Despite the lack of progress in the general regime, several works showed that it is in fact possible to obtain near-optimal compression results in restricted models of communication: When the input distribution μ is a *product distribution*, [16] show a near-optimal compression result, namely that π can be compressed into $O(I \cdot polylog(C))$ bits. Once again, using Claim 3.1.2 this yields the following direct sum theorem

Theorem 3.1.4 ([16])**.** *For every product distribution μ and any $\delta > 0$,*

$$\mathsf{D}_{\mu^n}(f^n, \varepsilon) = \tilde{\Omega}(n \cdot \mathsf{D}_{\mu}(f, \varepsilon + \delta)).$$

Near optimal compression results were recently proven for *public-coin protocols* (under arbitrary distributions) [45, 18], and for bounded-round protocols, leading to near-optimal direct sum theorems in corresponding communication models. We summarize these results in Table 5.1.

Reference	Regime	Communication Complexity
[35, 37]	r-round protocols	$I + O\left(\sqrt{r \cdot I}\right) + O(r \log 1/\delta)$
[18] (improved [45])	Public coin protocols	$O(I^2 \cdot \log\log(C)/\delta^2)$
[16]	Product distributions[a]	$O(I \cdot \mathrm{poly}\log(C)/\delta)$
[28, 16]	**General protocols**	$\min\{2^{O(I/\delta)}\ ,\ O(\sqrt{I \cdot C} \cdot \log(C)/\delta)\}$
[71]	**Best lower bound**	$\max\{2^{\Omega(I)}\ ,\ \Omega(I \cdot \log(C))$

Table 3.1: Best to date compression schemes, for various regimes.

[a] This result in fact holds for general (non-product) distributions as well, when the compression is with respect to I^{ext}, the external information cost of the original protocol π.

3.1.1 From Direct Sum to Direct Product

Note that Theorem 3.1.3 and Theorem 3.1.4 can be rephrased as stating that $\mathsf{suc}(\mu^n, f^n, o(\sqrt{n \cdot C})) \leq \varepsilon$ (in the general case) and $\mathsf{suc}(\mu^n, f^n, o(n \cdot C)) \leq \varepsilon$ (in the product case), where ε is *constant* (say, $2/3$). Our first result in this line of research is converting those direct sum results into *direct product* results, thus proving the first *exponentially small* upper bounds on the success probability of parallel computation in the two-party unbounded communication complexity model:

Theorem 3.1.5 (([36], [37] informally stated)**.** *There are universal constants $\alpha, \beta > 0$ such that for any two-party function f the following holds:*

- *If $\mathsf{suc}(\mu, f, C) = \frac{2}{3}$ and $T \log^{3/2} T \leq \alpha C \cdot \sqrt{n}$, then $\mathsf{suc}(\mu^n, f^n, T) \leq \exp(-\Omega(n))$.*
- *For product distributions μ, we show that if $\mathsf{suc}(\mu, f, C) \leq \frac{2}{3}$ and $T \log^2 T \leq \beta C n$, then $\mathsf{suc}(\mu^n, f^n, T) \leq \exp(-\Omega(n))$.*

- *For bounded-round protocols using only r rounds, we show that if* $\mathsf{suc}_{7r}(\mu, f, C) \leq \frac{2}{3}$ *and* $T \leq (C - \Omega(r \log r)) \cdot n$ *then* $\mathsf{suc}_r(\mu^n, f^n, T) \leq \exp(-\Omega(n))$ *(where* $\mathsf{suc}_r(\mu, f, C)$ *denotes the maximum success probability of an r-round communication protocol using $\leq C$ bits).*

We prove the above theorem in the next section (Section 3.2). Notice that the last two propositions essentially close the direct product conjecture in the bounded-round and product regimes.

3.1.2 Strong Direct Product in Terms of Information Complexity

In the same spirit as above, the "Information=Amortized Communication" theorem (Theorem 1.0.1 [35]) is merely a *direct sum* theorem: It asserts that any protocol for f^n under μ^n using $o_\varepsilon(n \cdot I_\mu)$ communication must fail with *overall* error ε simultaneously on all copies, namely

$$\mathsf{suc}(\mu^n, f^n, o(n \cdot I)) \leq \varepsilon.$$

In contrast, the *upper bound* in Theorem 1.0.1 guarantees that $O_\varepsilon(n \cdot I_\mu)$ communication suffice to compute f^n *only with per-copy* error of ε. Indeed, the upper bound is proved by executing independent copies (of the information-optimal protocol) in parallel, and therefore the overall success probability of this parallel protocol on all copies simultaneously is only $\approx (1-\varepsilon)^n$. Therefore, one could hope to prove the following direct product theorem: "a protocol which uses $\ll n \cdot \mathsf{IC}(f, \mu, \varepsilon)$ communication to solve n copies of f cannot succeed with probability more than $(1-\varepsilon)^{\Omega(n)}$". In this section we (essentially) prove this result.

Several prior works (e.g [92, 94, 36]) aim to get a generic direct product theorem for communication complexity. Other works prove a direct product theorem in terms of weaker complexity measures of the underlying function, such as the discrepancy $\mathsf{disc}_\mu(f)$ of the function ([115]) or the (stronger) smooth rectangle bound [93]. More precisely, Jain and Yao [95] show that any protocol attempting to compute f^n under μ^n using $\ll$

$n \cdot \mathsf{srect}_\mu(f)$ communication, will succeed with probability only $2^{-\Omega(n)}$, where $\mathsf{srect}_\mu(f)$ denotes the smooth rectangle bound of f under μ. Our direct product theorem implies all previous results in this category, since it has been shown that $\mathsf{IC}_\mu(f) \geq \mathsf{srect}_\mu(f) \geq \mathsf{disc}_\mu(f)$ (see [102]). Moreover, the discussion in the previous paragraph asserts that our direct product result (Theorem 3.1.6) is asymptotically tight (as communication and information are asymptotically equal), while such guarantee is not known to hold for the previous measures.

To turn the direct product result of [36] into a direct product theorem *in terms of information complexity*, one needs a generic way of turning a protocol that is statistically close to a low-information one into a low information protocol (this will become clearer in the proof of Theorem 3.1.5 in Section 3.2). Prior to the present paper, no such way was known. To do so, we introduce a primitive that allows for such conversion (we call this an "information odometer"). Combining this tool together with [36], we obtain an essentially optimal direct product theorem for communication complexity in terms of information complexity (Theorem 3.1.6 below).

To state the result more formally, let us denote by $\mathsf{suc}^{\mathsf{i}}(\mu, f, I)$ the maximum success probability of a protocol with *information* cost (at most) I in computing f under μ. We prove:

Theorem 3.1.6 ([40], informally stated)**.** *There is a global constant $\alpha > 0$ such that for any two-party function f the following holds: If $\mathsf{suc}^{\mathsf{i}}(\mu, f, I)) \leq 2/3$ and $T \log(T) < \alpha n \cdot I$, it holds that $\mathsf{suc}(\mu^n, f^n, T) \leq \exp(-\Omega(n))$.*

This theorem is tight (up to polylogarithmic factors), as mentioned above, since Theorem 1.0.1 proves that $O(n \cdot I)$ communication suffice to succeed on f^n under μ^n with probability $\approx (1-\varepsilon)^n$. Thus, the above theorem can be viewed as a sharpening of the celebrated "Information = amortized communication" theorem. In fact, Result 3 shows that direct product theorems in two-party communication complexity are equivalent to direct sum theorems, and are both equivalent to the interactive compression Problem 3.1.1 (in

the sense that if one can prove that for a T-bit protocol for f^n, $\mathsf{suc}(\mu^n, f^n, T) \leq 3/4$, then in fact $\mathsf{suc}(\mu^n, f^n, \tilde{\Omega}(T)) \leq exp(-\Omega(n))$). The proof of Theorem 3.1.6 appears in Section 3.4.

3.2 Proof of the Direct Product Result

Here we present the proof of Theorem 3.1.5. Our formal results are as follows:

Theorem 3.2.1 (General Direct Product [36]). *There is a universal constant $\alpha > 0$ such that if f is boolean*[4]*, $\gamma = 1-\mathsf{suc}(\mu, f, C)$, $T \geq 2$, and $T \log^{3/2} T < \alpha\gamma^{5/2}(C-1)\sqrt{n}$, then* $\mathsf{suc}(\mu^n, f^n, T) \leq \exp(-\alpha\gamma^2 n)$.

For product distributions, we obtain the following stronger result:

Theorem 3.2.2 (Strong Direct Product for Product Distributions [36]). *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* $\mathsf{suc}(\mu^n, f^n, T) \leq \exp(-\alpha\gamma^2 n)$.

Finally, for bounded-round protocols, we improve on the work of [94] and show a near-optimal direct product theorem:

Theorem 3.2.3 (Strong Direct Product for Bounded-Round Protocols [37]). *Let $\mathsf{suc}_r(\mu, f, C)$ denote the largest success probability of a protocol using at most r rounds in predicting f under μ. Then if $\mathsf{suc}_{7r}(\mu, f, C) \leq \frac{2}{3}$ and $T \leq (C - \Omega(r \log r)) \cdot n$ then $\mathsf{suc}_r(\mu^n, f^n, T) \leq \exp(-\Omega(n))$ (where $\mathsf{suc}_r(\mu, f, C)$ denotes the maximum success probability of an r-round communication protocol using $\leq C$ bits).*

Due to space constraints, here we only present the proofs of the main arguments. For missing proofs and a broader introduction, we refer the reader to the full version of this paper. The proof of Theorem 3.2.3 is based on Theorem 3.2.1, combined

[4]We remark that when f is a function that has a k-bit output, the above theorem is true with $(C-1)$ replaced by $(C-k)$. For simplicity, we focus on the case $k = 1$ throughout this paper. 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 $\mathsf{suc}(\mu^n, f^n, T) \leq \exp(-\Omega(n))$.

with a new compression result for bounded-round protocols, which allows a better round/communication tradeoff than [35]. We omit this result and refer the reader to the full paper [37].

Overview of the Proof

Our proof follows the logic of the direct sum theorem of Barak et. al. [17] (Theorems 3.1.3 and 3.1.4 above).Let T denote the communication complexity of the best n-fold protocol for f^n under μ^n. The first step of [17]'s proof 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 [134]. 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 (since $\pi(W_i|E) = 1 \; \forall \; i$). In fact, we shall prove that this holds for an arbitrary event W, not just E.

Lemma 3.2.4 (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 $\|\pi\| \log^{3/2} \|\pi\| < \alpha\gamma^{5/2}(C-1)\sqrt{n}$, then $(1/n)\sum_{i\in[n]} \pi(W_i|W) \leq \mathsf{suc}(\mu, f, C) + \gamma/\alpha$.*

Lemma 3.2.5 (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 [17]. However there are substantial differences in our proof, which are discussed in detail below. First let us see how Lemma 3.2.4 implies Theorem 3.2.1. Theorem 3.2.2 follows from Lemma 3.2.5 in the same way.

Proof of Theorem 3.2.1. 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 3.2.4, either $\|\pi\| < 2$, $\|\pi\| \log^{3/2} \|\pi\| \geq \alpha^{7/2} 2^{-5/2}(1-\mathsf{suc}(\mu, f, C))^{5/2} C\sqrt{n}$, or $\pi(E) < 2^{-\gamma^2 n}$. □

Due to space constraints, we leave out the formal proofs of the main lemmas (these can be found in Section 3 in the full version of this paper [36]). At a high level, the proofs of the lemmas are quite similar to each other, though there are some technical differences. We discuss Lemma 3.2.5 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:

1. **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 [17]), $\pi(m|x_i y_i W)$ 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.

2. **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 design a protocol θ. The parties publicly sample a uniformly random coordinate i in $[n]$ and set $x_i = x, y_i = y$. They also publicly sample a variable r_i that contains a subset of the variables $x_1, \ldots, x_n, y_1, \ldots, y_n$. Each message m_j sent by the first party in π is sampled according to the distribution $\pi(m_j|m_{<j}x_i r_i W)$, and each message sent by the second party is sampled according to the distribution $\pi(m_j|m_{<j}y_i r_i W)$ (the aforementioned notation is formally defined in the subsequent preliminaries section). We prove that for average i, $\theta(x_i y_i r_i m) \overset{\gamma}{\approx} \pi(x_i y_i r_i m|W)$. [94] analyzed a different protocol θ, which used a different definition of r_i, and showed that for average i, $\theta(x_i y_i r_i m) \overset{\gamma t}{\approx} \pi(x_i y_i r_i m|W)$, where here t is the number of rounds of communication in π. Our bound is independent of t, a feature that is essential to our results. A crucial technical feature of our protocol is the definition of r_i, which allows us to split the dependencies between inputs to π in a new way. This allows us to control the effect of the dependencies introduced by W using a bound that is independent of the number of rounds in π.

To solve the second challenge, we need to come up with a way to *compress* the protocol θ. To use the compression methods of [17], 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 [17] 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.

In Section 3.4, we exhibit an interactive technique for converting such a protocol to a protocol which *actually has low information* . Of course, the major challenge is doing so in a way that does not increase (by much) the information cost of the original protocol. We obtain the following result:

Theorem 3.2.6 (Conditional abort theorem, [40])**.** *Let θ be an alternatingprotocol with inputs $x, y \sim \mu$, public randomness r, and messages m, and suppose q is another distribution on these variables such that $\theta(xyrm) \overset{\epsilon}{\approx} q(xyrm)$. Denote $I_q := I_q(X; M|YR) + I_q(Y; M|XR)$. Then, there exists a protocol π that 15ϵ-simulates θ with $\|\pi\| \leq O(\|\theta\| \log(\|\theta\|))$ and*

$$\mathsf{IC}_\mu(\pi) \leq O\left(\frac{I_q + \log(\|\theta\| + 1)}{\varepsilon^2}\right).$$

As mentioned above, we outline the (highly nontrivial) proof of the above theorem in Section 3.4. For simplicity, let us argue how this "conversion" can be easily done, without interaction (!), for the *external information cost measure* (or alternatively, under product distributions). Indeed, 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 has significantly exceeded the typical amount of information (the crucial observation is that, under product distributions, this quantity can be *privately* estimated by any party, while for general (correlated) distributions this is impossible without interaction. See Section 3.4 for further discussion). This procedure 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.

With Theorem 3.2.6 in hand, the final protocol computing f is obtained by compressing τ using the methods of [17]. Indeed, for our choice of parameters, this compression would yield a single-copy protocol for f under μ, with success probability $> \mathsf{suc}(\mu, f, C)$ and communication $\leq C$, which is a contradiction.

We start by preparing our working horses for the proof of Lemmas 3.2.4 and 3.2.5, These tools and properties will enable us to sample and analyze the conditional "protocol" $\pi(m|x_i y_i W)$, and will also be used in Section 3.4, where we prove a strong direct product theorem in terms of information complexity. In Section 3.2 we present the proof of Lemma 3.2.4 (and outline the proof of Lemma 3.2.5). As noted above, the full proof of Theorems 3.2.1 and 3.2.2 can be found in the full version of the paper [36].

Preliminaries and Useful Inequalities

Notation

In what follows, 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 $A = a$).

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 $\mathbb{E}_{p(a)}[g(a)]$ the expected value of $g(a)$ with respect to $a \sim p$.

For two distributions p, q, we write $p \overset{\epsilon}{\approx} q$ if $|p - q|_1 \leq \epsilon$. Given distributions $p_1, \ldots, p_n$ and $q_1, \ldots, q_n$, we sometimes say "in expectation over i sampled according to $\eta(i)$, $p_i \overset{\gamma}{\approx} q_i$" when we mean that $\mathbb{E}_{\eta(i)}[|p_i - q_i|_1] \leq \gamma$.

In the above terminology, the *mutual information* between A, B conditioned on C in the joint probability space $p(\cdot)$ is

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

In this section, we will often analyze information terms conditioned on events W. In this case, we denote $I_p(A; B|CW) = I_q(A; B|C)$ where $q(abc) = p(abc|W)$.

Internal Simulation

Given a protocol π that operates on inputs x, y drawn from a distribution μ using public randomness[5] r and messages m, we write $\pi(xymr)$ to denote the joint distribution of these variables. For an an (arbitrary) distribution $q(x, y, a)$ in the same probability space of π, we say that π *δ-simulates* q, if there is a function g and a function h such that

$$\pi(x, y, g(x, r, m), h(y, r, m)) \overset{\delta}{\approx} q(x, y, a, a),$$

where $q(x, y, a, a)$ is the distribution on 4-tuples (x, y, a, a) where (x, y, a) are distributed according to q. Thus if π δ-simulates q, the protocol allows the parties to sample a ac-

[5]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.

cording to $q(a|xy)$. If in addition $g(x, r, m)$ does not depend on x, we say that π *strongly* δ-simulates q. Thus if π strongly simulates q, then the outcome of the simulation is apparent even to an observer that does not know x or 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 $\lambda(x, y, (r', m'))$. Similarly, we say that π strongly δ-simulates λ if π strongly δ-simulates $\lambda(x, y, (r', m'))$. We say that π computes f with success probability $1 - \delta$, if π strongly δ-simulates $\pi(x, y, f(x, y))$.

Useful inequalities

Missing proofs of the following simple facts can be found in [58].

Fact 3.2.7 (Projection minimizes divergence)**.** *Let $T, X, Y \sim p(txy)$ be (correlated) random variables in the same probability space. Then for any random variable $Z = Z(y) \sim q$, it holds that*

$$\forall y \quad \mathbb{E}_{x|y} \left[\mathbb{D} \left(T|xy \| T|y \right) \right] \leq \mathbb{E}_{x|y} \left[\mathbb{D} \left(T|xy \| Z \right) \right].$$

Proof. Fix any y and denote $T' := T|y$, $T'|x := T|xy$ and $p'(tx) := p(tx|y)$. Then

$$\begin{aligned} &\mathbb{E}_{x|y} \left[\mathbb{D} \left(T|xy \| T|y \right) \right] - \mathbb{E}_{x|y} \left[\mathbb{D} \left(T|xy \| Z \right) \right] = \mathbb{E} \left[\mathbb{D} \left(T'|x \| T' \right) \right] - \mathbb{E} \left[\mathbb{D} \left(T'|x \| Z \right) \right] \\ &= \sum_{xt} p'(xt) \left[\log \frac{p'(tx)}{p'(t)} - \log \frac{p'(tx)}{q(t)} \right] = \sum_{xt} p'(xt) \log \frac{q(t)}{p'(t)} = -\mathbb{D} \left(p'(t) \| q(t) \right) \leq 0 \end{aligned}$$

where the last transition is by Fact 1.1.15. Rearranging completes the proof. □

Proposition 3.2.8 (Properties of binary entropy)**.** *For any $x \in [0, 1]$, the binary entropy function $H(x)$ satisfies the following properties:*

1. *$H(x) \leq 2\sqrt{x(1-x)}$.*
2. *For any $y \in [0, 1]$, $y \cdot H(x) \leq H(yx)$.*
3. *For any $y \geq 1$, $y \cdot H(x) \geq H(yx)$.*

4. *If* $|x-y| \le \varepsilon$, $|H(x) - H(y)| \le H(\varepsilon)$.

All the above facts essentially follow from concavity of entropy ($H(x/2+y/2) \ge H(x)/2 + H(y)/2$). For detailed proofs see [58].

Proposition 3.2.9 (ℓ_1^2 approximates divergence)**.** *For any* $p, q \in [1/3, 2/3]$*, it holds that*

$$2(p-q)^2 \le \mathbb{D}(p\|q) \le \frac{9}{2} \cdot (p-q)^2.$$

Proof. The left hand side is Pinsker's inequality. To prove the right hand side, we have:

$$\begin{aligned}
&\mathbb{D}(p\|q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} = p\log\frac{q-(q-p)}{q} + (1-p)\log\frac{1-q+(q-p)}{1-q} \\
&= p\log\left(1+\frac{p-q}{q}\right) + (1-p)\log\left(1+\frac{q-p}{1-q}\right) \le p\cdot\frac{p-q}{q} + (1-p)\cdot\frac{q-p}{1-q} \\
&(\text{since } \log(1+x) \le x) \\
&= (p-q)\left(\frac{p}{q} - \frac{1-p}{1-q}\right) = (p-q)\left(\frac{p-pq-q+pq}{q(1-q)}\right) = \frac{(p-q)^2}{q(1-q)} \le \frac{9}{2}\cdot(p-q)^2
\end{aligned}$$

where the last inequality follows from the assumption that $q \in [1/3, 2/3]$, which implies that $q(1-q) \ge 2/9$. □

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

Lemma 3.2.10 (Reverse Pinsker)**.** *Let* $S = \left\{(a,b) : \log\frac{p(a|b)}{q(a|b)} > 1\right\}$*. Then,* $p(S) < 2|p(a,b) - q(a,b)|$.

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

Lemma 3.2.11. *Let* $S = \{a : p(a) < q(a)\}$*. Then,* $\sum_{a\in S} p(a)\log\frac{p(a)}{q(a)} \ge -1/(e\ln 2)$.

Inequalities that Involve Conditioning

The following lemmas upper bound the increase in divergence when extra conditioning is involved. Due to lack of space we omit all proofs. We note that these claims are central to the proof of theorem 3.1.5, and are used in a subtle way, and we encourage the reader to consult [36] for the complete proofs.

Lemma 3.2.12. *Let W be an event and A, B, M be random variables in the probability space p. Then,*

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

Lemma 3.2.13 (Conditioning does not decrease divergence)**.**

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

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 second argument of a divergence expression. The lemma states that, on average, $\mathbb{D}\left(p(a|brW)\|p(a|rW)\right)$ cannot be larger than $\mathbb{D}\left(p(a|brW)\|p(a|r)\right)$. 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 3.2.14. *Let W be an event and A, B, R be random variables. Then,*

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

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

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

The following lemma will be useful in our simulation protocols. It shows that messages sent by each party remain independent of the other party's input even after some part of the input is fixed.

Lemma 3.2.15. *Let* x, y *be inputs to a protocol* π *with public randomness* r *and let* r' *be a variable such that* $\pi(xy|rr') = \pi(x|rr')\pi(y|rr')$*. Let* $m_1, \ldots, m_j$ *be messages in* π *such that* m_j *is transmitted by Alice. Then* $\pi(m_j|m_{<j}rr') = \pi(m_j|m_{<j}rr'y)$*.*

Proof sketch. Conditioned on rr', the variables x, y are independent. Since $m_{<j}$ defines a rectangle over x, y, even conditioned on $m_{<j}rr'$, the variables x, y are independent. Since Alice sends the j'th message, $\pi(m_j|m_{<j}rr'xy) = \pi(m_j|m_{<j}rr'x)$. Thus:

$$\begin{aligned}
\pi(m_j|m_{<j}rr') &= \sum_x \pi(x|m_{<j}rr') \cdot \pi(m_j|m_{<j}rr'x) \\
&= \sum_x \pi(x|m_{<j}rr'y) \cdot \pi(m_j|m_{<j}rr'xy) \\
&= \sum_x \pi(xm_j|m_{<j}rr'y) = \pi(m_j|m_{<j}rr'y).
\end{aligned}$$

□

Useful Protocols

The following lemma was proved by Holenstein [82].

Lemma 3.2.16 (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* $\Pr[A \neq B] = |p - q|$*.*

The following compression theorem from [17] will be useful:

Theorem 3.2.17. *For every protocol π, and every $\epsilon > 0$, there exists a protocol λ that strongly ϵ-simulates π with*

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

Proof of the Main Result – Lemma 3.2.4

In this section we give a more detailed outline of the proof of Lemma 3.2.4, though still leaving out most of the technical proofs. Lemma 3.2.5 is proved in a similar fashion. The formal proofs of all the claims written below can be found in the full version of this paper.

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 $\mathcal{X} = X_1, \ldots, X_n$ and $\mathcal{Y} = Y_1, \ldots, Y_n$. Without loss of generality, we assume that n is even.

Consider the protocol η in Figure 3.1. We show that η computes f with good probability, although with a lot of communication. The protocol η has public randomness $i, \mathbf{g}, \mathbf{h}$ and runs protocol $\theta_{i,\mathbf{g},\mathbf{h}}$ given in Figure 3.2 as a subroutine with inputs $(x_i, r'_{i,\mathbf{g},\mathbf{h}}), (y_i, r''_{i,\mathbf{g},\mathbf{h}})$. Eventually, we shall argue that in expectation over $i, \mathbf{g}, \mathbf{h}$ sampled according to $\eta(i\mathbf{gh})$,

$$\eta\big((x_i, r'_{i,\mathbf{g},\mathbf{h}}), (y_i, r''_{i,\mathbf{g},\mathbf{h}})\big) \overset{O(\gamma)}{\approx}$$
$$\theta_{i,\mathbf{g},\mathbf{h}}((x_i, r'_{i,\mathbf{g},\mathbf{h}}), (y_i, r''_{i,\mathbf{g},\mathbf{h}})),$$

and that, on average, $\theta_{i,\mathbf{g},\mathbf{h}}$ is statistically close to having small internal information, and statistically close to having small external information in the case that μ is product. We shall apply Theorem 3.2.6 + the compression of [17] to compress the communication so as to obtain our final protocol for computing f and conclude the proof of Lemma 3.2.4 (Similarly, an analogues (yet much easier) theorem for external information (Theorem 5 in [36]) + Theorem 3.2.17 are used to obtain the protocol that proves Lemma 3.2.5).

Our first goal is to show that conditioning on the event W does not change the distribution in a typical coordinate. The following lemma is rather standard and follows from subadditivity of divergence and its relation to the ℓ_1 norm (Pinsker's inequality):

Lemma 3.3.18. *In expectation over i sampled according to $\eta(i)$, $\pi(x_i y_i) \overset{\gamma}{\approx} \pi(x_i y_i|W)$.*

Next we eliminate a corner case:

Lemma 3.3.19. *If $\|\pi\| \leq \gamma^2 n$, then in expectation over i sampled according to $\eta(i)$, $\pi(m x_i y_i|W) \overset{\sqrt{2}\gamma}{\approx} \pi(m|W) \cdot \pi(x_i y_i)$.*

The proof of Lemma 3.3.19 is also a straightforward application of subadditivity. Lemma 3.3.19 implies that if $\|\pi\| \leq \gamma^2 n$, then a protocol with 0 communication can approximate the messages of π conditioned on W, and so compute f with 1 additional bit of communication. So

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

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

Given subsets $\mathbf{g}, \mathbf{h} \subset [n]$, let $\mathcal{X}_{\mathbf{h}}$ and $\mathcal{Y}_{\mathbf{g}}$ denote $\mathcal{X}$ and $\mathcal{Y}$ projected on to the relevant coordinates. Define

$$R_{i,\mathbf{g},\mathbf{h}} = \mathcal{X}_{\mathbf{h}\setminus\{i\}}, \mathcal{Y}_{\mathbf{g}\setminus\{i\}}.$$

The random variable $R_{i,\mathbf{g},\mathbf{h}}$ helps to break the dependencies between Alice and Bob.

It turns out that choosing the right distribution for $i, \mathbf{g}, \mathbf{h}$ in η is crucial to our proofs. We need the distribution to be symmetric in $\mathbf{g}, \mathbf{h}$. It is important that $\mathbf{g} \cup \mathbf{h} = [n]$ so that $x_i, y_i, r_{i,\mathbf{g},\mathbf{h}}$ split the dependences between $\overline{x}, \overline{y}$. In the analysis we shall repeatedly use the fact that for every fixing of $\mathbf{h}$, $\eta(i\mathbf{g}|\mathbf{h})$ has the property that i is distributed uniformly over a large set, and $i \in \mathbf{g} \cap \mathbf{h}$. This allows us to apply the chain rule. For more intuition on the choice of the variables $r_{i,\mathbf{g},\mathbf{h}}$, see Section 3.3 in [36].

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

1. Let s_h, s_g be uniformly random numbers from the set $\{n/2+1,\ldots,n\}$. Let $\kappa : [n] \to [n]$ be a uniformly random permutation. Set $\mathbf{h} = \kappa([s_h])$ and $\mathbf{g} = \kappa(\{n - s_g + 1,\ldots,n\})$. Let i be a uniformly random element of $\mathbf{g} \cap \mathbf{h}$ (which must be non-empty by the choice of s_g, s_h).
2. Alice sets $x_i = x$ and Bob sets $y_i = y$.
3. Alice and Bob use Lemma 3.2.16 to sample $r_{i,\mathbf{g},\mathbf{h}}$: Alice uses the distribution $\pi(r_{i,\mathbf{g},\mathbf{h}}|x_i W)$ and Bob uses the distribution $\pi(r_{i,\mathbf{g},\mathbf{h}}|y_i W)$. Write $r'_{i,\mathbf{g},h}$ to denote Alice's sample and $r''_{i,\mathbf{g},\mathbf{h}}$ to denote Bob's sample.
4. Alice and Bob run protocol $\theta_{i,\mathbf{g},\mathbf{h}}$ from Figure 3.2 with inputs $(x_i, r'_{i,\mathbf{g},\mathbf{h}})$ and $(y_i, r''_{i,\mathbf{g},\mathbf{h}})$.

Figure 3.1: Protocol for computing f.

Protocol $\theta_{i,\mathbf{g},\mathbf{h}}$ for computing $f(x_i,y_i)$ when inputs $(x_i, r'_{i,\mathbf{g},\mathbf{h}}), (y_i, r''_{i,\mathbf{g},\mathbf{h}})$ are sampled according to $\pi((x_i, r_{i,\mathbf{g},\mathbf{h}}), (y_i, r_{i,\mathbf{g},\mathbf{h}})|W)$.

Alice sends each message M_j, j odd, according to the distribution $\pi(m_j|x_i r'_{i,\mathbf{g},\mathbf{h}} m_{<j} W)$. Bob sends each message M_j, j even, according to the distribution $\pi(m_j|y_i r''_{i,\mathbf{g},\mathbf{h}} m_{<j} W)$.

Figure 3.2: Simulation in the i'th coordinate.

Now we argue that $\eta(i\mathbf{gh})$ has the properties we need. Observe that we can sample $\eta(i\mathbf{gh})$ by the following different yet equivalent process. Let $\mathbf{h}$ be distributed as in η. For fixed $\mathbf{h}$, let $\kappa_{\mathbf{h}} : [n] \to [n]$ be a permutation sampled uniformly from the set of permutations that map $[|\mathbf{h}|]$ to $\mathbf{h}$. Let ℓ be a uniformly random element of $[n/2]$. Given $\mathbf{h}, \kappa_{\mathbf{h}}, \ell$, set $i = \kappa_{\mathbf{h}}(\ell)$ and $\mathbf{g} = \kappa_{\mathbf{h}}(\{\ell, \ell+1,\ldots,n\})$. Then note that $\mathbf{g}, \mathbf{h}, i$ are distributed as defined in the protocol η. Further, note that $(i, x_i, r_{i,\mathbf{g},\mathbf{h}})$ and $(\kappa_{\mathbf{h}}(\ell), \overline{x}_{\mathbf{h}}, \overline{y}_{\kappa_{\mathbf{h}}(\{\ell+1,\ldots,n\})})$ determine each other.

The following lemma asserts that the distribution of the public randomness $R_{i,\mathbf{g},\mathbf{h}}$ of π doesn't change much when conditioning on W:

Lemma 3.3.20. *In expectation over $i, \mathbf{g}, \mathbf{h}$ sampled according to $\eta(i\mathbf{gh})$,*
$\pi(x_i y_i)\pi(r_{i,\mathbf{g},\mathbf{h}}|x_i W) \overset{3\gamma}{\approx} \pi(x_i y_i r_{i,\mathbf{g},\mathbf{h}}|W) \overset{3\gamma}{\approx} \pi(x_i y_i)\pi(r_{i,\mathbf{g},\mathbf{h}}|y_i W)$.

The following claim is the heart of the proof. It asserts that indeed the distribution $(\pi|W)$, on an average coordinate i, is well approximated by the protocol θ.

Claim 3.3.21. *In expectation over $i, \mathbf{g}, \mathbf{h}$ sampled according to $\eta(i\mathbf{gh})$,*

$$\theta_{i,\mathbf{g},\mathbf{h}}(x_i y_i r_{i,\mathbf{g},\mathbf{h}} m) \overset{2\gamma}{\approx} \pi(x_i y_i r_{i,\mathbf{g},\mathbf{h}} m|W).$$

Proof. Consider

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\eta(i\mathbf{gh})}\left[\mathop{\mathbb{E}}_{\pi(x_i y_i r_{i,\mathbf{g},\mathbf{h}}|W)}\left[\mathbb{D}\left(\pi(m|x_i y_i r_{i,\mathbf{g},\mathbf{h}} W)\|\theta_{i,\mathbf{g},\mathbf{h}}(m|x_i y_i r_{i,\mathbf{g},\mathbf{h}})\right)\right]\right] \\
&= \sum_{j=1}^{2t} \mathbb{E}_{\eta(i\mathbf{gh})}\left[\mathop{\mathbb{E}}_{\pi(m_{<j} x_i y_i r_{i,\mathbf{g},\mathbf{h}}|W)}\left[\mathbb{D}\left(\pi(m_j|x_i y_i r_{i,\mathbf{g},\mathbf{h}} m_{<j} W)\|\theta_{i,\mathbf{g},\mathbf{h}}(m_j|x_i y_i r_{i,\mathbf{g},\mathbf{h}} m_{<j})\right)\right]\right] \qquad (3.6)
\end{aligned}$$

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

$$\sum_{\text{odd } j} \mathop{\mathbb{E}}_{\eta(i\mathbf{gh})}\left[I_\pi(M_j; Y_i|X_i R_{i,\mathbf{g},\mathbf{h}} M_{<j} W)\right].$$

As in the proof of Lemma 3.3.20, we can express this as

$$\frac{2}{n}\sum_{\text{odd } j} \mathbb{E}_{\eta(\mathbf{h}\kappa_\mathbf{h})}\left[I_\pi(M_j; \mathcal{Y}_{\kappa_\mathbf{h}([n/2])}|\mathcal{X}_\mathbf{h}\mathcal{Y}_{\kappa_\mathbf{h}(\{n/2+1,\ldots,n\})} M_{<j} W)\right]$$

by the chain rule. By Lemma 3.2.14, we can upper bound this by

$$\leq \frac{2}{n}\sum_{\text{odd } j} \mathbb{E}_{\eta(\mathbf{h}\kappa_\mathbf{h})}\left[\mathop{\mathbb{E}}_{\pi(m_{<j}\overline{x}_\mathbf{h}\overline{y}|W)}\left[\mathbb{D}\left(\pi(m_j|m_{<j}\overline{x}_\mathbf{h}\overline{y}W)\|\pi(m_j|m_{<j}\overline{x}_\mathbf{h}\overline{y}_{\kappa_\mathbf{h}(\{n/2+1,\ldots,n\})})\right)\right]\right].$$

Conditioned on $\overline{x}_{\mathbf{h}}\overline{y}_{\kappa_{\mathbf{h}}(\{n/2+1,\ldots,n\})}$, the inputs $\overline{x}, \overline{y}$ are independent. Thus Lemma 3.2.15 gives

$$\pi(m_j|m_{<j}\overline{x}_{\mathbf{h}}\overline{y}_{\kappa_{\mathbf{h}}(\{n/2+1,\ldots,n\})}) = \pi(m_j|m_{<j}\overline{x}_{\mathbf{h}}\overline{y}),$$

and we can continue to bound

$$= \frac{2}{n}\sum_{\text{odd } j}\mathbb{E}_{\eta(\mathbf{h}\kappa_{\mathbf{h}})}\left[\mathop{\mathbb{E}}_{\pi(m_{<j}\overline{x}_{\mathbf{h}}\overline{y}|W)}\left[\mathbb{D}\left(\pi(m_j|m_{<j}\overline{x}_{\mathbf{h}}\overline{y}W)\|\pi(m_j|m_{<j}\overline{x}_{\mathbf{h}}\overline{y})\right)\right]\right].$$

Since the divergence is always non-negative, we can add in the even terms in the sum over j to bound

$$\begin{aligned}
&\leq \frac{2}{n}\sum_{j=1}^{2t}\mathbb{E}_{\eta(\mathbf{h}\kappa_{\mathbf{h}})}\left[\mathop{\mathbb{E}}_{\pi(m_{<j}\overline{x}_{\mathbf{h}}\overline{y}|W)}\left[\mathbb{D}\left(\pi(m_j|m_{<j}\overline{x}_{\mathbf{h}}\overline{y}W)\|\pi(m_j|m_{<j}\overline{x}_{\mathbf{h}}\overline{y})\right)\right]\right]\\
&= \frac{2}{n}\mathop{\mathbb{E}}_{\eta(\mathbf{h}\kappa_{\mathbf{h}})}\left[\mathop{\mathbb{E}}_{\pi(\overline{x}_{\mathbf{h}}\overline{y}|W)}\left[\mathbb{D}\left(\pi(m|\overline{x}_{\mathbf{h}}\overline{y}W)\|\pi(m|\overline{x}_{\mathbf{h}}\overline{y})\right)\right]\right] \quad \text{(by the chain rule)}\\
&\leq \frac{2}{n}\mathop{\mathbb{E}}_{\eta(\mathbf{h}\kappa_{\mathbf{h}})}\left[\gamma^2 n\right] = 2\gamma^2,
\end{aligned}$$

by Lemma 3.2.12. Repeating the same argument for even j gives (3.6) $\leq 4\gamma^2$. We apply Lemma 1.1.12 to conclude the proof. □

Completing the Proof of Lemma 3.2.4

Claim 3.3.22. *The expected value of the expression for the internal information cost according to π conditioned on W can be bounded:*

$$\mathbb{E}_{\eta(i\mathbf{gh})}[(I_\pi(X_i; M|Y_i R_{i,\mathbf{g},\mathbf{h}}W) + I_\pi(Y_i; M|X_i R_{i,\mathbf{g},\mathbf{h}}W))] \leq 4\|\pi\|/n.$$

In the probability space of π, let $i, \mathbf{g}, \mathbf{h}$ be independent of all other variables, and distributed as in η. Let $x' = (i, \mathbf{g}, \mathbf{h}, x_i, r_{i,\mathbf{g},\mathbf{h}})$ and $y' = (i, \mathbf{g}, \mathbf{h}, y_i, r_{i,\mathbf{g},\mathbf{h}})$. Define the protocol θ that gets inputs $(i, \mathbf{g}, \mathbf{h}, x_i, r'_{i,\mathbf{g},\mathbf{h}})$ and $(i, \mathbf{g}, \mathbf{h}, y_i, r''_{i,\mathbf{g},\mathbf{h}})$, where the inputs are distributed

according to

$$\pi((i, \mathbf{g}, \mathbf{h}, x_i, r_{i,\mathbf{g},\mathbf{h}}), (i, \mathbf{g}, \mathbf{h}, y_i, r_{i,\mathbf{g},\mathbf{h}})|W),$$

and executes $\theta_{i,\mathbf{g},\mathbf{h}}((x_i, r'_{i,\mathbf{g},\mathbf{h}}), (y_i, r''_{i,\mathbf{g},\mathbf{h}}))$.

By Lemma 3.2.16 and Lemma 3.3.20, $\Pr_\eta[R'_{i,\mathbf{g},\mathbf{h}} \neq R''_{i,\mathbf{g},\mathbf{h}}] \leq O(\gamma)$. Thus in expectation over $i, \mathbf{g}, \mathbf{h}$ sampled according to $\eta(i\mathbf{gh})$,

$$\eta((x_i, r'_{i,\mathbf{g},\mathbf{h}}), (y_i, r''_{i,\mathbf{g},\mathbf{h}})) \overset{O(\gamma)}{\approx} \eta((x_i, r'_{i,\mathbf{g},\mathbf{h}}), (y_i, r'_{i,\mathbf{g},\mathbf{h}})),$$

where here $\eta((x_i, r'_{i,\mathbf{g},\mathbf{h}}), (y_i, r'_{i,\mathbf{g},\mathbf{h}}))$ denotes the distribution where Bob's sample for $r''_{i,\mathbf{g},\mathbf{h}}$ is set to be the same as Alice's sample. By Lemma 3.3.20 and Lemma 3.3.18,

$$\eta(i\mathbf{gh}xyr'_{i,\mathbf{g},\mathbf{h}}) \overset{O(\gamma)}{\approx} \pi(i\mathbf{gh}x_iy_ir_{i,\mathbf{g},\mathbf{h}}|W).$$

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 3.3.21 implies that $\theta(x'y'm) \overset{O(\gamma)}{\approx} \pi(x'y'm|W)$. Claim 3.3.22 implies that

$$\begin{aligned}
&I_\pi(X'; M|Y'W) + I_\pi(Y'; M|X'W) \\
&= \mathop{\mathbb{E}}_{\eta(i\mathbf{gh})} [I_\pi(X_i; M|Y_iR_{i,\mathbf{g},\mathbf{h}}W) + I_\pi(Y_i; M|X_iR_{i,\mathbf{g},\mathbf{h}}W)] \\
&\leq 4\|\pi\|/n \quad \text{(since } \|\pi\| \geq \gamma^2 n\text{)}.
\end{aligned}$$

To prove Lemma 3.2.4, we apply Theorem 3.2.6 to conclude that there exists a protocol that $O(\gamma)$-simulates θ with information cost at most

$$\leq O\left(\frac{(4\|\pi\|/n) + \log(\|\pi\| + 1)}{\gamma^2}\right).$$

We can now compress the resulting protocol using the compression scheme of [17], which for the choice of α in the statement of Lemma 3.2.4, gives a protocol with overall

communication complexity

$$O\left(\frac{\log\|\pi\|\sqrt{(4\|\pi\|/n+1+\log\|\pi\|)\|\pi\|}}{\gamma^{3/2}}\right) < O\left(\frac{\|\pi\|\cdot\log^{3/2}\|\pi\|}{\sqrt{n}\gamma^{5/2}}\right) < C-1.$$

The proof of Lemma 3.2.4 is complete, since with one additional bit of communication to send the value of f, the protocol η computes f with probability of success at least $(1/n)\sum_{i=1}^{n}\pi(W_i|W) - O(\gamma)$. □

3.4 Proof of Theorem 3.1.6 (DP in Terms of IC)

As promised, in this Section we prove Theorems 3.2.6 and 3.1.6, which together complete the proof of Theorem 3.1.6. The main ingredient of the proof is a construction of an "information odometer": An interactive procedure which allows Alice and Bob to keep an *online* estimate of the information of their conversation, and "abort" the protocol when too much information was revealed. Constructing such a primitive is highly non-trivial task, and is the main content of this section. We will see that such a primitive has further applications to direct products, to the interactive compression problem (Subsection 3.5.2) and to privacy (Chapter 7).

Overview of the Odometer Construction

The "information odometer" problem can be put as follows. Alice and Bob are given inputs $x, y \sim \mu$, and are executing a communication protocol π. During the course of the execution of π they wish to maintain an information odometer — an *online* estimate (say, within a factor of 2) of the amount of information they have revealed to each other about their inputs. In more technical terms, they wish to maintain an estimate on the internal information cost of the protocol π so far. Moreover, since applications of this primitive involve limiting the amount of information revealed, they wish to implement it without

revealing too much additional information about their inputs in the process. Ideally, the information overhead of implementing it up to any point in time should scale with the information cost of π so far. In this paper, we introduce a technique that enables such an implementation.

Before discussing applications, let us discuss the challenges in implementing such an information odometer. Firstly, we note that even if the original protocol π does not involve interaction, estimating information revealed requires interaction. Consider the following simple scenario. Alice is given a sequence of blocks $X_1, X_2, \ldots, X_k$ and a subset $S \subset \{1, \ldots, k\}$. Bob is also given a sequence of blocks $Y_1, \ldots, Y_k$ and a subset $T \subset \{1, \ldots, k\}$ for $i \in T$, $X_i = Y_i$, and for $i \notin T$, X_i and Y_i are statistically independent. In the protocol π, Alice performs the following action: For each $i \in [k]$ she sends the block X_i if $i \in S$, and sends a random block R_i otherwise. Thus π is a one-round protocol. The amount of information revealed by π is proportional to $|S \setminus T|$, and the amount of information revealed by the first t blocks is proportional to $a_t := |(S \setminus T) \cap \{1, \ldots, t\}|$. Note that maintaining an estimate on a_t requires the parties to compute $S \setminus T$, which would require Alice and Bob to interact.

The fact that interaction is required means that no "simple" unilateral solution (where Alice and Bob keep some counters separately) is possible, and makes a generic information odometer more difficult to construct. Luckily, while the protocol π can be quite complex, we can always break it down into the individual bits that are being transmitted. Therefore, we can focus on estimating the amount of information transmitted in a single bit sent, say, from Alice to Bob. The distribution of Alice's message M in this case is described by one number $p = \Pr[M = 1 | \text{history}, X = x] \in (0, 1)$, such that her message is given by the Bernoulli distribution B_p: 1 with probability p and 0 with probability $1 - p$. For technical convenience, we will only focus on the case when $p \in (1/3, 2/3)$ — this can be done essentially without loss of generality, since the sending of a highly-biased bit can be simulated by the majority of several, slightly-biased bits, without increasing the infor-

mation cost of the protocol (see Section 6.1 in the full version of this result [40]). Note that the value of the probability p depends on Alice's input x, as well as on the transcript so far. The actual sampling of $M \sim B_p$ is done using Alice's private randomness.

What does Bob learn about x from a message $M \sim B_p$? Not surprisingly, the answer depends on what Bob already knows. More specifically, it is given by the KullbackLeibler divergence between the actual distribution of M, and Bob's belief about this distribution. Note that since $M \in \{0,1\}$ is a binary message, Bob's belief is given by a Bernoulli variable B_q (where $q = \Pr[M = 1|\text{history}, Y = y]$). Since $p \in (1/3, 2/3)$, we must also have $q \in (1/3, 2/3)$. The amount of information learned by Bob is given by $\mathbb{D}(B_p \| B_q)$. For $p, q \in (1/3, 2/3)$ it is the case that $\mathbb{D}(B_p \| B_q) = \Theta((p-q)^2)$. In particular, Bob learns nothing if $q = p$ (i.e. if he already knows p). Therefore, the odometer problem reduces to the task of estimating $I := (p-q)^2$, while revealing not much more than I bits of information to the players in the process. More specifically, we show how to sample a Bernoulli random variable $B_{(p-q)^2}$, while revealing at most $O(H((p-q)^2)) = O((p-q)^2 \log 1/(p-q)^2)$ bits of information. While this quantity is more than $(p-q)^2$ by a $\log 1/(p-q)$ factor, this will be sufficient for most applications. Our test produces an (essentially) unbiased estimator on the amount of information revealed in a given round. By running this estimator on a *subsample* of the rounds, rather than on all the rounds of π, we can keep the overhead below the information cost of π itself, while maintaining a good unbiased estimate of the amount of information revealed so far.

We have therefore reduced the odometer problem to the following scenario. Alice and Bob are given numbers $p \in (1/3, 2/3)$ and $q \in (1/3, 2/3)$, respectively. Their goal is to sample $B_{(p-q)^2}$, while revealing at most $O(H(B_{(p-q)^2}))$ information to each other. The simplest strategy that clearly doesn't work is to have Alice send Bob p and have Bob sample $B_{(p-q)^2}$ (or vice versa). This does not work since p may reveal many bits of information about x (and q may reveal many bits of information about y). A slightly less naïve approach is based on the idea of correlated sampling of [82]. We can sample

a number $Z \in_U [0,1]$ uniformly at random. Alice and Bob then exchange information on whether $p < Z$ and $q < Z$, respectively. If the answers do not match, they output 1, otherwise they output 0. It is not hard to see that this procedure produces a sample from the distribution $B_{|p-q|}$. By repeating it twice and outputing the conjunction of the two answers, we can get a sample from $B_{|p-q|^2}$. Unfortunately, it is not hard to see that this procedure may reveal as much as $\Omega(H(B_{|p-q|})) = \Omega(|p-q| \log 1/|p-q|)$ to the parties, which is prohibitively high.

Our approach is based on the correlated sampling above. Instead of Z being chosen using public randomness, Z is chosen by Alice from a distribution Z_p which depends on the value of p. Alice then sends Z_p to Bob. The distribution Z_p is designed to meet the following two conditions: (1) a sample $Z \sim Z_p$ reveals at most $O((p-q)^2)$ bits of information about p (and thus about x) to someone who knows q; (2) the probability that Z falls between p and q is $\sim (p-q)^2$ (note that for $Z \in_U [0,1]$ this probability was $\sim |p-q|$). Satisfying these two conditions allow us to sample from $B_{(p-q)^2}$ by seeing whether Z falls between p and q (using condition (2)). Condition (1) ensures that the value of Z does not reveal too much information to Bob about x in the process.

As discussed above, we primarily apply this basic primitive as follows. At each step i of π we execute the protocol above with some probability α, obtain a sample $S_i \sim B_{(p-q)^2}$, and maintain the sum Σ_i of the S_i's so far. This way, if I_i^π is the amount of information revealed by π by round i, we have that Σ_i is an unbiased estimator of $\alpha \cdot I_i^\pi$. Therefore Σ_i implements an information odometer for π. While Σ_i is stochastic, by choosing $\alpha < 1$ that is not too small, we can also ensure that Σ_i has sufficiently nice concentration properties for our applications we discuss below.

The Formal Results

We begin by showing how to construct a single-round information odometer. The following sampling lemma serves as the main building block in subsequent applications and constructions in this paper.

Theorem 3.5.1 (One round information odometer)**.** *Let $(p, q) \sim \mathcal{D}$ be two numbers $\in (1/3, 2/3)$, such that $\forall\ q\ \mathbb{E}_{p|q}[p] = q$. Suppose that Alice is given p (not known to Bob), and Bob is given q (not known to Alice). Then there is a a (2-round) protocol τ such that:*

- *(Correctness) At the end of execution, players output "1" w.p exactly $2(p-q)^2$.*
- *(Low information) If $T = T(p, q)$ denotes the transcript of τ, then*

$$\mathsf{IC}(\tau) := \mathbb{E}_{p,q}\left[\mathbb{D}\left((T|p)\|(T|q)\right)\right] \leq 16\mathbb{E}_{p,q}\left[\mathbb{D}\left(p\|q\right)\right] + 2\mathbb{E}_{p,q}\left[H_{2(p-q)^2}\right].$$

Theorem 3.5.1 is the central ingredient in the proofs of Theorems 3.2.6 and 3.1.6, which together establish a strong (essentially tight) direct product theorem for communication complexity in terms of information complexity, as discussed in the introduction of this chapter.

In Section 3.5.2, we discuss the implications of our odometer construction to the interactive compression problem, in light of the recent (exponential) separation result of [71]. We outline a potential strategy for improving state of the art compression results, which uses the odometer to "break" the underlying protocol into low-information pieces ($\sim \log C$), and compress each one separately.

3.5.1 An Interactive Information Odometer

In this section we prove Theorem 3.5.1, the main building block of our information odometer.

Proof of Theorem 3.5.1. The players run the protocol τ from Figure 3.3.

The protocol τ

1. Given her number p, Alice samples a number $Z_p \in [0,1]$, according to the following probability density function:

$$\mu_p(z) = \begin{cases} 4(p-z) & \text{if } 0 \le z < p \\ 4(z-p) & \text{if } p \le z \le p + 1/2 \\ 2 - 4(z-p-1/2) & \text{if } p + 1/2 < z \le 1 \end{cases}$$

   If $p > 1/2$, Alice samples from $\mu_{1-p}(1-z)$.

2. Alice sends Z_p to Bob.
3. Alice sends Bob a bit I^p indicating whether "$Z_p > p$".
4. Bob responds by sending a bit I^q indicating whether "$Z_p > q$".
5. The players output "1" iff $I^p \neq I^q$.

Figure 3.3: A single round information odometer. The probability that the protocol outputs "1" is $2(p-q)^2$.

Analysis of τ: Throughout the entire analysis, we assume that $p \le 1/2$, as it is straightforward to verify that $\mu_p(z) = \mu_{1-p}(1-z)$. First, let us analyze the probability with which the players output "1". Note that the assumption that $p, q \in [1/3, 2/3]$ implies that either $q \in [0,p]$ or $q \in [p, p+1/2]$. If $q \in [p, p+1/2]$, then by construction we have

$$\Pr[\text{players output "1"}] = \Pr[I^p \neq I^q] = \Pr_{\mu_p}[Z \in [p,q]] = \int_p^q \mu(z)dz = \int_p^q 4(z-p)dz =$$

$$= \left[2z^2 - 4pz)\right]_p^q = 2q^2 - 4pq - 2p^2 + 4p^2 = 2(p-q)^2. \tag{3.7}$$

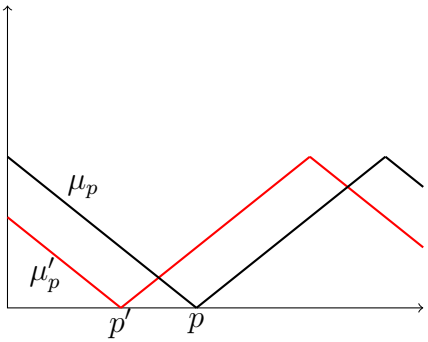

Figure 3.4: The distribution μ_p of Z_p for $p = 0.5, p' = 0.3$. The divergence between μ_p and $\mu_{p'}$ is proportional to $(p-p')^2$. The structure of the density function μ_p ensures that the log-ratio between the distributions mostly cancels out, up to second order terms.

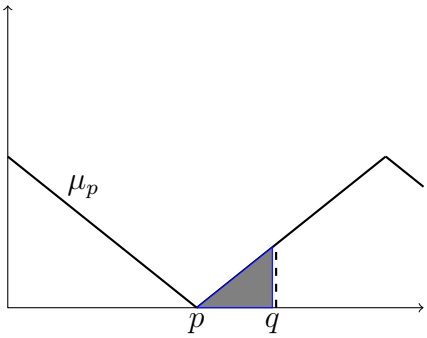

Figure 3.5: The distribution μ_p for $p = 0.5$. For any q, the probability that $p < Z_p < q$ is equal to the area of the triangle enclosing $p, q, \mu_p(q)$.

Similarly, if $q \in [0, p]$, then

$$\Pr[\text{players output "1"}] = \int_q^p \mu(z)dz = \int_p^q 4(p-z)dz = 2(p-q)^2,$$

as claimed in the first proposition of the Theorem.

We turn to analyze the information cost of τ. We analyze step 2 of the protocol and steps 3,4 separately. **Step 2:** The heart of the proof is showing that the information Z_p conveys to Bob (with input q) about Alice's input p, is in fact comparable to the divergence between p and q:

Lemma 3.5.2. $I(Z_p; p|q) \leq 16 \cdot \mathbb{E}_{p,q}\left[\mathbb{D}\left(p \| q\right)\right]$.

The key step is the following technical lemma which asserts that the divergence between the distribution of Z_p and a "shift" of it $Z_{p'}$ is proportional to $(p-p')^2$ (see Figure 2):

Lemma 3.5.3. *For any $p, p' \in (1/3, 2/3)$, it holds that $\mathbb{D}\left(Z_p \| Z_{p'}\right) \leq 8(p-p')^2$.*

This technical lemma is proved by a direct calculation of the divergence, for details see the appendix section of [40]. We now show how Lemma 3.5.3 implies Lemma 3.5.2:

Proof of Lemma 3.5.2.

$$I(Z_p; p|q) = \mathbb{E}_{p,q}\left[\mathbb{D}\left(Z_p \| \mathbb{E}_{p|q}[Z_p]\right)\right] \leq \mathbb{E}_{p,q}\left[\mathbb{D}\left(Z_p \| Z_q\right)\right] \quad \leq 8 \cdot \mathbb{E}_{p,q}[(p-q)^2] \qquad \text{(by Lemma 3.5.3)}$$
$$\leq 16 \cdot \mathbb{E}_{p,q}\left[\mathbb{D}\left(p \| q\right)\right],$$

where the second transition is by Fact 3.2.7 (taken with $T = Z, X = p, Y = q, Z(q) = Z_q$), and the last transition is by Pinsker's inequality (Fact 3.2.9). □

We continue to bound the information of the remaining steps of the protocol τ. **Steps 3 and 4:** Let W denote the indicator random variable of the event "$Z_p \in [p, q]$". Note that at this point, both players already know Z_p, and conditioned on I^p and Z_p, W determines I^q (and vice versa for I^p). Thus, the data processing inequality implies that the information cost of the above steps is upper bounded by

$$\mathbb{E}_{pq}[H(I^p|I^q Z_p) + H(I^q|I^p Z_p)] \leq \mathbb{E}_{pq}[H(W|I^q Z_p) + H(W|I^p Z_p)] \leq 2\mathbb{E}_{pq}[H(W)] = 2\mathbb{E}_{pq}[H(2(p-q)^2)], \tag{3.8}$$

where the last transition is by (3.7). By Lemma 3.5.2 and (3.8), we conclude that

$$\mathsf{IC}(\tau) \leq I(Z_p; p|q) + \mathbb{E}_{pq}\left[H(I^p|I^q Z_p) + H(I^q|I^p Z_p)\right] \leq 16\mathbb{E}_{p,q}\left[\mathbb{D}\left(p \| q\right)\right] + 2\mathbb{E}_{p,q}\left[H(2(p-q)^2)\right],$$

which concludes the second proposition and thus the whole proof of Theorem 3.5.1.

□

Due to space constraints, we leave the formal proofs of Theorems 3.2.6 and 3.1.6 to the full version of the paper [40]. We conclude this chapter by describing the implications of

the information odometer to the interactive compression problem. This is the content of the next Section.

3.5.2 Towards better interactive compression?

In this section we discuss a potential application of the information odometer to the interactive compression question. While we do not prove new compression results, our construction helps clarify the main challenges involved in improving the current state-of-the-art in compressing interactive communication, and suggests a "meta"-approach for making progress on this fascinating problem.

As mentioned in the introduction, the problem of compressing interactive communication can be summarized as follows: "Given a protocol π whose information cost $\mathsf{IC}(\pi, \mu)$ is I and whose communication cost is C, is there an equivalent — compressed — protocol π' that only uses $O(I)$ communication?". Note that if π is non-interactive then the answer to this question is positive [87]. A more modest goal would be to compress π into a protocol π' that uses some function $g(I, C)$ of communication, such as $g(I, C) = \mathrm{p}oly(I) \cdot \mathrm{p}olylog(C)$. The compression question is closely related to the direct sum problem for randomized communication complexity. In fact, these questions are essentially equivalent to each other [35] — the better we can compress, the stronger direct sum holds for communication complexity

The two best general compression results to date are incomparable to each other. The first one, due to [16], gives $g(I, C) = \tilde{O}(C^{1/2} \cdot I^{1/2})$. The second one, due to [28], gives $g(I, C) = 2^{O(I)}$, and was recently shown to be tight in a breakthrough result of Ganor et al. [71]. Note that the second bound becomes non-trivial once $I \ll \log C$. More precisely, the compression scheme of [28] starts with an information-I protocol, and produces a $2^{O(I/\varepsilon)}$-communication protocol while failing with probability ε. Failing with probability ε is inevitable, since I is an average-case quantity, and thus with a small probability ε the

information cost of π will be very high (potentially as high as I/ε) making it impossible to compress in less

than $2^{O(I/\varepsilon)}$-communication with existing techniques. However, one can easily extend the compression result of [28] to show that if we are given a π whose information cost is *uniformly* bounded by I (i.e. with high probability over paths taken by the protocol, the divergence cost is bounded by I), then one can compress π into $2^{O(I)}$ communication while only introducing a negligible amount of additional error:

Claim 3.5.4 (Adapted from [28])**.** *Let $\rho, \varepsilon > 0$ be error parameters, and let π an ε-error protocol for f, such that $\Pr_\mu[\mathbb{D}^\pi_{xyr}(m) > I] \leq \rho$. Then for any distribution μ, $CC^\mu_{\rho+\varepsilon}(f) \leq 2^{O(I)}$.*

Claim 3.5.4 gives rise to the following strategy for compressing a protocol π: (1) partition π into pieces $\pi_1, \pi_2, \ldots$, such that each piece reveals only I_1 bits of information (thus the total number of pieces is $\sim I/I_1$); (2) compress each piece using $2^{O(I_1)}$ communication. Such a plan, if successful, would yield a total communication cost of $O(2^{O(I_1)} \cdot I/I_1)$. If one can make I_1 as small as $O(1)$, this would give a method for interactive compression.

Indeed, this strategy has been successfully carried out in [16] for compressing to *external information cost* of π. The external information cost $\mathsf{IC}^{ext}(\pi)$ of π measures the amount of information $\pi(X, Y)$ reveals about X, Y to an external observer. It is always the case that $\mathsf{IC}^{ext}(\pi) \geq \mathsf{IC}(\pi) = I$, and thus compressing a protocol to $\mathsf{IC}^{ext}(\pi) := I^{ext}$ is easier than compressing it to I. Since step (2) of the strategy is guaranteed by Claim 3.5.4, the main challenge is executing step (1). "Partitioning" means terminating π after $\sim I_1$ information has been revealed. This produces the first piece π_1. Then terminating after another $\sim I_1$ information is revealed produces π_2 etc. In the case of external information Alice can privately estimate the amount of information learnt by an external observed from her messages (since she has the ability to take the external observer's point of view). A similar statement holds about Bob. This allows Alice and Bob to partition the proto-

col into pieces of low ($O(1)$) *external* information cost, thus enabling compression of π to $O(I^{ext} \cdot polylog(C))$ communication.

Is a similar partitioning possible for internal information instead of external? This question is essentially equivalent to the odometer problem studied in this paper. In particular, we can use our odometer construction to pause the protocol π after $O(1)$ bits of information have been communication. Unfortunately, in the process we reveal an additive overhead of $O(\log C)$ bits of information, and thus the resulting information complexity of each part $\pi_1, \pi_2, \ldots$ is $O(\log C)$ rather than $O(1)$. Thus after applying step (2) of the compression plan we get a total communication cost of $I \cdot 2^{O(\log C)}$, which is not better than the original cost C. Unfortunately, [71] asserts it is hopeless to improve the exponential dependence on I in the result of [28], so this the latter approach will not work. Nevertheless, it is still hopeful to use the above approach to improve [16]'s compression result (see the following subsection). This statement can be generalized as follows: Each chunk π_i has information complexity $I_1 = O(\log C)$, and communication complexity $C_1 \leq C$. Therefore, if we could compress π_i into a protocol π_i' of communication complexity $g(I_1, C_1)$, the odometer will imply that any π can be compressed to $O(I \cdot g(I_1, C_1))$ communication. We thus obtain the following claim[6]:

Claim 3.5.5. *Suppose there is a compression protocol that takes as an input a protocol π_1 with communication cost C_1 and* worst case *information cost I_1, and compresses it into a protocol π_1' of communication complexity $g(C_1, I_1)$. Then a protocol π with communication cost C and information cost I can be compressed into a protocol with communication cost $\tilde{O}(I \cdot g(C, \log C))$.*

Claim 3.5.5 implies that it is sufficient to compress protocols whose information cost is logarithmic in their communication cost. In particular, if one could compress a protocol with communication cost C and information cost $\log C$ to a protocol with communication cost $g(C, \log C) = C^{o(1)}$, it would imply that *any* protocol with communication cost C

[6]Since, at this point, this is a qualitative statement, we leave errors out of the statement to avoid complicating the notation.

and information cost I can be compressed to communication $I \cdot C^{o(1)}$. Note that both the scheme from [16] and [28] yield only an upper bound of $g(C, \log C) = C^{O(1)}$ in this case.

Comparison and implications of [71]'s separation result

As argued above, the result in this section shows is that the information odometer reduces the task of interactive compression to the regime where information is only logarithmic in communication ($I = O(\log C)$). Thus, if one could compress a protocol whose information cost is $O(\log C)$ and whose communication cost is $\leq C$ into a protocol which uses $g(C, \log C)$ communication, then one could compress π which uses C communication and I information into a protocol which uses $O(I \cdot g(C, \log C))$ communication in total. In particular, if we could compress into $g(C, \log C) = C^{o(1)}$ bits, then we could compress any π into $I \cdot C^{o(1)}$ communication, which would improve over the current state of the art ($g(C, \log C) = C^{O(1)}$ due to [16]).

The result of Ganor et. al. [71] does not, by any means, rule out such a compression scheme, but only a compression scheme with subexponential dependence on I, if one insists that $g(C, I)$ depends solely on I (and at most sub-logarithmically in C). Therefore, our result (Claim 39) still carries the hope that the odometer may lead to improved interactive compression by focusing one's efforts on the $I = \log C$ regime (notice that in this regime, even a modest compression on the order of $2^{O(\varepsilon) \cdot I} \cdot C^{1/2-\varepsilon}$ would already improve the $C^{1/2}$ dependence in [16]).

Part II

Applications to Other Computational Models

Many computational systems, both in theory and practice, are interactive in nature. Standard real-life examples include client-server interaction for information exchange, distributed computing, and online auctions to mention a few. In theory, the study of interactive systems has led to striking consequences in computational complexity, showing that interaction can be surprisingly powerful: In the context of proof complexity, the celebrated $\mathsf{IP} = \mathsf{PSPACE}$ theorem [142] states that, even problems which will forever remain infeasible to solve from scratch (such as finding a winning strategy in chess), can nevertheless be verified efficiently if we allow the prover and verifier a small number of rounds of interaction (i.e., assuming a player has a winning strategy, he can convince the other in this fact, beyond reasonable doubt, in polynomial time). In algorithmic game theory and mechanism design, it has been shown that multi-round interaction improves the efficiency (social welfare and revenue) of systems, both in pricing mechanism and combinatorial auctions contexts [11, 62] (Result 8 we obtain below improves on the latter work).

The abstractness of the communication complexity model enables it to capture many other computational models. The standard approach is to distribute the input for the (non-interactive) problem into two or more players, and define an appropriate communication game in which the number of bits (or rounds) exchanged correspond to the complexity measure (space, queries, running time etc.) of the original model. This approach for proving lower bounds has been particularly successful in the fields of streaming algorithms (as demonstrated by one of our own results below) and extension complexity [33], and constitutes one of the promising approaches to make progress in the notoriously difficult field of circuit lower bounds (see Section Section 4 below). Other connections between communication complexity and computational complexity have been fruitful. Pătraşcu and Williams [130] showed that a (computationally efficient) protocol with sublinear ($o(n)$) communication for the 3-party Disjointnss problem in

the Number-on-Forehead (NOF) model[7] implies a sub exponential SAT-solver. Later, Pătraşcu showed that the (notoriously difficult) task of obtaining lower bounds in the NOF model would imply strong data structure lower bounds in the cell-probe model, which has been stuck for more than two decades [129].

In the following four chapters, we explore applications of information complexity to the fields of circuit complexity, privacy, streaming and economics. Despite the apparent dissimilarity of these fields of research, the common feature of the problems we study in each of them is that they underly an interactive setup in which information is distributed among multiple agents who are required to solve or optimize some objective function. These agents may be honest, strategic or even adversarial (malicious). The philosophy of this chapter is to explore the role of information and interaction in obtaining efficient solutions in those various interactive systems, where efficiency may be measured in terms of social welfare, revenue, space, privacy or communication, depending on the context.

[7] In this model, each player gets to see every input of the other players, except his own.

Chapter 4

Applications to Circuit Complexity: Towards the KRW Conjecture

One of the holy grails of complexity theory is showing that $\mathbf{NP}$ cannot be computed by polynomial-size circuits, namely, that $\mathbf{NP} \not\subseteq \mathbf{P}/\mathrm{poly}$. Unfortunately, it currently seems that even finding a function in $\mathbf{NP}$ that cannot be computed by circuits of linear size is beyond our reach. Thus, it makes sense to try to prove lower bounds against weaker models of computation, in the hope that such study would eventually lead to lower bounds against general circuits.

This paper focuses on (de-Morgan) formulas, which are one such weaker model. Intuitively, formulas model computations that cannot store intermediate results. Formally, they are circuits with AND, OR, and NOT gates that have fan-out 1, or in other words, their underlying graph is a tree.

For our purposes, it is useful to note that formulas are polynomially related to circuits[1] of depth $O(\log n)$: It is easy to show that circuits of depth $O(\log n)$ can be converted into formulas of polynomially-related size. On the other hand, every formula of size s can be converted into a formula of depth $O(\log s)$ and size $\mathrm{poly}(s)$ [147, 44, 26]. In particular, the

[1] All the circuits in this paper are assumed to have constant fan-in.

complexity class[2] $\mathbf{NC}$ can be defined both as the class of polynomial-size formulas, and as the class of polynomial-size circuits of depth $O(\log n)$.

It is a major open problem to find an explicit function that requires formulas of super-polynomial size, that is, to prove that $\mathbf{P} \not\subseteq \mathbf{NC}$. In fact, even proving that $\mathbf{NEXP} \not\subseteq \mathbf{NC}$ would be a big breakthrough. The state-of-the-art in this direction is the work of Håstad [78], which provided an explicit function whose formula complexity is $n^{3-o(1)}$. Improving over this lower bound is an important challenge.

One strategy for separating $\mathbf{P}$ from $\mathbf{NC}$ was suggested by Karchmer, Raz, and Wigderson [99]. They made a conjecture on the depth complexity of composition, and showed that this conjecture implies that $\mathbf{P} \not\subseteq \mathbf{NC}$. In order to introduce their conjecture, we need some notation:

Definition 4.0.6 (Composition)**.** *Let $f : \{0,1\}^n \to \{0,1\}$ and $g : \{0,1\}^m \to \{0,1\}$ be boolean functions. Their composition $g \circ f : (\{0,1\}^n)^m \to \{0,1\}$ is defined by*

$$(g \circ f)(x_1, \ldots, x_m) \triangleq g(f(x_1), \ldots, f(x_m)),$$

where $x_1, \ldots, x_m \in \{0,1\}^n$.

Definition 4.0.7 (Depth complexity)**.** *Let $f : \{0,1\}^n \to \{0,1\}$. The depth complexity of f, denoted $\mathsf{D}(f)$, is the smallest depth of a circuit of fan-in 2 that computes f using AND, OR and NOT gates.*

Conjecture 4.0.8 (The KRW conjecture [99])**.** *Let $f : \{0,1\}^n \to \{0,1\}$ and $g : \{0,1\}^m \to \{0,1\}$. Then,*

$$\mathsf{D}(g \circ f) \approx \mathsf{D}(g) + \mathsf{D}(f). \tag{4.1}$$

As noted above, [99] then showed that this conjecture could be used to prove that $\mathbf{P} \not\subseteq \mathbf{NC}$: the basic idea is that one could apply $O(\log n)$ compositions of a random func-

[2] In this paper, $\mathbf{NC}$ always denotes the *non-uniform* version of $\mathbf{NC}$, which is sometimes denoted $\mathbf{NC}^1/\mathrm{poly}$.

tion $f : \{0,1\}^{\log n} \to \{0,1\}$, thus obtaining a new function over n bits that is computable in polynomial time yet requires depth $\tilde{\Omega}(\log^2 n)$. The key point here is that a random function on $\log n$ bits has depth complexity $\log n - o(\log n)$, and can be described explicitly using n bits.

In this paper, we solve a natural milestone toward proving the KRW conjecture, using a new information-theoretic approach[3]. We also suggest a candidate for the next milestone, and provide some initial results toward solving it. The rest of this introduction is organized as follows: In Section 4.1, we review the background relevant to our results. In Section 4.2, we describe our main result and our techniques. In Section 4.3, we describe the next milestone candidate, and our initial results in this direction.

4.1 Background: A Communication Complexity Approach to KRW

Karchmer-Wigderson relations

Karchmer and Wigderson [100] observed an interesting connection between depth complexity and communication complexity: for every boolean function f, there exists a corresponding communication problem R_f, such that the depth complexity of f is equal to the deterministic[4] communication complexity of R_f. The communication problem R_f is often called the Karchmer-Wigderson relation of f, and we will refer to it as a KW relation for short.

The communication problem R_f is defined as follows: Alice gets an input $x \in f^{-1}(0)$, and Bob gets as input $y \in f^{-1}(1)$. Clearly, it holds that $x \neq y$. The goal of Alice and Bob

[3]We note that the works [100, 63] on the KRW conjecture also use a (different) information-theoretic argument.

[4]In this paper, we always refer to *deterministic* communication complexity, unless stated explicitly otherwise.

is to find a coordinate i such that $x_i \neq y_i$. Note that there may be more than one possible choice for i, which means that R_f is a relation rather than a function.

This connection between functions and KW relations allows us to study the formula and depth complexity of functions using techniques from communication complexity. In the past, this approach has proved very fruitful in the setting of *monotone* formulas [100, 76, 135, 99], and in particular [99] used it to prove a monotone analogue of the KRW conjecture.

Intuitively, Information Complexity seems a plausible tool for tackling the KRW conjecture, as it involves a "direct-sum" type communication problem, and we already saw in Chapter 3 that this tool has led to (partial) resolution of a similar (but different) composition problem in communication complexity. Indeed, one of the contributions of this work is formalizing this intuition by showing how some of those ideas carry over to the setting of KW relations (see Section Section 4.4).

KW relations and the KRW conjecture

In order to prove the KRW conjecture, one could study the KW relation that corresponds to the composition $g \circ f$. Let us describe how the KW relation $R_{g \circ f}$ looks like. Let $f : \{0,1\}^n \to \{0,1\}$ and $g : \{0,1\}^m \to \{0,1\}$. For every $m \times n$ matrix X, let us denote by $f(X)$ the vector in $\{0,1\}^m$ obtained by applying f to each row of X. In the KW relation $R_{g \circ f}$, Alice and Bob get as inputs $m \times n$ matrices X, Y respectively, such that $f(X) \in g^{-1}(0)$ and $f(Y) \in g^{-1}(1)$, and their goal is to find an entry (j, i) such that $X_{j,i} \neq Y_{j,i}$.

Let us denote the (deterministic) communication complexity of a problem R by $\mathsf{C}(R)$. Clearly, it holds that

$$\mathsf{C}(R_{g \circ f}) \leq \mathsf{C}(R_g) + \mathsf{C}(R_f). \tag{4.2}$$

This upper bound is achieved by the following protocol: For every $j \in [m]$, let X_j denote the j-th row of X, and same for Y. Alice and Bob first use the optimal protocol of g on inputs $f(X)$ and $f(Y)$, and thus find an index $j \in [m]$ such that $f(X_j) \neq f(Y_j)$. Then, they

use the optimal protocol of f on inputs $f(X_j)$ and $f(Y_j)$ to find a coordinate i on which the j-th rows differ, thus obtaining an entry (j, i) on which X and Y differ.

The KRW conjecture says that the above protocol is essentially optimal. One intuition for that conjecture is the following: the best way for Alice and Bob to solve $R_{g \circ f}$ is to solve R_f on some row j such that $f(X_j) \neq f(Y_j)$, since otherwise they are not using the guarantee they have on X and Y. However, in order to do that, they must find such a row j, and to this end they have to solve R_g. Thus, they have to transmit $\mathsf{C}(R_g)$ bits in order to find j, and another $\mathsf{C}(R_f)$ bits to solve f on the j-th row. This intuition was made rigorous in the proof of the monotone version of the KRW conjecture [99], and a similar intuition underly our argument as well as the works of [63, 88] that are to be discussed later.

The universal relation and its composition

Since proving the KRW conjecture seems difficult, [99] suggested studying a simpler problem as a starting point. To describe this simpler problem, we first need to define a communication problem called the universal relation, and its composition with itself. The universal relation R_{U_n} is a communication problem in which Alice and Bob get as inputs $x, y \in \{0,1\}^n$ with the sole guarantee that $x \neq y$, and their goal is to find a coordinate i such that $x_i \neq y_i$. The universal relation R_{U_n} is universal in the sense that every KW relation reduces to it, and indeed, it is not hard to prove that $\mathsf{C}(R_{\mathrm{U}_n}) \geq n$.

The composition of two universal relations R_{U_m} and R_{U_n}, denoted $R_{\mathrm{U}_m \circ \mathrm{U}_n}$, is defined as follows. Alice gets as an input an $m \times n$ matrix X and a string $a \in \{0,1\}^m$, and Bob gets as an input an $m \times n$ matrix Y and a string $b \in \{0,1\}^m$. Their inputs satisfy the following conditions:

1. $a \neq b$.
2. for every $j \in [n]$ such that $a_j \neq b_j$, it holds that $X_j \neq Y_j$.

Their goal, as before, is to find an entry on which X and Y differ. The vectors a and b are analogues of the vectors $f(X)$ and $f(Y)$ in the KW relation $R_{g \circ f}$.

The relation $R_{\mathrm{U}_m \circ \mathrm{U}_n}$ is a universal version of composition problems $R_{g \circ f}$, in the sense that every composition problem $R_{g \circ f}$ reduces to $R_{\mathrm{U}_m \circ \mathrm{U}_n}$. Now, [99] suggested to prove that

$$\mathsf{C}(R_{\mathrm{U}_m \circ \mathrm{U}_n}) \approx \mathsf{C}(R_{\mathrm{U}_m}) + \mathsf{C}(R_{\mathrm{U}_n}) \geq m + n \tag{4.3}$$

as a milestone toward proving the KRW conjecture. This challenge was met by [63] up to a small additive loss, and an alternative proof was given later in [88]. Since then, there has been no further progress on this problem for about two decades.

4.2 Main Result: The composition of a function with the universal relation

Summing up, the KRW conjecture is about the composition of two functions $R_{g \circ f}$, but it was only known how to prove it for the composition of two universal relations $R_{U_m \circ U_n}$. In this work we go a step further: We prove an analogue of the KRW conjecture for relations of the form $R_{g \circ \mathrm{U}_n}$, where $g \in \{0,1\}^m \to \{0,1\}$ is an arbitrary function; and where $R_{g \circ \mathrm{U}_n}$ is a problem that can be naturally viewed as the composition of g with the universal relation.

We define the communication problem $R_{g \circ \mathrm{U}_n}$ as follows. Alice gets as an input an $m \times n$ matrix X and a string $a \in g^{-1}(0)$, and Bob gets as an input an $m \times n$ matrix Y and a string $b \in g^{-1}(1)$. Their inputs are guaranteed to satisfy Condition 2 from above, i.e., for every $j \in [n]$ such that $a_j \neq b_j$, it holds that $X_j \neq Y_j$. Clearly, their inputs also satisfy $a \neq b$, as in Condition 1 above. The goal of Alice and Bob, as usual, is to find an entry on which X and Y differ.

Note that $R_{g \circ U_n}$ is universal, in the sense that for any $f : \{0,1\}^n \to \{0,1\}$, the communication problem $R_{g \circ f}$ reduces to $R_{g \circ \mathrm{U}_n}$. An ideal analogue of the KRW conjecture for

$R_{g \circ U_n}$ would be

$$\mathsf{C}(R_{g \circ U_n}) = \mathsf{C}(R_g) + \mathsf{C}(R_{\mathrm{U}_n}) = \mathsf{C}(R_g) + n. \tag{4.4}$$

We prove the following closely related result. Let $\mathsf{L}(g)$ denote the formula complexity of g, and recall that $\log \mathsf{L}(g) \geq \Omega\left(\mathsf{C}(R_g)\right)$ due to the correspondence between formula size and circuit depth.

Theorem 4.2.1 (Main theorem, [72])**.** *Let $m, n \in \mathcal{N}$ be such that $m \leq 2^{n/2}$, and let $g : \{0,1\}^m \to \{0,1\}$. Then,*

$$\mathsf{C}(R_{g \circ \mathrm{U}_n}) \geq \log \mathsf{L}(g) + n - O(\frac{m \cdot \log m}{n}) \geq \Omega\left(\mathsf{C}(R_g)\right) + n - O(\frac{m \cdot \log m}{n}),$$

Moreover, the same lower bound applies to the logarithm of the number of leaves of any protocol for $R_{g \circ \mathrm{U}_n}$.

There is a good reason why the formula complexity $\mathsf{L}(g)$ appears in Theorem Theorem 4.2.1, as will be made clear in the following discussion on our techniques.

Remark 4.2.2. *In the target application of the KRW conjecture, namely the proof that $\mathbf{P} \not\subseteq \mathbf{NC}$, the parameters can be chosen such that $m \ll n$, so the loss of $O(\frac{m \cdot \log m}{n})$ in Theorem Theorem 4.2.1 is not very important.*

Remark 4.2.3. *We note that Theorem Theorem 4.2.1 also implies lower bound on the composition $R_{U_m \circ U_n}$ of two universal games, thus giving yet another proof for the results of [63, 88].*

Proof Techniques

Our starting point is the simple observation that (the logarithm of) the size of a formula ϕ for any function f can be reinterpreted as the *external* information cost of the corresponding (determinestic) protocol for R_f.

To see why this is helpful, consider the KW relation $R_{g \circ \mathrm{U}_n}$. Intuitively, we would like to argue that in order to solve $R_{g \circ \mathrm{U}_n}$, Alice and Bob must solve R_g (incurring a cost of

$\mathsf{C}(R_g)$), and also solve the universal relation on one of the rows their matrices (incurring a cost of n). Such an argument requires decomposing the communication of Alice and Bob to a communication "about" R_g and a communication "about" R_{U_n}. However, it is not clear how to do that, because Alice and Bob may "talk" simultaneously about R_g and R_{U_n} (e.g. by sending the XOR of a bit of a and a bit of X).

On the other hand, when considering the *information* transmitted by Alice and Bob, such a decomposition comes up naturally: the information that Alice and Bob transmit can be decomposed, using the chain rule, into the information they transmit on the strings a, b (which are inputs of R_g) and the information they transmit on the matrices X and Y (which consist of inputs of R_{U_n}). Of course, implementing this argument is far from trivial, and in particular, we do not know how to extend this argument to the full KRW conjecture, i.e., KW relations of the form $R_{g \circ f}$.

This suggests that information complexity may be the "right" tool to study the KRW conjecture. In particular, since in the setting of KW relations, the information cost is analogous to the formula size, the "correct" way to state the KRW conjecture may be using formula size:

$$\mathsf{L}(g \circ f) \approx \mathsf{L}(g) \cdot \mathsf{L}(f).$$

On hard distributions

One significant difference between our work and previous works on information complexity and direct sum (e.g. [17]) is the following: In order to define the information complexity of a communication problem, one must specify a distribution on the inputs. The reason is information-theoretic notions such as entropy are only defined with respect to a distribution. The previous works use distributions that are protocol independent, that is, they first choose a distribution μ, and then prove that every protocol π for the problem must have a large information cost with respect to μ.

In the setting of KW relations, this is impossible: for every distribution μ there exists a protocol π that has a small information cost with respect to μ (see the full version of this paper [72]). Therefore, the only way to apply information-complexity techniques to KW relations is to use protocol-dependent distributions, that is, to tailor a different distribution for each protocol.

4.3 A next candidate milestone: The composition $\oplus_m \circ f$

In order to make further progress toward the KRW conjecture, we would like to replace the universal relation with a function. One possible approach to this question would be to start with compositions $g \circ f$ where g is a some known simple function. Perhaps the simplest such example is the composition $\vee_m \circ f$, where $\vee_m$ is the disjunction of m bits, and f is an arbirary function. For this example, an analogue of the KRW conjecture is already known, that is, we know that

$$\mathsf{L}(\vee_m \circ f) = \mathsf{L}(\vee_m) \cdot \mathsf{L}(f) = m \cdot \mathsf{L}(f),$$

(see, e.g., [150], and also discussion in Section 3.2.1 in [72]). The next simplest example would be $\oplus_m \circ f$, where $\oplus_m$ is the parity of m bits. For this example, an analogue of the KRW conjecture would be

$$\mathsf{L}(\oplus_m \circ f) \approx \mathsf{L}(\oplus_m) \cdot \mathsf{L}(f) = m^2 \cdot \mathsf{L}(f), \tag{4.5}$$

where the second equality follows from [105]. We therefore suggest the following conjecture as a next milestone toward the KRW conjecture:

Conjecture 4.3.1. *For every function $f : \{0,1\}^n \to \{0,1\}$ and every $m \in \mathcal{N}$, it holds that*

$$\mathsf{L}(\oplus_m \circ f) = \tilde{\Omega}\left(m^2 \cdot \mathsf{L}(f)\right).$$

We note that Conjecture 4.3.1 is not only interesting as a milestone toward the KRW conjecture, but is also interesting on its own right. In particular, if Conjecture 4.3.1 is proved, it will yield an alternative proof of the state-of-the-art lower bound of of $n^{3-o(1)}$ by [78].

In this work, we provide two preliminary results toward proving Conjecture 4.3.1:

- **A lower bound for $R_{\oplus_m \circ U_n}$:** A natural first step toward solving Conjecture 4.3.1 would be to prove a corresponding lower bound on $R_{\oplus_m \circ \mathrm{U}_n}$, the composition of parity with the universal relation. Though in principle we could apply Theorem Theorem 4.2.1 with $g = \oplus_m$, in this case it would not give a meaningful lower bound unless $m \ll n$.. On the other hand, in the target application described above (proving a lower bound of $\tilde{\Omega}(n^3)$), we would like to set $m = 2^n/n$.
  One contribution of this work is proving the following tight analogue of Conjecture 4.3.1 for $R_{\oplus_m \circ \mathrm{U}_n}$:

  Theorem 4.3.2 ([72]). *For every $m, n \in \mathcal{N}$ it holds that*

  $$\mathsf{C}(R_{\oplus_m \circ U_n}) \geq 2 \log m + n - O(\log \log m).$$

  Moreover, the same lower bound applies to the logarithm of the number of leaves of any protocol for $R_{\oplus_m \circ U_n}$.

- **A candidate hard distribution:** We would like to use information-complexity techniques in order to study the KW relation $R_{\oplus_m \circ f}$. In order to define the information complexity of a protocol, we must first define an appropriate distribution over the inputs. We would therefore like to find a "hard distribution" over the inputs of $R_{\oplus_m \circ f}$ that will have a large information cost. As discussed in Section 4.2, this requires tailoring a different hard distribution for each protocol, which is a non-trivial task.

One contribution of this work is suggesting a way to construct a candidate hard distribution for each protocol for $R_{\oplus_m \circ f}$.

4.4 External Information Complexity and Formula Lower Bounds

In this section, we provide a short discussion of the connection between information complexity and formula lower bounds. For a more thorough discussion, the reader is referred to Section 3 of [72] We use the following notation:

Definition 4.4.1. *The size of a formula ϕ, denoted $\mathsf{L}(\phi)$, is the number of the leaves in the underlying tree of ϕ. For a function $f : \{0,1\}^n \to \{0,1\}$, we denote by $\mathsf{L}(f)$ the size of the smallest formula that computes it.*

Definition 4.4.2. *The size of a protocol Π, denoted $\mathsf{L}(\Pi)$, is the number of leaves in the underlying tree of Π. Alternatively, $\mathsf{L}(\Pi)$ is the number of distinct transcripts the protocol may exhibit. For a relation R, we denote by $\mathsf{L}(R)$ the size of the smallest protocol that solves the communication problem R.*

As discussed in the introduction, there is a tight connection between formulas for a function f, and protocols for the KW relation R_f. In particular, it holds that $\mathsf{L}(f) = \mathsf{L}(R_f)$. Fix a function f and a protocol Π for R_f. We can prove formula lower bounds for f by proving lower bounds on $\mathsf{L}(\Pi)$.

Recall that the (external) information cost of Π with respect to μ [17] is

$$\mathsf{IC}^{\mathsf{ext}}{}_{\mu}(\Pi) \triangleq I(\Pi : X, Y),$$

where $I(\Pi : X, Y)$ is the mutual information between Π and (X, Y). Now, we claim that $\mathsf{IC}^{\mathsf{ext}}{}_{\mu}(\Pi)$ gives a lower bound for $\mathsf{L}(\Pi)$. The reason is that

$$I(\Pi : X, Y) \leq H(\Pi) \leq \log \mathsf{L}(\Pi), \tag{4.6}$$

where the second inequality follows since the entropy of a random variable is upper bounded by the logarithm of its support's size. Indeed, in Section 3.1.1 of [72], it is shown that the proofs of [105, 100] of a lower bound for the parity function can be viewed very naturally as lower bounds on $\mathsf{IC}^{\mathsf{ext}}{}_{\mu}(\Pi)$.

On hard distributions. A key point to proving lower bounds on $\mathsf{IC}^{\mathsf{ext}}{}_{\mu}(\Pi)$ is choosing the right distribution μ. For start, observe that for every protocol Π there exists some distribution μ that achieves $\mathsf{IC}^{\mathsf{ext}}{}_{\mu}(\Pi) = \log \mathsf{L}(\Pi)$: this is the distribution that chooses a transcript of Π uniformly at random, and then chooses an arbitrary pair of inputs (x, y) that generates this transcript. We refer to such a distribution μ as a hardest distribution for Π.

A curious feature of this construction of hardest distribution μ is that μ depends on Π. On the other hand, one could hope to construct hard distributions that depend only on the function f, and not on the specific protocol Π. Indeed, all the previous works on information complexity used protocol-independent distributions. Unfortunately, this is not the case: The fact that a $O(\log n)$-information *zero-error randomized* protocol for all KW relations, implies by Yao's minimax theorem that it is impossible to prove formula lower bounds better than $\Omega(n^4)$ using protocol-independent distributions (See proof and further discussion in Section 3.2 of [72]).

It follows that in order to prove formula lower bounds via information complexity, one has to tailor a different hard distribution μ for each protocol Π. One of our contributions in this work is developing tools for performing such a "tailoring". For start, in Section 3.2.1 in [72], we show how a known lower bound on the formula complexity of

the composition $\vee_m \circ f$ can be viewed as a natural way of tailoring hard distributions for protocols for $R_{\vee_m \circ f}$ (here, $\vee_m$ is the boolean OR function over m bits).

In Section 5 of [72], we propose a way for performing such "tailoring" for $R_{\oplus_m \circ f}$. To this end, we develop some generic tools that may be of independent interest. For example, we show that for every function f that is average-case hard, there are hardest distributions μ such that X and Y have large entropy.

Proof of the main result

In this section, we prove our main result, namely, a lower bound on the complexity of the relation $R_{g \circ \mathrm{U}_n}$.

We start by defining $R_{g \circ \mathrm{U}_n}$ formally. Let $g : \{0,1\}^m \to \{0,1\}$, and recall that in the introduction we defined the relation $R_{g \circ \mathrm{U}_n}$ as the following communication problem: Alice gets as an input a matrix $X \in \{0,1\}^{m \times n}$ and a string $a \in g^{-1}(0)$. Bob gets a matrix $Y \in \{0,1\}^{m \times n}$ and a vector $b \in g^{-1}(0)$. Their goal is to find an entry (j,i) on which X and Y differ, and they are promised that for every $j \in [m]$ such that $a_j \neq b_j$, it holds that $X_j \neq Y_j$.

In what follows, we will use a slightly different definition, following [88]: We will allow the players to get inputs that violate the above promise, but in such case, the players are allowed to fail, in which case they should output the special failure symbol $\perp$. As was noted in [88], this modification does not change much the complexity of the communication problem, and simplifies the analysis considerably.

Theorem 4.5.3 ([72]). *[Theorem 4.2.1, main theorem, restated]Let $m, n \in \mathcal{N}$ be such that $m \leq 2^{n/2}$, and let $g : \{0,1\}^m \to \{0,1\}$. Then,*

$$\mathsf{C}(R_{g \circ U}) \geq \log \mathsf{L}(R_{g \circ \mathrm{U}_n}) \geq \log \mathsf{L}(g) + n - O(\frac{m \cdot \log m}{n}).$$

In the rest of this section, we prove Theorem Theorem 4.2.1. We note that only the second inequality requires a proof, whereas the first inequality is trivial since a binary tree of depth c has at most 2^c leaves.

Fix a protocol Π for $R_{g\circ \mathrm{U}_n}$. We prove a lower bound for the number of leaves in the protocol tree of Π. To prove it, we will define a distribution μ on the inputs of the protocol, and analyze the information complexity of Π with respect to μ. More specifically, below we define random variables $\mathbf{X} \in \{0,1\}^{m\times n}$, $\mathbf{a} \in g^{-1}(0)$, $\mathbf{b} \in g^{-1}(1)$, and give Alice and Bob the inputs $(\mathbf{X}, \mathbf{a})$ and $(\mathbf{X}, \mathbf{b})$ respectively. We will prove that

$$\mathsf{IC}^{\mathsf{ext}}{}_{\mu}(\Pi) \geq \log \mathsf{L}(g) + n - O(\frac{m \cdot \log m}{n}).$$

Definition 4.5.4. *Let ℓ be a leaf of Π and let $\mathcal{X}_\ell \times \mathcal{Y}_\ell$ be its corresponding rectangle.*

- *We say that the leaf ℓ supports a matrix $X \in \{0,1\}^{m\times n}$ if X can be given as an input to both players at ℓ. Formally, ℓ supports X if there exist $a, b \in \{0,1\}^m$ such that $(X, a) \in \mathcal{X}_\ell$ and $(X, b) \in \mathcal{Y}_\ell$. We also say that X is supported by ℓ and a, or by ℓ and b. Note that the leaf ℓ must be a leaf that outputs $\perp$.*

- *We say that the leaf ℓ supports $a \in g^{-1}(0)$ if a can be given as an input to Alice at ℓ. Formally, ℓ supports a if there exists a matrix $X \in \{0,1\}^{m\times n}$ such that $(X, a) \in \mathcal{X}_\ell$. A similar definition applies to strings $b \in g^{-1}(1)$.*

Definition 4.5.5. *Let $X \in \{0,1\}^{m\times n}$ be a matrix. Then, the sub-tree of X, denoted T_X, is the sub-tree of Π that consists of the leaves that support X. Note that all those leaves output $\perp$.*

The distribution μ is sampled as follows:

1. Choose a uniformly distributed matrix $\mathbf{X} \in \{0,1\}^{m\times n}$.

2. Choose a uniformly distributed leaf ℓ of $T_{\mathbf{X}}$, and let $\mathcal{X}_\ell \times \mathcal{Y}_\ell$ denote its rectangle.

3. Choose an arbitrary pair $(\mathbf{a}, \mathbf{b})$ such that $(\mathbf{X}, \mathbf{a}) \in \mathcal{X}_\ell$ and $(\mathbf{X}, \mathbf{b}) \in \mathcal{Y}_\ell$.

4. Give Alice the input $(\mathbf{X}, \mathbf{a})$ and to Bob the input $(\mathbf{X}, \mathbf{b})$.

We proceed to analyze the information cost of Π with respect to μ. We begin by applying the chain rule to the information cost:

$$\begin{aligned} \mathsf{IC}^{\mathsf{ext}}{}_{\mu}(\Pi) &\triangleq I(\Pi : \mathbf{X}, \mathbf{a}, \mathbf{b}) \\ &= I(\Pi : \mathbf{X}) + I(\Pi : \mathbf{a}, \mathbf{b} | \mathbf{X}). \end{aligned}$$

As was discussed in the introduction, the second equality can be thought of as a decomposition of the information that the protocol Π gives on $R_{g \circ \mathrm{U}_n}$ (which is $I(\Pi : \mathbf{X}, \mathbf{a}, \mathbf{b})$), into information on the universal game (which is $I(\Pi : \mathbf{X})$), and information on R_g (which is $I(\Pi : \mathbf{a}, \mathbf{b} | \mathbf{X})$). In the following two lemmas, we show that the terms $I(\Pi : \mathbf{X})$ and $I(\Pi : \mathbf{a}, \mathbf{b} | \mathbf{X})$ behave as expected.

Lemma 4.5.6. $I(\Pi : \mathbf{X}) \geq n$.

Lemma 4.5.7. $I(\Pi : \mathbf{a}, \mathbf{b} | \mathbf{X}) \geq \log \mathsf{L}(g) - O(\frac{m \cdot \log m}{n})$.

We prove the lemmas in the following two subsections.

4.6 Proof of the Central Lemmas

In this section we prove the main lemmas required to establish Theorem Theorem 4.2.1.

Proof of Lemma Lemma 4.5.6

We prove that $I(\Pi : \mathbf{X}) \geq n$. As discussed above, the intuition for the lower bound $I(\Pi : \mathbf{X}) \geq n$ is that by the end of the protocol, Alice and Bob must be convinced that their matrices agree on at least one row, and we will show that this requires transmitting

n bits of information. By the definition of mutual information, it holds that

$$\begin{aligned} I(\Pi : \mathbf{X}) &= H(\mathbf{X}) - H(\mathbf{X}|\Pi) \\ &= m \cdot n - H(\mathbf{X}|\Pi). \end{aligned}$$

Thus, it suffices to prove that $H(\mathbf{X}|\Pi) \leq (m-1) \cdot n$. We prove the following stronger claim: for every fixed transcript π in the support of Π, the number of matrices that are supported by π is at most $2^{(m-1)\cdot n}$.

Fix a transcript π, and let $\mathcal{T}$ be the set of matrices X that are supported by π (see Definition 4.5.4). We prove the following claim on $\mathcal{T}$, which is equivalent to saying that Alice and Bob must be convinced that their matrices agree on at least one row.

Claim 4.6.1. *Every two matrices X, X' in $\mathcal{T}$ agree on at least one row.*

Proof. We use a standard "fooling set" argument. Let $\mathcal{X}_\pi \times \mathcal{Y}_\pi$ denote the rectangle that corresponds to π. Suppose, for the sake of contradiction, that there exist $X, X' \in \mathcal{T}$ that do not agree on any row. By definition of $\mathcal{T}$, it follows that there exist strings $a, b \in \{0,1\}^m$ of even and odd weights respectively such that $(X, a) \in \mathcal{X}_\pi$ and $(X', b) \in \mathcal{Y}_\pi$. In particular, this means that if we give Alice and Bob the inputs (X, a) and (X', b) respectively, the resulting transcript of the protocol will be π.

However, this is a contradiction: on the one hand, π is a transcript on which the protocol outputs $\perp$, since it was generated by the distribution μ. On the other hand, the players are not allowed to output $\perp$ on inputs (X, a), (X', b), since X and X' differ on all their rows, and in particular differ on the all the rows j for which $a_j \neq b_j$. The claim follows. □

We conclude the proof of Lemma Lemma 4.5.6 by applying to $\mathcal{T}$ the following combinatorial lemma, which is a corollary of [75, Theorem 1]. For completeness, we provide a proof in the full version of this paper, which rephrases a proof of [4].

Lemma 4.6.2. *Let $\mathcal{S} \subseteq \{0,1\}^{m \times n}$ be a set of matrices such that every two matrices agree on at least one row. Then, $|\mathcal{S}| \leq 2^{(m-1)\cdot n}$.*

Proof of Lemma Lemma 4.5.7

We prove that $I(\Pi : \mathbf{a}, \mathbf{b} | \mathbf{X}) \geq \log \mathsf{L}(g) - O(\frac{m \cdot \log m}{n})$. To this end, we define the notion of "good" and "bad" matrices X, which intuitively are matrices for which the protocol solves R_g and does not solve R_g respectively. We will then show that good matrices contribute have a high information cost, while bad matrices occur with low probability and therefore do not affect the information cost by much.

We start by defining the notion that a string is good for a leaf ℓ. Intuitively, $a \in g^{-1}(0)$ is good for a leaf ℓ if, when the protocol reaches the leaf ℓ and Alice is given the string a, Alice "almost knows" an index j such that $a_j \neq b_j$. More specifically, Alice knows a small set $\mathcal{J} \subseteq [m]$ that contains a coordinate j on which $a_j \neq b_j$.

Definition 4.6.3. *Let ℓ be a leaf of Π, and let $a \in g^{-1}(0)$ be a string supported by ℓ. We say that a is good for ℓ if there exists a set $\mathcal{J} \subseteq [m]$ such that $|\mathcal{J}| < t \triangleq \frac{6m}{n} + 2$, and such that the following holds: for every $b \in g^{-1}(1)$ that is supported by ℓ, there exists $j \in \mathcal{J}$ such that $a_j \neq b_j$.*

If a is not good for ℓ, we say that it is bad for ℓ. Alternatively, we say that a is bad for ℓ if for every set $\mathcal{J}$ of size at most t, there exists $b \in g^{-1}(1)$ that is supported by ℓ such that $a|_{\mathcal{J}} = b|_{\mathcal{J}}$. This definition is inspired by [79].

Definition 4.6.4. *Let ℓ be a leaf of Π and let $\mathcal{X}_\ell \times \mathcal{Y}_\ell$ be its rectangle. A matrix $X \in \{0,1\}^{m \times n}$ is good for ℓ if for every a such that $(X, a) \in \mathcal{X}_\ell$, it holds that a is good for ℓ. Otherwise, we say that X is bad for ℓ. We say that a matrix $X \in \{0,1\}^{m \times n}$ is good if it is good for all the leaves in T_X.*

In the next two subsections, we will prove the following propositions, which together finish the proof of Lemma Lemma 4.5.7.

Proposition 4.6.5. *For every good matrix X, it holds that $I(\Pi : \mathbf{a}, \mathbf{b} | \mathbf{X} = X) \geq \log \mathsf{L}(g) - t \cdot (\log m + 2)$.*

Proposition 4.6.6. *The probability that* $\mathbf{X}$ *is a bad matrix is at most* 2^{-m}.

Proof of Proposition 4.6.5

Let X be a good matrix. We prove that $I(\Pi : \mathbf{a}, \mathbf{b} | \mathbf{X} = X) \geq \log \mathsf{L}(g) - t \cdot (\log m + 2)$. We start by noting that

$$I(\Pi : \mathbf{a}, \mathbf{b} | \mathbf{X} = X) \stackrel{\Delta}{=} H(\Pi | \mathbf{X} = X) - H(\Pi | \mathbf{a}, \mathbf{b}, \mathbf{X} = X) = H(\Pi | \mathbf{X} = X)$$

where the second equality holds since the transcript Π is determined by $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{X}$. Thus, it suffices to lower bound the entropy $H(\Pi | \mathbf{X} = X)$.

Next, observe that by the definition of μ, it holds that conditioned on $\mathbf{X} = X$, the transcript $\Pi = \Pi((\mathbf{X}, \mathbf{a}), (\mathbf{X}, \mathbf{b}))$ is distributed uniformly over the leaves of T_X. Therefore, it suffices to prove that the tree T_X has at least $2^{-t \cdot (\log m + 2)} \cdot \mathsf{L}(g)$ leaves. Let s denote the number of leaves of T_X. We prove the lower bound on s by reduction to R_g: we show that, using T_X, one can construct a protocol Π_X for R_G of size $\mathsf{L}(\Pi_X) \leq 2^{t \cdot (\log m + 2)} \cdot s$, and this would imply the required bound.

The protocol Π_X for R_g is defined as follows: When Alice and Bob get inputs a and b respectively, they invoke the protocol Π on inputs (X, a) and (X, b) respectively, thus reaching a leaf ℓ. Since X is a good matrix, it follows that a is good for ℓ. This implies that there exists a set $\mathcal{J} \subseteq [m]$, $|\mathcal{J}| < t$, such that $a|_{\mathcal{J}} \neq b'|_{\mathcal{J}}$ for every b' that is supported by ℓ. Alice now sends the set $\mathcal{J}$ and the string $a|_{\mathcal{J}}$ to Bob, and Bob replies with $b|_{\mathcal{J}}$. At this point, they both know a coordinate on which a and b differ, and the protocol ends.

The correctness of the protocol Π_X is easy to verify. To analyze its size, observe that after reaching the leaf ℓ, Alice and Bob transmit at most $t \cdot (\log m + 2)$ bits. This implies that the protocol tree of Π_X can be obtained from the tree T_X by replacing each leaf with a binary tree with at most $2^{t \cdot (\log m + 2)}$ leaves. It follows that $\mathsf{L}(\Pi_X) \leq 2^{t \cdot (\log m + 2)} \cdot s$, as required.

Proof of Proposition 4.6.6

We prove that the probability that X is a bad matrix is at most 2^{-m}, or in other words, that there are at most $2^{-m} \cdot 2^{m \cdot n}$ bad matrices. The intuition for the proof is the following: Recall that Alice and Bob output $\perp$, and this means that they have to be convinced that their matrices agree on some row j for which $\mathbf{a}_j \neq \mathbf{b}_j$. However, when X is bad, Alice and Bob do not know an index j such that $\mathbf{a}_j \neq \mathbf{b}_j$ at the end of the protocol. This forces them to agree on many rows, and they can only do so for few matrices. Details follow.

Without loss of generality, we may assume that

$$\mathsf{L}(\Pi) \leq \mathsf{L}(g) \cdot 2^n \leq 2^{m+n},$$

since otherwise Theorem Theorem 4.2.1 would follow immediately. Next, recall that a matrix X is bad if and only if it is supported by a leaf ℓ and a string $a \in g^{-1}(0)$ such that a is bad for ℓ. Therefore, it suffices to prove that every such pair of a leaf ℓ and a string a are "responsible" for at most $2^{-(3 \cdot m+n)} \cdot 2^{m \cdot n}$ bad matrices. This would imply that there are at most $2^{-m} \cdot 2^{m \cdot n}$ bad matrices, by taking union bound over all leaves of Π (at most 2^{m+n}) and all strings a (at most 2^m).

Fix a leaf ℓ of Π and a string $a \in g^{-1}(0)$. Let $\mathcal{T}$ be the set of matrices that are supported by ℓ and a. We prove that $|\mathcal{T}| \leq 2^{-(3 \cdot m+n)} \cdot 2^{m \cdot n}$. Our proof works in two steps: first, we define a combinatorial property of sets of matrices, called the h-agreement property, and show that $\mathcal{T}$ has this property. This property embodies the intuition that Alice and Bob must agree on many rows. Then, we prove a combinatorial lemma that upper bounds the size of sets of matrices that have the h-agreement property.

Definition 4.6.7. *Let $\mathcal{S} \subseteq \{0,1\}^{m \times n}$ be a set of matrices, and let $h \in \mathcal{N}$. We say that $\mathcal{S}$ satisfies the h-agreement property if for every set $\mathcal{J} \subseteq [m]$ such that $|\mathcal{J}| < h$, there exists a matrix $Y^{\mathcal{J}}$, such that every matrix $X \in \mathcal{S}$ agrees with $Y^{\mathcal{J}}$ on a row* outside *$\mathcal{J}$.*

We now show that $\mathcal{T}$ satisfies the h-agreement property for $h = t$. The intuition for this claim is the following: at the leaf ℓ, the protocol outputs $\perp$, and therefore Alice and Bob must be convinced that their matrices agree on at least one row j such that $a_j \neq b_j$. However, Alice does not know an index $j \in [m]$ such that $a_j \neq b_j$, and cannot even isolate it to a small set $\mathcal{J}$ (since a is bad for ℓ). Therefore, for every small set $\mathcal{J}$, Alice must be convinced that her matrix agrees with Bob's matrix on some row outside $\mathcal{J}$.

Claim 4.6.8. *$\mathcal{T}$ satisfies the t-agreement property.*

Proof. Let $\mathcal{J} \subseteq [m]$ be a set such that $|\mathcal{J}| < t$, and let $\mathcal{X}_\ell \times \mathcal{Y}_\ell$ be the rectangle of ℓ. Since a is bad for ℓ, there exists some $b \in g^{-1}(1)$ that is supported by ℓ such that $a|_{\mathcal{J}} = b|_{\mathcal{J}}$. This implies that there exists some matrix Y such that $(Y, b) \in \mathcal{Y}_\ell$. We set $Y^{\mathcal{J}} \triangleq Y$, and prove that every matrix $X \in \mathcal{T}$ agrees with Y on a row outside $\mathcal{J}$.

Let $X \in \mathcal{T}$. This implies that $(X, a) \in \mathcal{X}_\ell$, and therefore, if we give to Alice and Bob the inputs (X, a) and (Y, b) respectively, the protocol will reach the leaf ℓ. Now, recall that ℓ outputs $\perp$, and therefore, there must exist some $j \in [m]$ such that $a_j \neq b_j$ but $X_j = Y_j$. However, we know that $a|_{\mathcal{J}} = b|_{\mathcal{J}}$, and therefore $j \notin \mathcal{J}$. It follows that X and Y agree on a row outside $\mathcal{J}$, as required. □

To upper bound the size of $\mathcal{T}$, we use the following combinatorial lemma, whose proof can be found in Section 2.7 of [72].

Lemma 4.6.9. *Let $\mathcal{S}$ be a set of matrices which satisfies the h-agreement property. Then,*

$$|\mathcal{S}| \leq \frac{m!}{(m-h)!} \cdot 2^{(m-h)\cdot n}.$$

By combining Claim 4.6.8 and Lemma Lemma 4.6.9, it is easy to show that $|\mathcal{T}| \leq \frac{1}{2^{3\cdot m+n}} \cdot 2^{m\cdot n}$, as required.

Chapter 5

Applications to Streaming: Tight Space Bounds for Frequency Moments

In the turnstile model of data streams, an integer vector x is initialized to 0^n and undergoes a long sequence of additive updates to its coordinates. The t-th update in the stream has the form $x_i \leftarrow x_i + \delta_t$, where δ_t is an arbitrary (positive or negative) integer. At the end of the stream we are promised that $x \in \{-M, -M+1, \ldots, M\}^n$ for some bound M which is typically assumed to be at least n (and which we assume here).

Approximating the frequency moments $F_p = \sum_{i=1}^n |x_i|^p$ is one of the most fundamental problems studied in data streams, starting with the seminal work of Alon, Matias, and Szegedy [5]. The goal is to output a number $\hat{F}_p \in [(1-\varepsilon)F_p, (1+\varepsilon)F_p]$ with probability at least $1-\delta$ using as little memory in bits as possible. It is known that for $0 < p \leq 2$, $\Theta(\varepsilon^{-1} \log(M) \log 1/\delta)$ bits of space is necessary and sufficient [98, 97]. Obtaining this optimal bound required a number of ideas, on the upper bound front from p-stable distributions [90] to Nisan's pseudorandom generator [90] to FT-Mollification [98], and on the lower bound front from gap-Hamming [91] to augmented indexing [56] to indexing with low error [97]. These ideas have been the basis of many other streaming algorithms and

lower bounds, with connections to other areas, e.g., linear algebra [146, 55] and information complexity [51, 30].

Perhaps surprisingly, for $p > 2$ a polynomial (in n) amount of space is required [139, 15, 49]. The best known upper bound is due to Ganguly and achieves space

$$O(n^{1-2/p}\epsilon^{-2}\log n \cdot \log(M)\log(1/\delta))/\min(\log n, \epsilon^{4/p-2}))).$$

In the case that $\epsilon \leq 1/\text{poly}(\log n)$, this simplifies to $O(n^{1-2/p}\epsilon^{-2}\log M \log(1/\delta))$. On the other hand, if ϵ is a constant, this simplifies to $O(n^{1-2/p}\log n \log M \log(1/\delta))$. The latter complexity is also achieved by algorithms of [7, 6]. The lower bound, on the other hand, for any ε, δ is only $\Omega(n^{1-2/p}\varepsilon^{-2}\log M)$ [118]. A natural question is whether there are algorithms using less space and achieving a high success probability, that is, if one can do better than just repeating the constant probability data structure and taking a median of $\Theta(\log 1/\delta)$ independent estimates. While there is some work on tightening the bounds in the context of linear sketches over the reals [8, 118], these lower bounds do not yield lower bounds in the streaming setting; for more discussion on this, see below.

Our Results. We show that for any $\varepsilon \in (0, 1)$, any $\delta \geq 2^{-o(n^{1/p})}$, and constant $p > 2$, any algorithm obtaining a $(1+\varepsilon)$-approximation to F_p in the turnstile streaming model requires $\Omega(n^{1-2/p}\varepsilon^{-2}\log M \log(1/\delta))$ bits of space. In light of the upper bounds above, our lower bound is optimal for any $\epsilon \leq 1/\text{poly}(\log n)$. As argued in [118], this is an important regime of parameters. Namely, if $\varepsilon = 1\%$, we have that for, e.g., $n = 2^{32}$, $\varepsilon^{-1} \geq \log n$. Our result is a direct strengthening of the $\Omega(n^{1-2/p}\varepsilon^{-2}\log M)$ lower bound of [118] which cannot be made sensitive to the error probability δ. Moreover, even for constant ϵ, our lower bound of $\Omega(n^{1-2/p}\log M \log(1/\delta))$ bits improves prior work by a $\log(1/\delta)$ factor. We note that for constant ε, the upper bounds still have space $O(n^{1-2/k}\log n \log M \log(1/\delta))$ bits, so while we obtain an improvement, there is still a gap in this case.

While the ultimate goal in this line of research is to obtain tight space bounds simultaneously for any $\varepsilon, \delta \in (0,1)$ and $p > 2$, our result is the first to obtain tight bounds simultaneously in ε and δ for a wide range of parameters. Our proof technique is also quite different than previous work. This seems necessary as the problem considered in [118] has a protocol with information cost $O(n^{1-2/p}\varepsilon^{-2}\log M)$ with 0 error probability, which can be compressed to a protocol with this amount of communication and exponentially small error probability. A description of this protocol, explainaining why the problem considered in [118] does not give stronger lower bounds, can be found in the full version of this paper (see the appendix chapter).

Our Techniques. *A Communication Problem:* Our result is obtained by proving a lower bound for a promise version of k-party set disjointness in the public-coin simultaneous model of communication. In this model there are k players each with a bit string $x^i \in \{0,1\}^n$, $i \in [k] = \{1, 2, \ldots, k\}$, who are promised that their inputs satisfy one of the following cases:

- (NO instance) for all $j \in [n]$, the number of $i \in [k]$ for which $x^i_j = 1$ is distributed as Bin$(k, 1/k)$, or
- (YES instance) there is a unique $j^* \in [n]$ for which $x^i_{j^*} = 1$ for all $i \in [k]$, and for all $j \neq j^*$, the number of $i \in [n]$ for which $x^i_j = 1$ is distributed as Bin$(k, 1/k)$.

The players simultaneously send a message $M^i(x^i, R)$ to a referee, where R is a public-coin that the players share. The referee then outputs a function $f(M^1(x^1, R), \ldots, M^k(x^k, R), R)$, which should equal 1 if the inputs form a YES instance, and equal 0 otherwise. Notice that if $X \sim$ Bin$(k, 1/k)$, then $\Pr[X > \ell] \leq (e/\ell)^\ell$, and so by a union bound for all coordinates j in a NO instance, the number of $i \in [k]$ for which $X^i_j = 1$ is $O(\log n / \log\log n)$. Thus, for $k = \Omega(\log n/\log\log n)$, and in fact $k = n^{\Omega(1)}$ in our reduction, NO and YES instances are distinguishable.

We first show an $\Omega(n \min(\log 1/\delta, \log k)/k)$ total communication lower bound for any protocol which succeeds with probability at least $1 - \delta$ in solving this promise problem in the public-coin simultaneous model of communication (SMP). To do so, we use the information complexity paradigm, first proving a direct sum theorem and then proving a lower bound on the δ-error SMP complexity of a primitive problem (the k-party AND function with the aforementioned promise).

The lower bound for the primitive problem involves lower bounding the information a player's message reveals about his/her input in a NO instance, which is an independent bit with bias $1/k$. To get a handle on this, we ask how many independent messages (over a player's private randomness) the player would need to send for someone to be able to tell if his/her input is 0 or 1. We use the product structure of Hellinger distance to lower bound this quantity, and relate it back to the amount of information a single message of the player reveals via the Maximum Likelihood Estimation principle. This quantity differs from player to player but we again use the product structure of Hellinger distance to show the total information revealed across all players is large.

The proof above does not give a lower bound stronger than $\Omega(n \min(\log 1/\delta, \log k)/k)$ since in this case $O(1)$ players can send their entire input to the referee (assuming $k = n^{\Omega(1)}$), and parts of the argument above require that not all information about a player's input is revealed. To obtain our stronger bound of $\Omega(n \log(1/\delta)/k)$ for any $\delta \geq 2^{-o(n^{1/p})}$, we restrict all players to have the same (randomized) message function, which implies that if one player sends his/her input to the referee, then all k players send their input to the referee, resulting in much higher communication. This same message function restriction turns out to be possible in our reduction, see below.

A Reduction to Streaming: To lower bound the space complexity of a streaming algorithm we need a way of relating it to the communication cost of a protocol for this disjointness problem. We use a recent result of Li, Nguyen, and Woodruff [117] showing

there is a near-optimal streaming algorithm for any problem in the turnstile model which can be implemented by maintaining $A \cdot x$ in the stream, where A is a matrix with poly(n)-bounded integer entries, and A is sampled from a fixed set of $O(n \log m)$ hardwired matrices. In [117] near-optimal meant up to an $O(\log n)$ multiplicative factor in space, which would not suffice here. However, their proof shows if one maintains $A \cdot x \mod q$, where q is a vector of integers one for each coordinate (which depends on A but not on x), then this is optimal up to a constant factor. Notice that this need not be optimal for a *specific family of streams*, such as those arising in our communication game, though we use the fact that by results in [117] an algorithm which succeeds with good probability *for any* fixed stream has this form, and therefore we can assume this form in our reduction. This implies a public-coin simultaneous protocol since the players can use the public coin to choose an (A, q) pair, then each communicate $A \cdot x^i \mod q$ to the referee, who can combine these (using linearity) to obtain $A \cdot (\sum_{i=1}^{k} x^i) \mod q$. This simulation also implies all players have the same message function, even conditioned on the public coin, i.e., it does not depend on the identity of the player.

We stress that the use of a public-coin simultaneous communication model is essential for our result, as there is an $O(n/k)$ total communication upper bound with exponentially small error probability in the one-way communication model (in which player 1 talks to player 2, who talks to player 3, etc., and player k announces the output) for this disjointness problem. The idea is similar to the multi-round 2-player protocol of Håstad and Wigderson [81], in which the players interpret the public coin as a sequence of random subsets of $[n]$ and use it to whittle down their sets until they find the intersection. A proof of this can be found in the full version of this paper (see appendix).

Given this reduction, one of the player's messages must be $\Omega(n \log(1/\delta)/k^2)$ bits long, which lower bounds the space complexity of the streaming algorithm. By setting $k = \varepsilon n^{1/p}$, and by having the referee add $n^{1/p} e_{j*}$ to the stream, where e_{j*} is the standard unit vector in direction j^*, one can show with probability $1 - \delta$, YES and NO instances differ

by a $(1+\varepsilon)$-factor in $F_p(x)$. This is true even given our relaxed definition of disjointness, in which we allow some coordinates to be as large as $\Theta(\log n/\log\log n)$, provided the average of the k-th powers of these coordinates is $\Theta(1)$.

We are not done though, as we seek an extra $\log M$ factor in the lower bound, and for this we superimpose $\Theta(\log M)$ independent copies of this problem at different scales, in a similar way as done for communication problems in previous work [118], and ask the referee to solve a random scaling. There are some technical differences needed to execute this approach in the high $(1-\delta)$ probability regime.

Related Work: We summarize the previous work on this problem in Table 5.1.

F_p Algorithm	Space Complexity
[89]	$O(n^{1-2/p}\epsilon^{-O(1)}\log^{O(1)} n\log(M))$
[24]	$O(n^{1-2/p}\epsilon^{-2-4/p}\log n\log^2(M))$
[126]	$O(n^{1-2/p}\epsilon^{-O(1)}\log^{O(1)} n\log(M))$
[7]	$O(n^{1-2/p}\epsilon^{-2-6/p}\log n\log(M))$
[42]	$O(n^{1-2/p}\epsilon^{-2-4/p}\log n\cdot g(p,n)\log(M))$
[6]	$O(n^{1-2/p}\log n\log(M)\epsilon^{-O(1)})$
[69], **Best upper bound**	$O(n^{1-2/p}\epsilon^{-2}\log n\cdot\log(M)/\min(\log n,\epsilon^{4/p-2}))$
[5]	$\Omega(n^{1-5/p})$
[152]	$\Omega(\epsilon^{-2})$
[15]	$\Omega(n^{1-2/p-\gamma}\epsilon^{-2/p})$, any constant $\gamma>0$
[49]	$\Omega(n^{1-2/p}\epsilon^{-2/p})$
[153]	$\Omega(n^{1-2/p}\epsilon^{-4/p}/\log^{O(1)} n)$
[70]	$\Omega(n^{1-2/p}\epsilon^{-2}/\log n)$
[118]	$\Omega(n^{1-2/p}\epsilon^{-2}\log(M))$

Table 5.1: Results are in bits and for constant $p>2$. The results are stated for constant probability; all results can be made to achieve $1-\delta$ success probability by repeating the data structure independently $O(\log 1/\delta)$ times and taking the median of estimates; this blows up the space by a multiplicative $O(\log 1/\delta)$ factor. Here, $g(p,n)=\min_{c\text{ constant}} g_c(n)$, where $g_1(n)=\log n$, $g_c(n)=\log(g_{c-1}(n))/(1-2/p)$. We start the upper bound timeline with [89], since that is the first work which achieved an exponent of $1-2/p$ for n. For earlier works which achieved worse exponents for n, see [5, 57, 66, 67]. We note that [5] initiated the problem and obtained an $O(n^{1-1/p}\epsilon^{-2}\log(M))$ bound in the insertion-only model (see also [43, 41] for work in the insertion model).

A few papers [9, 132, 118] study the "sketching model" of F_p-estimation in which the underlying vector x is in $\mathbb{R}^n$, rather than in the discrete set $\{-M, -M+1, \ldots, M\}^n$. The goal is to design a distribution over linear maps $A : \mathbb{R}^n \to \mathbb{R}^s$, for some $s \ll n$, so that for any fixed vector $x \in \mathbb{R}^n$, one can $(1+\varepsilon)$-approximate $\|x\|_p^p$ with constant probability by applying an estimation procedure $E : \mathbb{R}^s \to \mathbb{R}$ to Ax. We want the smallest s for a given ε and n. Lower bounds in the sketching model do not imply lower bounds in the turnstile model; this is even true given the recent work [117] characterizing turnstile streaming algorithms as linear sketches. The main issue is that dimension lower bounds in the sketching model are shown for input vectors *over the reals*, while it is conceivable that a linear sketch with fewer dimensions does in fact exist if the input is restricted to be in the integer box $\{-M, -M+1, \ldots, M+1\}^n$. For instance, the inner product of x with the single vector $(1, 1/(M+1), 1/(M+1)^2, \ldots, 1/(M+1)^{n-1})$ is enough to recover x, so a sketching dimension of $s = 1$ suffices. What we are really interested in is a linear sketch with polynomially bounded integer entries, and it is an open question to transport dimension lower bounds in the sketching model to space lower bounds in the turnstile streaming model.

Other related work is that of Jayram and Woodruff [97] which gives lower bounds in terms of δ for F_p for $p \leq 2$. This regime, as mentioned, is fundamentally different and the communication problems there are based on two-player gap-Hamming and Index problems, which have hard product distributions. In contrast we study multi-player communication problems under non-product distributions.

There is also work on direct sums by Molinaro, Woodruff, and Yaroslavtsev [125], which shows that for some problems, solving all n copies of the problem simultaneously with probability $2/3$, is as hard as solving each copy independently with probability $1 - 1/n$. The techniques in that paper do not seem to apply here, since we are interested in solving an OR rather than all copies, and so the output reveals a lot less information about

the inputs. As observed in [46], there is a quite substantial difference in solving the OR versus all copies of a problem.

Braverman and Oshman (private communication) recently obtained an $\Omega(\log k)$ lower bound on the (multi-round) Number-In-Hand communication complexity of the k-party AND function. Of course, this lower bound applies in particular to simultaneous protocols and is much stronger than the one proven in this paper ($\Omega\left(\log(1/\delta)/k\right)$). However, this stronger lower bound holds only for distributions which (prohibitively) violate the promise required for our streaming application, and therefore their lower bound cannot be used to prove our main result.

Preliminaries and Useful Statistical Measures

We henceforth use $[-M, M]^n$ to denote the set $\{-M, \ldots, 0, \ldots, M\}^n$. Since the problem considered in this section is a multi-party communication problem, we will often use vector random variables. Thus, in this section we abuse the notation and reserve bold capital letters for random variables, and use calligraphic letters to sets (for example, $\mathbf{X} \in \mathcal{X}$ represents a random variable with support $\mathbf{X}$). We write $\mathbf{X} \sim B(p)$ to denote a Bernoulli-distributed random variable, taking the value 1 with probability p and 0 with probability $1 - p$.

We use the following distance measures in our arguments.

Definition 5.1.10 (Total Variation distance and Hellinger distance)**.** *The* Total Variation *distance between two probability distributions P, Q over the same universe $\mathcal{U}$ is $\Delta(P, Q) := \sup_A |P(A) - Q(A)|$, where A ranges over all measurable events in the probability space.*

The (squared) Hellinger distance *between P and Q is denoted as*

$$h^2(P,Q) = 1 - \sum_{x\in\mathcal{U}} \sqrt{P(x)Q(x)} = \frac{1}{2}\cdot\sum_{x\in\mathcal{U}} \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2.$$

By a slight abuse of notation, we sometimes use the above distance measures with random variables instead of their underlying distributions. For example, if A, B are two random variables in the joint probability space $p(a,b)$, then $\Delta(A,B) = \Delta(p(a),p(b))$, and $h(A,B) = h(p(a),p(b))$.

The following properties and relationships between the above measures will be used throughout the paper. For missing proofs see [14] and references therein.

Fact 5.1.11 (Product structure of Hellinger distance)**.** *Let $P := P_1,\ldots,P_t$, $Q := Q_1,\ldots,Q_t$ be two product distributions over the same universe, (i.e., $P(x) = \prod_i P_i(x_i)$, $Q(x) = \prod_i Q_i(x_i)$). Then $h^2(P,Q) = 1 - \prod_{i=1}^t (1 - h^2(P_i,Q_i))$.*

Lemma 5.1.12 (Hellinger vs. Total Variation)**.** *For any two distributions P, Q it holds that*

$$h^2(P,Q) \le \Delta(P,Q) \le h(P,Q)\cdot\sqrt{2 - h^2(P,Q)}.$$

Corollary 5.1.13. *If $\Delta(P,Q) \ge 1-\alpha$, then $h^2(P,Q) \ge 1 - 2\sqrt{\alpha}$.*

Proof. Rearranging the RHS inequality of Lemma 5.1.12 and substituting $x := h^2(P,Q), C := \Delta(P,Q)$, we get the following quadratic equation: $x^2 - 2x + C^2 \le 0$. Solving this equation for x yields

$$h^2(P,Q) = x \ge 1 - \sqrt{1-C^2} = 1 - \sqrt{(1+C)(1-C)} \ge 1 - 2\sqrt{1-C} \ge 1 - 2\sqrt{\alpha},$$

where the last two inequalities follow since $1-\alpha \le C = \Delta(P,Q) \le 1$. □

We will need the following fact about the moments of sums of independent random variables (For a proof see [113] Corollary 3).

Lemma 5.1.14 (Moments of sums of independent random variables)**.** *Let $X_1, X_2, \ldots, X_n$ be independent non-negative random variables, and define $X := \sum_{i=1}^{n} X_i$, $\Delta_\ell(X) := \left(\sum_i \mathbb{E}[X_i^\ell]\right)^{1/\ell}$. Then for every $m > 1$,*

$$\left(\mathbb{E}[X^\ell]\right)^{1/\ell} \leq 3e \cdot \frac{m}{\log m} \cdot \max\left\{\Delta_2(X), \Delta_m(X)\right\}.$$

Lemma 5.1.15 (Chebychev inequality for higher moments)**.** *For any $\lambda > 0$ and $m \geq 2$, it holds that $\Pr\left[|X - \mathbb{E}[X]| > \lambda \cdot \sigma_m(X)\right] \leq \frac{1}{\lambda^m}$, where $\sigma_m(X) := \left(\mathbb{E}[|X - \mathbb{E}[X]|^m]\right)^{\frac{1}{m}}$.*

Fact 5.1.16. *If $x \leq \varepsilon$, then $\log(1-x) \geq -\frac{x}{1-\varepsilon}$. (The proof follows by the known inequality $\log(1+y) \geq 1 - 1/y$).*

Fact 5.1.17 (Binary entropy)**.** $\forall p \in [0, 1/2] \;\; , \;\; H(p) \leq p \log(e/p) \leq 2p \log(1/p)$.

Multiparty Communication and Information Complexity in the SMP Model

We use the framework of communication complexity in the Simultaneous Message-Passing model:

Definition 5.1.18 (Multiparty SMP Model)**.** *Let P be a k-ary relation with domain $\mathbf{X}^k := \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_k$ and range $\mathcal{Z}$. In the* SMP *communication model, k parties receive inputs $\mathbf{X}_1, \mathbf{X}_2, \ldots \mathbf{X}_k$, jointly distributed according to some prior distribution μ, and are allowed to share a public random tape R. Each of the players simultaneously sends a message $M_j(\mathbf{X}_j, R)$ to an external party called the referee* , and the referee needs to output an answer $v = v(M_1(\mathbf{X}_1, R), \ldots, M_k(\mathbf{X}_k, R), R)$ such that $v = 1$ iff $(\mathbf{X}_1, \ldots \mathbf{X}_k) \in P$.

The *communication cost* of an SMP protocol π is the sum of the (worst-case) lengths of its messages $\|\pi\| := \sum_{j \in [k]} |M_j|$. For a fixed error parameter $\delta > 0$, the *distributional* SMP

communication complexity of a function f, denoted $\vec{\mathsf{D}}^{\delta}_{\mu}(f)$, is the communication complexity of the cheapest deterministic SMP protocol which computes f correctly with error at most δ under input distribution μ.

The *randomized* SMP *communication complexity* of f, denoted $\mathsf{R}(\delta, f)$, denotes the communication of the cheapest (public-coin) randomized SMP protocol which computes f correctly with error at most δ on *any* input $x \in \mathcal{X}^k$, under the randomness of the protocol (R).

By Yao's minimax theorem, $\mathsf{R}(\delta, f) = \max_{\mu} \vec{\mathsf{D}}^{\delta}_{\mu}(f)$, and therefore it suffices to prove our lower bound for some "hard" distribution μ in the distributional model.

Remark 5.1.19. *To facilitate our proof techniques, we will sometimes need to give the referee an auxiliary input as well. In the distributional model, this input is jointly distributed with the inputs of the k players. The referee's answer is then a function of the k messages he receives as well as his own input. As a convention, in our definitions we typically ignore this artificial feature of the model, and include it implicitly.*

In this section we use the notion of (external) Information Cost (rather than internal). Recall that the *external information cost* of a protocol π with respect to inputs $\mathbf{X}_1, \ldots \mathbf{X}_k \sim \mu$ is defined as

$$\mathsf{IC}^{\mathsf{ext}}_{\pi}(\mu) := I_{\mu}(\pi; \mathbf{X}_1, \ldots \mathbf{X}_k).$$

Again, when the distribution μ is clear from the context, we omit the subscript and simply write $I(\pi; \mathbf{X}_1, \ldots \mathbf{X}_k)$. The *external information complexity of f* under μ is the least amount of information the players need to disclose to the referee about their inputs under μ, if their SMP protocol is required to solve f *on every input* with high probability:

$$\mathsf{IC}^{\mathsf{ext}}_{\delta}(\mu, f) := inf_{\pi:\ \pi \text{ is an SMP protocol and } \forall\ (x_1, \ldots, x_k) \in \mathbf{X}^k\ \ \Pr_{\pi}[\pi(x_1, \ldots, x_k) \neq f(x_1, \ldots, x_k)] \leq \delta}\ \mathsf{IC}^{\mathsf{ext}}_{\pi}(\mu).$$

Remark 5.1.20. *(a) The requirement in the definition above that π is correct everywhere* , i.e., even outside the support of the distribution μ, is crucial: Our lower bounds will rely on

analyzing the information cost of protocols under "trivial" distributions, and the only reason these lower bounds will be meaningful (in particular, non-zero) is that these protocols are required to succeed uniformly. (b) We remark that, unlike communication complexity, the usage of private randomness may be crucial to achieve low information cost, and therefore we assume π is randomized even against a fixed prior distribution μ.

Since one bit of communication can never reveal more than one bit of information, the external information cost of a protocol is always upper bounded by its communication:

Fact 5.1.21. *For any (k-party) communication protocol π and any distribution μ , $\|\pi\| \geq \mathsf{IC}^{\mathsf{ext}}_{\pi}(\mu)$.*

A special class of communication protocols are protocols in which players are restricted to use the same function when sending their messages to the referee. This class will be relevant to our main result.

Definition 5.1.22 (Symmetric SMP protocols)**.** *A k-party* SMP *protocol π is called symmetric* if for any fixed input $\mathbf{X} = x$ and fixing of the public randomness $R = r$,

$$M_1(x, r) = M_2(x, r) = \ldots = M_k(x, r).$$

For a function f, we denote the distributional and randomized communication complexity of f with respect to symmetric SMP protocols by by $\vec{\mathsf{D}}^{\mathsf{SYM},\delta}_{\mu}(f)$ and $\vec{\mathsf{R}}^{\mathsf{SYM}}_{\delta}(f)$. Similarly, we denote by $\mathsf{IC}^{\mathsf{SYM},\delta}_{\mu}(f)$ the (external) information complexity of f with respect to symmetric SMP protocols.

5.2 Multiparty SMP complexity of Set-Disjointness

In this section we prove our lower bound on the SMP communication complexity of the k-party Set-Disjointness function. We obtain the following theorem.

Theorem 5.2.1 (SMP communication complexity of multiparty Set-Disjointness)**.** *For any* $\delta \geq n \cdot 2^{-k}$,

$$\mathsf{R}(\delta, \mathsf{Disj}_k^n) \geq \Omega\left(n \cdot \frac{\min\{\log(1/\delta), \log k\}}{k}\right).$$

$$\vec{\mathsf{R}}_\delta^{\mathsf{SYM}}(\mathsf{Disj}_k^n) \geq \Omega\left(n \cdot \min\left\{\frac{\log(1/\delta)}{k}, \log k\right\}\right).$$

Recall the k-party Set-Disjointness problem is defined as follows:

Definition 5.2.2 (Disj_k^n)**.** *Denote by* Disj_k^n *the multiparty Set-Disjointness problem in which* k *players each receive an* n*-dimensional input vector* $\mathbf{X}_j = \{\mathbf{X}_{j,i}\}_{i=1}^n$ *(where* $\mathbf{X}_{j,i} \in \{0,1\}$*). By the end of the protocol, the referee needs to distinguish between the following cases:*

- **(The "NO" case)** $\forall \quad i \in [n], \quad \sum_j \mathbf{X}_{j,i} < k$, or
- **(The "YES" case)** $\exists \quad i \in [n]$ for which $\sum_j \mathbf{X}_{j,i} = k$.

Denote $\mathsf{AND}_k(x_1, x_2, \ldots, x_k) := \bigwedge_{j=1}^k x_j$*. Note that* $\overline{\mathsf{Disj}_k^n}(\mathbf{X}_1, \ldots, \mathbf{X}_k) = \bigvee_{i=1}^n \mathsf{AND}_k\left(\mathbf{X}_{1,i}, \ldots, \mathbf{X}_{k,i}\right)$.

We start by defining a "hard" distribution for Disj_k^n which still satisfies the promise (gap) required for our streaming application. Consider the distribution η on n-bit string inputs, defined by the following process.

The distribution η:

- For each $i \in [n], j \in [k]$ set $\mathbf{X}_{j,i} \sim B(1/k)$, independently at random.
- Pick a uniformly random coordinate $I \in_R [n]$.
- Pick $Z \in_R \{0,1\}$. If $Z = 1$, set all the values $\mathbf{X}_{j,I}$ to 1, for all $j \in [k]$
  (If $Z = 0$, keep all coordinates as before.).
- The referee receives the index I (this feature will only be used in Section 5.5).

Denote by η_0 the distribution of $\eta \mid$ "$Z = 0$", and by μ_0 the projection of η_0 on a single coordinate (this is well defined since the distribution over all coordinates is i.i.d). In particular, notice that $\eta_0 = \mu_0^n$ is a product distribution, and for every $i \in [n]$, $\Pr_{\mu_0}[\mathbf{X}_{i,j} = 1 \text{ for all } j \in [k]] = (1/k)^k$. Thus, by a union bound over all n coordinates and our assumption on δ,

$$\Pr_{\mu_0^n}[\mathsf{Disj}_k^n(\mathbf{X}_1, \ldots, \mathbf{X}_k)] \leq n \cdot (1/k)^k \leq n \cdot 2^{-k} \leq \delta. \tag{5.1}$$

Direct sum and the SMP complexity of AND_k

To prove Theorem 5.2.1, we first use a direct sum argument, asserting that under product distributions, solving Set Disjointness is essentially equivalent to solving n copies of the 1-bit AND_k function. The following direct sum argument is well known (See e.g., [15]):

Claim 5.2.3 (Direct sum for Disj_k^n)**.** *For any $\delta \geq n \cdot 2^{-k}$, $\mathsf{IC}_\delta(\eta_0, \mathsf{Disj}_k^n) \geq n \cdot \mathsf{IC}^{\mathsf{ext}}_{2\delta}(\mu_0, \mathsf{AND}_k)$.*

We defer the proof of this claim to the full version of the paper (see appendix). With Claim 5.2.3 in hand, it suffices to prove that any (randomized) SMP protocol solving AND_k with error at most δ, must have a large information cost *under* μ_0 . This is the content of the next theorem, which is one of our central technical contributions.

Theorem 5.2.4. *For every $\delta > 0$,*

$$\mathsf{IC}^{\mathsf{ext}}_\delta(\mu_0, \mathsf{AND}_k) \geq \Omega\left(\min\left\{\frac{\log 1/\delta}{k}, \frac{\log k}{k}\right\}\right) \quad \textit{and} \quad \mathsf{IC}^{\mathsf{SYM},\delta}_{\mu_0}(\mathsf{AND}_k) \geq \Omega\left(\min\left\{\frac{\log 1/\delta}{k}, \log k\right\}\right).$$

Proof. Let π be a (randomized) SMP protocol which solves $\mathsf{AND}_k(\mathbf{X}_1, \ldots, \mathbf{X}_k)$ for all inputs in $\{0,1\}^k$ with success probability at least $1 - \delta$. For the rest of the analysis, we fix the public randomness of the protocol. Indeed, proving the lower bound for every fixing of the tape suffices as the chain rule for mutual information implies $\mathsf{IC}^{\mathsf{ext}}_\pi(\mu_0) = \mathbb{E}_R[\mathsf{IC}^{\mathsf{ext}}_{\pi_R}(\mu_0)]$. For each player $j \in [k]$, let M_j denote the transcript of player j's message,

and let $M_0^j := M_j|\text{“}\mathbf{X}_j = 0\text{”}$, $M_1^j := M_j|\text{“}\mathbf{X}_j = 1\text{”}$ (note that if π is further a symmetric protocol, then M_0^j and M_1^j are the same for every player $j \in [k]$). Since the $\mathbf{X}_j$'s are independent under μ_0, and therefore so are the messages M_j, the chain rule implies that $\mathsf{IC}_\pi^{\mathsf{ext}}(\mu_0) = \sum_{j=1}^k I(M_j; \mathbf{X}_j)$. We shall argue that $\sum_{j=1}^k I(M_j; \mathbf{X}_j) \geq \Omega\left(\frac{\log 1/\delta}{k}\ ,\ \frac{\log k}{k}\right)$, and if π is further a symmetric protocol, then $\sum_{j=1}^k I(M_j; \mathbf{X}_j) \geq \Omega\left(\frac{\log 1/\delta}{k}\ ,\ \log k\right)$. To this end, let us denote by

$$h^2(M_1^j, M_0^j) := 1 - z_j$$

the (squared) Hellinger distance between player j's message distributions in both cases. There are two cases: If there is a player j for which $z_j = 0$, then $h^2(M_1^j, M_0^j) = 1$, which means that $I(M_j; \mathbf{X}_j) = H(\mathbf{X}_j) = H(1/k) = \Omega(\log(k)/k)$ and thus $\mathsf{IC}_\delta^{\mathsf{ext}}(\mu_0, \mathsf{AND}_k) \geq \Omega(\log(k)/k)$. Furthermore, if π is symmetric, then $z_1 = z_2 = \ldots = z_j$, which in this case implies by the same reasoning that $I(M_j; \mathbf{X}_j) = \Omega(\log(k)/k)$ for *all* players $j \in [k]$, and thus $\mathsf{IC}_{\mu_0}^{\mathsf{SYM},\delta}(\mathsf{AND}_k) \geq \Omega(\log k)$, as desired.

We may henceforth assume that all z_j's are non-zero, and the rest of the analysis applies for general (not necessarily symmetric) SMP protocols. To this end, let us introduce one final notation: For a *fixed* input $\mathbf{X}_j$, let $M_j^{\oplus t}$ denote (the concatenation of) t *independent* copies of $M_j|\mathbf{X}_j$ (so $M_j^{\oplus t} = (M_0^j)^t$ whenever $\mathbf{X}_j = 0$ and $M_j^{\oplus t} = (M_1^j)^t$ whenever $\mathbf{X}_j = 1$). By the conditional independence of the t copies of M_j (conditioned on $\mathbf{X}_j$) and the product structure of the Hellinger distance (Fact 5.1.11), we have that for each $j \in [k]$, the total variation distance between the t-fold message copies in the “YES” and “NO” cases is at least

$$\Delta\Big((M_1^j)^t, (M_0^j)^t\Big) \geq h^2\Big((M_1^j)^t, (M_0^j)^t\Big) = 1 - (z_j)^t, \tag{5.2}$$

where the first inequality follows from Lemma 5.1.12. Set $t_j = O(\log k / \log(1/z_j))$ (note that this is well defined as we assumed $z_j \neq 0$). Thus, for each player $j \in [k]$,

$$\Delta\left((M_1^j)^{t_j}, (M_0^j)^t\right) \geq 1 - \frac{1}{10k}. \tag{5.3}$$

Equation (5.3) implies that the error probability of the MLE predictor[1] for predicting $\mathbf{X}_j$ given $M_j^{\oplus t_j}$ is at most $\varepsilon := 1/(10k)$. Therefore, Fano's inequality (Lemma 1.1.19) and the data processing inequality together imply that

$$\forall\ j \in [k], \quad I(M_j^{\oplus t_j}; \mathbf{X}_j) \geq H(\mathbf{X}_j) - H(\varepsilon) \geq H\left(\frac{1}{k}\right) - H\left(\frac{1}{10k}\right) \geq \Omega\left(\frac{\log k}{k}\right), \tag{5.4}$$

since $\mathbf{X}_j \sim B(1/k)$ under μ_0, and $H(1/(10k)) \leq \frac{2}{10k}\log(10k) \leq \frac{4}{5}k\log(k)$ by Fact 5.1.17.

Now, by the chain rule for mutual information (Fact 1.1.16) we know that

$$I(M_j^{\oplus t_j}; \mathbf{X}_j) = \sum_{s=1}^{t_j} I((M_j)_s; \mathbf{X}_j | (M_j)_{<s}) \leq \sum_{s=1}^{t_j} I((M_j)_s; \mathbf{X}_j), \tag{5.5}$$

where the last inequality follows from Fact 1.1.14, as the messages $(M_j)_s$ and $(M_j)_{<s}$ are independent conditioned on $\mathbf{X}_i$ (by construction). Notice that $(M_j)_s \sim M_j$ for all $s \in [t]$, as all the messages are equally distributed conditioned on $\mathbf{X}_j$). Combining equations (5.4) and (5.5) therefore implies

$$I(M_j; \mathbf{X}_j) \geq \Omega\left(\frac{\log k}{k \cdot t_j}\right) \geq \Omega\left(\frac{\log(1/z_j)}{k}\right), \tag{5.6}$$

recalling that $t_j = O(\log k / \log(1/z_j))$. Since (5.6) holds for any player $j \in [k]$, we have

$$\sum_{j=1}^{k} I(M_j; \mathbf{X}_j) \geq \Omega\left(\frac{1}{k} \cdot \sum_{j=1}^{k} \log\left(\frac{1}{z_j}\right)\right). \tag{5.7}$$

[1]That is, the predictor which given $M_j^{\oplus t} = m$, outputs $Y := argmax_{x \in \{0,1\}} \Pr[(M_x^j)^t = m]$.

We finish the proof by showing that

$$\sum_{j=1}^{k} \log\left(\frac{1}{z_j}\right) \geq \Omega(\log(1/\delta)). \tag{5.8}$$

To this end, we first claim that the correctness of π implies that the total variation distance between the transcript distributions of π on the input 0^k and on the input 1^k must be large (notice that below we crucially use the fact that our information complexity definition requires the protocol to be correct on all inputs, so in particular, a δ-error protocol must distinguish with comparable error, between "YES" and "NO" inputs):

Proposition 5.2.5. $\Delta(\pi(0^k), \pi(1^k)) \geq 1 - 2\delta$.

Proof. Let $\mathcal{Y}$ be the set of transcripts τ for which $\pi(\tau) = \mathsf{AND}_k(1^k) = 1$. By the correctness assumption, $\Pr[\pi(1^k) \in \mathcal{Y}] \geq 1 - \delta$, and $\Pr[\pi(0^k) \in \mathcal{Y}] \leq \delta$, so the above follows by definition of the total variation distance. □

Since μ_0 is a product distribution (the $\mathbf{X}_j$'s are i.i.d), it holds that $\pi(0^k) = \times_{j=1}^{k} M_0^j$, and $\pi(1^k) = \times_{j=1}^{k} M_0^j$. Therefore, recalling that $z_j := 1 - h^2(M_0^j, M_1^j)$, the product structure of the Hellinger distance (Fact 5.1.11) implies

$$1 - \Pi_{j=1}^{k}\, z_j = 1 - \Pi_{j=1}^{k}(1 - h^2(M_0^j, M_1^j)) = h^2(\pi(0^k), \pi(1^k)) \geq 1 - 4\sqrt{\delta} \tag{5.9}$$

where the last transition follows from the combination of Proposition 5.2.5 with Corollary 5.1.13 (taken with $\alpha = 2\delta$). Rearranging (5.9), we get $\Pi_{j=1}^{k}\, z_j \leq 4\sqrt{\delta}$, or equivalently,

$$\sum_{j=1}^{k} \log\left(\frac{1}{z_j}\right) \geq \frac{1}{2}\log\left(\frac{1}{\delta}\right) - 2 = \Omega\left(\log 1/\delta\right), \tag{5.10}$$

as desired. Combining equations (5.8) and (5.7), we conclude that $\mathsf{IC}_{\pi}^{\mathsf{ext}}(\mu_0) \geq \Omega\left(\frac{\log 1/\delta}{k}\right)$, which completes the proof of Theorem 5.2.4.

□

Since communication is always lower bounded by information (Fact 5.1.21), combining Theorem 5.2.4 and Claim 5.2.3 directly implies Theorem 5.2.1:

Corollary 5.2.6. *For any $\delta \geq n \cdot 2^{-k}$,*

$$\mathsf{R}(\delta, \mathsf{Disj}_k^n) \geq \Omega\left(n \cdot \min\left\{\frac{\log 1/\delta}{k}, \frac{\log k}{k}\right\}\right) \quad \textit{and} \quad \vec{\mathsf{R}}_\delta^{\mathsf{SYM}}(\mathsf{Disj}_k^n) \geq \Omega\left(n \cdot \min\left\{\frac{\log 1/\delta}{k}, \log k\right\}\right).$$

Path-independent stream automata [117]

As mentioned in the introduction, a central fact which facilitates our lower bound is the recent result of [117], asserting that in the turnstile streaming model, linear sketching algorithms achieve optimal space complexity, up to a logarithmic factor. Since we cannot even afford losing a $\log n$ factor in our lower bound, we use the following intermediate result of [117], which shows that oblivious streaming algorithms are optimal up to a *constant* factor. The following exposition largely follows that of [117], from which a number of definitions also occur in the earlier work of [68].

The work of [117] considers problems in which the input is a vector $x \in \mathcal{Z}^n$ represented as a data stream $\sigma = (\sigma_1, \sigma_2, \dots)$ in which each element σ_i belongs to $\Sigma = \{e_1, \dots, e_n, -e_i, \dots, -e_n\}$ (where the e_i's are canonical basis vectors) such that $\sum_i \sigma_i = x$. We write $x = \operatorname{freq} \sigma$.

Definition 5.3.7 (Deterministic stream automata)**.** *A deterministic stream automaton $\mathcal{A}$ is a deterministic Turing machine that uses two tapes, a one-way (unidirectional) read-only input tape and a (bidirectional) two way work-tape. The input tape contains the input stream σ. After processing its input, the automaton writes an output, denoted by $\phi_{\mathcal{A}}(\sigma)$, on the work-tape.*

A configuration of a stream automaton $\mathcal{A}$ is modeled as a triple (q, h, w), where, q is a state of the finite control, h the current head position of the work-tape and w the content of the work-tape. The set of configurations of a stream automaton $\mathcal{A}$ that are reachable from the initial configuration o on some input stream is denoted by $C(\mathcal{A})$. A stream

automaton is a tuple $(n, C, o, \oplus, \phi)$, where n specifies the dimension of the underlying vector, $\oplus : C \times \Sigma \to C$ is the configuration transition function, o is the initial position of the automaton and $\phi : C \to \mathcal{Z}^{p(n)}$ is the output function and $p(n)$ is the dimension of the output. For a stream σ we also write $\phi(o \oplus \sigma)$ as $\phi(\sigma)$ for simplicity.

The set of configurations of an automaton $\mathcal{A}$ that is reachable from the origin o for some input stream σ with $\|\operatorname{freq}\sigma\|_\infty \leq m$ is denoted by $C(\mathcal{A}, m)$. The space of the automaton $\mathcal{A}$ with stream parameter m is defined as $S(A, m) = \log|C(\mathcal{A}, m)|$. An algorithm is said to be a correct randomized algorithm with error probability δ if for any fixed stream σ with $\|\operatorname{freq}\sigma\|_\infty \leq m$, with probability at least $1 - \delta$ the algorithm outputs the correct answer to a relation P for the underlying vector x represented by σ. Note that the streaming algorithm should be correct even if for a substream σ' of σ we have $\|\operatorname{freq}\sigma'\|_\infty > m$, provided that $\|\operatorname{freq}\sigma\|_\infty \leq m$. In this case we say $\mathcal{A}$ solves P on $\mathcal{Z}^n_{|m|}$.

Definition 5.3.8 (Path-independent stream automata)**.** *A stream automaton $\mathcal{A}$ is said to be* path independent *(*PIA*) if for each configuration s and input stream σ, $s \oplus \sigma$ is dependent only on $\operatorname{freq}\sigma$ and s.*

Suppose that $\mathcal{A}$ is a path independent automaton. We can define a function $+ : \mathcal{Z}^n \times C \to C$ as $x + a = a \oplus \sigma$, where $\operatorname{freq}\sigma = x$. Since $\mathcal{A}$ is a path independent automaton, the function $+$ is well-defined. In [68] it is proved that

Theorem 5.3.9. *Suppose that $\mathcal{A}$ is a path independent automaton with initial configuration o. Let $M = \{x \in \mathcal{Z}^n : x + o = 0 + o\}$, then M is a submodule of $\mathcal{Z}^n$, and the mapping $x + M \mapsto x + o$ is a set isomorphism between $\mathcal{Z}^n/M$ and the set of reachable configurations $\{x + o : x \in \mathcal{Z}^n\}$.*

Definition 5.3.10 (Randomized stream automata)**.** *A randomized stream automaton is a deterministic stream automaton with one additional tape for the random bits. The random bit string R is initialized on the random bit tape before any input record is read; thereafter the random bit string is used in a two way read-only manner. The rest of the execution proceeds as in a deterministic stream automaton.*

A randomized stream automaton $\mathcal{A}$ is said to be *path-independent* if for each randomness R the deterministic instance $\mathcal{A}_R$ is path-independent. The space complexity of $\mathcal{A}$ is defined to be

$$S(\mathcal{A}, m) = \max_R \{|R| + S(\mathcal{A}_R, m)\}.$$

Theorem 5.3.11 ([117] Theorems 9 and 10)**.** *Suppose that a randomized algorithm $\mathcal{A}$ solves a relation P on any stream σ with probability at least $1 - \delta$. There exists a randomized path-independent automaton (*PIA *) $\mathcal{B}$ which solves P on $\mathcal{Z}^n_{|m|}$ with probability at least $1 - 7\delta$ such that $S(\mathcal{B}, m) \leq S(\mathcal{A}, m) + O(\log n + \log\log m + \log \frac{1}{\delta})$. Further, the number of random bits used by the algorithm is is $O(\log 1/\delta + \log n + \log\log m)$.*

Here we record the corollary of Theorem 5.3.11 that will be used in the proof of our main result (Theorem 5.5.1). To this end, we will need the following (refined) restatement of the SMP communication model used in our paper:

Definition 5.3.12. *Let $P(x_1, \ldots, x_k)$ be a k-ary relation. In the public-coin* SMP *communication model, k players receive inputs $x_1, \ldots, x_k \in \mathcal{Z}^n$ respectively, such that $x := \sum_j x_j \in Z^n_{|m|}$ (for some $m \in \mathbb{N}$). The players share a public random tape R of $O(\log 1/\delta + \log n + \log\log m)$ uniformly random bits. Each of the players simultaneously sends a message $M_j(x_j, R)$ to an external party called the referee ,* and the referee outputs an answer $v = v(M_1(x_1, R), \ldots, M_k(x_k, R), R)$, such that $\Pr_R[v = P(x_1, \ldots, x_k)] \geq 1 - \delta$. Recall that the *symmetric* SMP communication complexity of P is $\vec{\mathsf{R}}^{\mathsf{SYM}}_\delta(P) := \min_{\pi \,:\, \pi \text{ is symmetric and } \delta\text{-solves } P} \sum_{j=1}^k |M_j(x_j, R)|$ where $|\cdot|$ denotes the worst-case length of the messages, over all choices of $x^1, \ldots, x^s$ and R.

Corollary 5.3.13. *Let $P(x_1, \ldots, x_k)$ be a relation such that $\vec{\mathsf{R}}^{\mathsf{SYM}}_\delta(P) = c$. Let $\mathcal{A}$ be a space-optimal streaming algorithm in the turnstile model from which the output of $\mathcal{A}$ on an input stream σ with underlying vector x, can be used to solve P with probability at least $1 - \delta$. Then the space complexity of $\mathcal{A}$ is at least c/k.*

Proof. By Theorem 5.3.11, we can assume that $\mathcal{A}$ is a randomized path-independent automaton using $O(\log 1/\delta + \log n + \log\log m)$ random bits. The players in the public-coin simultaneous model of communication can therefore use the public coin R to agree upon a deterministic path-independent automaton $\mathcal{B}$. Each player can run $\mathcal{B}$ on his/her local input vector x_j, and transmit the state of $\mathcal{B}$ to the referee. Notice that each player uses the same function to compute his message, and therefore this SMP protocol is also symmetric. By Theorem 5.3.9, the referee can associate these states with elements of the quotient group Z^n/M, where M is determined from the description of $\mathcal{B}$ (which is in turn determined by R), and perform arithmetic in Z^n/M to add up the states to obtain the result of the execution of $\mathcal{B}$ on the concatenation of streams $\sigma^1, \ldots, \sigma^k$, where σ^j is a stream generating x_j. It follows that $\mathcal{B}$ will be executed on σ with underlying vector x, and by hypothesis can be used to solve P with probability at least $1 - \delta$. As k times the space complexity of $\mathcal{B}$ is the communication cost, the corollary follows. □

5.4 The Augmented Disj_k^n problem

In this section we define the multiparty communication problem which we use as a proxy for our main lower bound. This communication problem is constructed using a fairly standard hardness-amplification technique ("augmentation") of the k-party Disjointness problem, in a similar fashion to the work of [118] (who used this technique for the L_∞ communication problem).

Definition 5.4.1 (Aug-Disj(r, k, δ))**.** Aug-Disj(r, k, δ) *is the following k-party communication problem: The players receive r instances of* Disj_k^n*:*

$$(\mathbf{X}_1^1, \ldots, \mathbf{X}_k^1), (\mathbf{X}_1^2, \ldots, \mathbf{X}_k^2), \ldots, (\mathbf{X}_1^r, \ldots, \mathbf{X}_k^r)$$

In addition, the referee receives an index $T \in [r]$ which is unknown to the players, along with the last $(r-T)$ inputs $\{(\mathbf{X}_1^\ell, \ldots, \mathbf{X}_k^\ell)\}_{\ell=T+1}^r$. By the end of the protocol, the referee should output the answer to the T'th instance, i.e, the players need to solve $\mathsf{Disj}_k^n(\mathbf{X}_1^T, \ldots, \mathbf{X}_k^T)$ with probability $1-\delta$.

For convenience, we henceforth denote $(\mathbf{X}_1^{>t}, \ldots, \mathbf{X}_k^{>t}) := \{(\mathbf{X}_1^\ell, \ldots, \mathbf{X}_k^\ell)\}_{\ell=t+1}^r$, and $(\mathbf{X}_1^{<t}, \ldots, \mathbf{X}_k^{<t})$, $\left(\mathbf{X}_1^{\{-t\}}, \ldots, \mathbf{X}_k^{\{-t\}}\right)$ are defined analogously.

We now define a "hard" distribution ν for $\mathsf{Aug\text{-}Disj}(r,k,\delta)$. To this end, recall the distributions η, η_0 for Disj_k^n from Section 5.2. The index $T \in [r]$ is chosen independently and uniformly at random. All copies but the T'th copy are independently chosen according to the "NO" distribution for Disj_k^n, i.e., $\left(\mathbf{X}_1^{\{-T\}}, \ldots, \mathbf{X}_k^{\{-T\}}\right) \sim \eta_0^{r-1}$, while $(\mathbf{X}_1^T, \ldots, \mathbf{X}_k^T) \sim \eta$. The next lemma asserts that the the r-augmented Disjointness problem under the distribution ν is r times harder than solving a single instance Disj_k^n under η.

Lemma 5.4.2 (Direct Sum for $\mathsf{Aug\text{-}Disj}(r,k,\delta)$)**.**

$$\mathsf{R}(\delta, \mathsf{Aug\text{-}Disj}(r,k,\delta)) \geq r \cdot \mathsf{IC}_\delta^{\mathsf{ext}}(\eta_0, \mathsf{Disj}_k^n).$$

Proof. The proof is essentially the same as that of Claim 5.2.3, using a standard "embedding" argument: Let Π be a protocol for $\mathsf{Aug\text{-}Disj}(r,k,\delta)$ under ν, such that $\|\Pi\| = \mathsf{R}(\delta, \mathsf{Aug\text{-}Disj}(r,k,\delta))$. The k players will use public randomness to sample a random $t \in_R [r]$ along with $(r-t)$ "dummy" inputs $(\mathbf{X}_1^{>t}, \ldots, \mathbf{X}_k^{>t})$ where each copy is independently drawn from η_0, and "embed" their inputs $(x_1, \ldots, x_k) \sim \eta$ (to Disj_k^n) to the t'th coordinate of Π, having the referee set $T = t$. Since in the augmented problem player's inputs to each of the r copies are independent, and since η_0 is a product distribution, they can use *private randomness* to "fill in" their inputs to the rest of the coordinates $(\mathbf{X}_1^{<t}, \ldots, \mathbf{X}_k^{<t})$ (for a formal argument see the essentially identical proof of Claim 5.2.3). This process defines a legal input $\{(\mathbf{X}_1^\ell, \ldots, \mathbf{X}_k^\ell)\}_{\ell=1}^r \sim \nu$ for Π, and so the players can now run Π on this input and output its answer. Call this protocol π. By the premise $\left(\mathbf{X}_1^{\{-T\}}, \ldots, \mathbf{X}_k^{\{-T\}} \sim \eta_0^{r-1}\right)$, π

outputs the correct answer to the t'th copy with probability at least $1-\delta$. Furthermore, we may analyze the information complexity of Π under the distribution η_0^r (notice that $\vec{\mathsf{D}}_{\eta_0^r}(\mathsf{Aug\text{-}Disj}(r,k,\delta)) = 0$ trivially, but we are analyzing the information cost of Π which must be correct with probability $1-\delta$ over all inputs !). We have

$$
\begin{aligned}
&\mathsf{IC}^{\mathsf{ext}}_{\delta}(\eta_0, \mathsf{Disj}^n_k) \leq \mathsf{IC}^{\mathsf{ext}}_{\pi}(\eta_0) = I_{\eta_0}(\pi; x_1, \dots x_k) \\
&= \mathbb{E}_{t\in_R[r]}\left[I_{\eta_0}(\Pi; \mathbf{X}^t_1, \dots \mathbf{X}^t_k)\right] \\
&\leq \mathbb{E}_{t\in_R[r]}\left[I_{\eta_0}(\Pi; \mathbf{X}^t_1, \dots \mathbf{X}^t_k \mid \mathbf{X}^{>t}_1, \dots, \mathbf{X}^{>t}_k)\right] \quad \text{(By Lemma 1.1.14)} \\
&= \frac{1}{r}\sum_{t=1}^{r} I_{\eta_0}(\Pi; \mathbf{X}^t_1, \dots \mathbf{X}^t_k \mid \mathbf{X}^{>t}_1, \dots, \mathbf{X}^{>t}_k) \\
&= \frac{1}{r}\cdot I_{\eta_0^r}(\Pi; (\mathbf{X}^1_1, \dots, \mathbf{X}^1_k), (\mathbf{X}^2_1, \dots, \mathbf{X}^2_k), \dots, (\mathbf{X}^r_1, \dots, \mathbf{X}^r_k)) \\
&\leq \frac{\|\Pi\|}{r} = \mathsf{R}(\delta, \mathsf{Aug\text{-}Disj}(r,k,\delta)),
\end{aligned}
$$

where the last equality follows from the chain rule for mutual information.

□

5.5 Improved Space Bounds for Frequency Moments

Let $x \in \mathbb{R}^n$ represent a data stream in turnstile streaming model. We say that an algorithm solves the (p,ε,δ)-Norm problem if its output v satisfies $v \in (1\pm\varepsilon)\|x\|_p^p$ with probability at least $1-\delta$. Our main result is as follows:

Theorem 5.5.1. *For any constant $p > 2$, there exists an absolute constant $\alpha > 1$ such that for any $\varepsilon > n^{-\Omega(1)}$ and $\delta \geq 2^{-o(n^{1/p})}$, any randomized streaming algorithm that solves the (p,ε,δ)-Norm problem for $x \in [-M, M]^n$ where $M = \Omega(n^{\alpha/p})$, requires $\Omega\left(\varepsilon^{-2}\cdot n^{1-2/p}(\log M)\log 1/\delta\right)$ bits of space. In particular, the space complexity of any ε-approximate high-probability streaming algorithm (i.e, where $\delta = O(1/n)$) is at least $\Omega\left(\varepsilon^{-2}\cdot n^{1-2/p}(\log n)(\log M)\right)$.*

Proof. Let s be the space complexity of a space-optimal streaming algorithm that solves the (p, ε, δ)-Norm. By Theorem 5.3.11, there is a *path-independent* streaming algorithm $\mathcal{A}$ which solves $(p, \varepsilon, 7\delta)$-Norm using $s' = s + O(\log n + \log\log M + \log\frac{1}{\delta})$ bits of space. We will show that $\mathcal{A}$ can be used to (produce a symmetric SMP protocol) solving Aug-Disj$(r, k, 18\delta)$ (on $\mathbb{Z}_M^n$), for $r = (1 - 1/\alpha)\log_{10} M$, $k = \Theta(\varepsilon \cdot n^{1/p})$, under the hard distribution ν. Corollary 5.3.13 then implies

$$s' \geq \frac{\vec{\mathsf{R}}_\delta^{\mathsf{SYM}}(\text{Aug-Disj}(r, k, 18\delta))}{k} \geq \Omega\left(\frac{rn}{k} \cdot \min\left\{\frac{\log 1/\delta}{k}, \log k\right\}\right) = \Omega\left(\min\left\{\frac{n^{1-\frac{2}{p}}(\log M)\log 1/\delta}{\varepsilon^2} \quad , \quad \frac{n^{1-\frac{1}{p}}\log n(\log M)}{\varepsilon}\right\}\right),$$

where the second inequality follows from Lemma 5.4.2 and Corollary 5.2.6 (as in our regime of parameters $\delta \geq n \cdot 2^{-k} = 2^{-\Omega(n^{1/p})}$), and the last transition follows by substituting the values of the parameters k, M and noting that $\log k = \Theta(\log n)$ in our regime. Since $s' = s + O(\log n + \log\log M + \log\frac{1}{\delta})$, the last equation implies that so long as $\delta \geq 2^{-o(n^{1/p})}$,

$$s \geq \Omega\left(\frac{n^{1-\frac{2}{p}}(\log M)\log 1/\delta}{\varepsilon^2}\right) \quad , \quad \text{as claimed.}$$

It therefore remains to prove that $\mathcal{A}$ can be used to solve Aug-Disj$(r, k, 18\delta)$ under the input distribution ν. To this end, recall that in the Aug-Disj(r, k, δ) problem under the distribution ν, k players receive r instances each, where all instances but a single random instance $t \in_R [r]$ are independently distributed according to η_0, while $(\mathbf{X}_1^t, \ldots, \mathbf{X}_k^t) \sim \eta$. The referee receives t along with $(\mathbf{X}_1^{>t}, \ldots, \mathbf{X}_k^{>t})$ and needs to solve $\mathsf{Disj}_k^n(\mathbf{X}_1^t, \ldots, \mathbf{X}_k^t)$, i.e., distinguish between the "NO" case and the "YES" case in definition 5.2.2. Recall that for every instance $(\mathbf{X}_1^\ell, \ldots, \mathbf{X}_k^\ell)$, the referee also receives the "spiked" coordinate $I^\ell \in [n]$ of this instance (see the definition of η, η_0 in Subsection 5.2). The players will use the PIA algorithm $\mathcal{A}$ to design the SMP protocol π for Aug-Disj described in Figure 5.1.

The SMP protocol π

Input : $(\mathbf{X}_1^1, \ldots, \mathbf{X}_k^1), (\mathbf{X}_1^2, \ldots, \mathbf{X}_k^2), \ldots, (\mathbf{X}_1^r, \ldots, \mathbf{X}_k^r)$.

1. Set $\gamma \longleftarrow 4ep$, $C \longleftarrow 10^t \cdot \gamma \cdot n^{1/p}$, $\rho \longleftarrow (1+\varepsilon)(\mathcal{E}_p + (1+7\varepsilon)C^p)$ ($\mathcal{E}_p$ is defined below).
2. Each player $j \in [k]$ locally defines $\mathcal{Y}_j^\ell := 10^{\ell-1} \cdot \mathbf{X}_j^\ell \quad \forall \ell \in [r]$, and generates the stream
$$\sigma_j := \mathcal{Y}_j^1, \ldots, \mathcal{Y}_j^r$$
according to his input, and sends the referee the message $\mathcal{A}(\sigma_j)$.
3. The referee locally computes $A^{>t} := \sum_{j=1}^k \mathcal{A}(\mathcal{Y}_j^{t+1}, \mathcal{Y}_j^{t+2}, \ldots, \mathcal{Y}_j^r)$ where the addition is over the quotient ring $\mathbb{Z}^n/M$ (and M is the module representing the kernel of the automaton[a] $\mathcal{A}$).
Notice that he can do so as he has t and $(\mathbf{X}_1^{>t}, \ldots, \mathbf{X}_k^{>t})$.
4. The referee adds the value C to the "spiked" coordinate I^ℓ of the ℓ-th instance, for each $\ell \in [r]$. Let $\mathsf{C} := \mathsf{C}^1, \ldots, \mathsf{C}^r$ denote the underlying stream representing this vector (notice that he can do so since he receives the "spiked" coordinate I^ℓ of each instance).
5. The referee adds up the messages he receive from the players, over the quotient ring $\mathbb{Z}^n/M$, and outputs 1 ("YES") iff
$v := \left(\sum_{j=1}^k \mathcal{A}(\sigma_j)\right) + \mathcal{A}(\mathsf{C}^{\leq t}) - A^{>t} > \rho$.

[a] See Section 5.2 for the formal definitions and statement.

Figure 5.1: An SMP protocol for Aug-Disj$(r, k, 18\delta)$ using the PIA $\mathcal{A}$

We now turn to analyze the correctness of π. For the rest of this analysis, we fix the value of the "special" coordinate $T = t$. Notice that the value v the referee computes in π corresponds to the p-norm of the stream (with underlying frequency vector) $z := (\mathcal{Y}_1^{\leq t}, \ldots, \mathcal{Y}_k^{\leq t}, \mathsf{C}^{\leq t})$. Furthermore, $\|v\|_\infty \leq \left(\sum_{\ell=1}^t \sum_{j=1}^k \mathcal{Y}_j^\ell\right) + C \leq \sum_{\ell=1}^r \sum_{j=1}^k 10^{\ell-1} + C \leq 10^r \cdot k + C \leq O(10^r \cdot n^{1/p}) \leq M$ for a sufficiently small constant $\alpha > 1$, by our assumption $k \leq O(n^{1/p}), C \leq O(10^r \cdot n^{1/p})$, and our assumption that $M = \Omega(n^{\alpha/p})$. Therefore, the correctness of the streaming algorithm $\mathcal{A}$ guarantees that the output v of the referee

satisfies

$$\Pr\left[v \notin (1 \pm \varepsilon)\|z\|_p^p\right] \leq 7\delta. \tag{5.11}$$

Define $L_i := \sum_{j\in[k]} \sum_{\ell\in[t]} \mathcal{Y}_{j,i}^{\ell} + \mathbf{1}_I \cdot C$, where $\mathbf{1}_I$ is the indicator random variable for the event $I = i$. In this notation, $\|z\|_p^p = \sum_{i=1}^{n} (L_i)^p$. Recall that for any $i \neq I$, both in the "NO" and "YES" distributions, $\mathbf{X}_{j,i}^{\ell} \sim B(1/k)$ independently of each other, and in particular, these L_i's are independent random variables. We will need the following concentration bounds on the contribution of the L_i's:

Claim 5.5.2 (Concentration bounds)**.** *It holds that:*

- $\mathbb{E}_{\eta_0^r}[\sum_{i\neq I}(L_i)^p] \leq 2n \cdot 10^{tp}(2ep)^p$.
- *For every* $m \in \mathbb{N}$,

$$\sigma_m\left(\sum_{i\neq I}(L_i)^p\right) := \left(\mathbb{E}\left[\left|\sum_{i\neq I}(L_i)^p - \mathbb{E}\left[\sum_{i\neq I}(L_i)^p\right]\right|^m\right]\right)^{\frac{1}{m}} \leq n^{1/m} \cdot (4emp)^p \cdot 10^{tp}.$$

- *For every* $m \leq o(n^{1/p})$, $\Pr_{\eta_0}[(L_I)^p \geq (1+7\varepsilon)C^p] \leq \delta$.

We defer the proof of this technical claim to the end of this argument. In the following denote $\mathcal{E}_p := \mathbb{E}_{\eta_0}\left[\sum_{i\neq I}(L_i)^p\right]$ (note that δ can be computed by the referee as it is public knowledge). Applying the generalized Chebychev's inequality (Lemma 5.1.15) with $m = \log 1/\delta$ and $\lambda = 2$, the first two propositions of Claim 5.5.2 guarantee that

$$\begin{aligned}
&\Pr\left[\left|\sum_{i\neq I}(L_i)^p - \mathcal{E}_p\right| > \varepsilon \cdot C^p\right] = \Pr\left[\left|\sum_{i\neq I}(L_i)^p - \mathbb{E}\left[\sum_{i\neq I}(L_i)^p\right]\right| > \varepsilon \cdot C^p\right] \\
&\leq \Pr\left[\left|\sum_{i\neq I}(L_i)^p - \mathbb{E}\left[\sum_{i\neq I}(L_i)^p\right]\right| > 2 \cdot \sigma_m\left(\sum_{i\neq I}(L_i)^p\right)\right] \leq 2^{-m} = 2^{-\log 1/\delta} < \delta,
\end{aligned} \tag{5.12}$$

where the first inequality is by the second proposition of Claim 5.5.2 (which implies that $2 \cdot \sigma_m \left(\sum_{i \neq I} (L_i)^p \right) \leq \varepsilon C^p$ whenever $\varepsilon \geq \Omega(n^{1/m-1}) = n^{-\Omega(1)}$), and the second inequality holds by our assumption that $\delta > 2^{-o(n^{1/p})}$. Combining (5.12) with the third proposition of Claim 5.5.2 implies

$$\Pr_{\eta_0} \left[\|z\|_p^p > \mathcal{E}_p + C^p(1+7\varepsilon) \right] = \Pr_{\eta_0} \left[\|z\|_p^p > (\mathcal{E}_p + \varepsilon C^p) + C^p(1+7\varepsilon) \right] \leq 2\delta.$$

Hence, by definition of the "threshold" $\rho := (1+\varepsilon)[\mathcal{E}_p + (1+7\varepsilon)C^p]$, we conclude by (5.11) that in the "NO" case,

$$\Pr_{\eta_0} [v > \rho] \leq \Pr \left[\|z\|_p^p > \mathcal{E}_p + (1+7\varepsilon)C^p \right] \leq 9\delta. \tag{5.13}$$

On the other hand, in the "YES" case, the coordinate I is such that $\mathbf{X}_{j,I}^{\ell} = 1$ for all $j \in [k], \ell \in [r]$. Setting $k = 128e\varepsilon \cdot n^{1/p} = \Theta(\varepsilon \cdot n^{1/p})$, the contribution of this coordinate to the p-norm of z is

$$\begin{aligned}
&\left(C + \sum_{\ell=1}^{t} 10^{\ell-1} \cdot k \right)^p \geq \left(10^t \cdot \gamma \cdot n^{1/p} + 10^t \cdot 128e\varepsilon \cdot n^{1/p} \right)^p = \\
&= n \cdot 10^{tp} \cdot \gamma^p \left(1 + 128\varepsilon/\gamma\right)^p \geq n \cdot 10^{tp} \cdot \gamma^p \cdot e^{\frac{128e\varepsilon p}{2\gamma}} \quad \text{(since } 128e\varepsilon/\gamma < 1/2\text{)} \\
&= n \cdot 10^{tp} \cdot \gamma^p \cdot e^{\frac{128e\varepsilon p}{8ep}} \geq n \cdot 10^{tp} \cdot \gamma^p \cdot (1+16\varepsilon) = (1+16\varepsilon)C^p.
\end{aligned}$$

Furthermore, (5.12) ensures that, except with probability δ, the contribution of all the rest coordinates ($i \neq I$) is at least $\mathcal{E}_p - \varepsilon C^p$, and thus in the "YES" case,

$$\Pr \left[\|z\|_p^p > \mathcal{E}_p + (1+15\varepsilon)C^p \right] = \Pr \left[\|z\|_p^p > (\mathcal{E}_p - \varepsilon C^p) + C^p(1+16\varepsilon) \right] \geq 1 - 2\delta$$

Finally, (5.11) implies that under the "YES" distribution,

$$\begin{aligned}
&\Pr\left[v > \rho\right] \\
&\geq \Pr\left[\|z\|_p^p > (1+\varepsilon)\rho\right] - 7\delta = \Pr\left[v > (1+\varepsilon)^2 \cdot (\mathcal{E}_p + (1+7\varepsilon)C^p)\right] - 7\delta \\
&\geq \Pr\left[\|z\|_p^p > \mathcal{E}_p + 3\varepsilon\mathcal{E}_p + (1+3\varepsilon)(1+7\varepsilon)C^p\right] - 7\delta \\
&\geq \Pr\left[\|z\|_p^p > \mathcal{E}_p + 3\varepsilon C^p + (1+11\varepsilon)C^p\right] - 7\delta \qquad \text{(since } \mathcal{E}_p \leq C^p\text{)} \\
&= \Pr\left[\|z\|_p^p > \mathcal{E}_p + (1+14\varepsilon)C^p\right] \geq 1 - 9\delta
\end{aligned} \tag{5.14}$$

We conclude from equations (5.13) and (5.14) that setting the threshold ρ guarantees that π is correct with probability at least $1 - 18\delta$, which completes the reduction and the proof of Theorem 5.5.1.

It remains to prove Claim 5.5.2.

Proof of Claim 5.5.2. **First proposition:** Since all coordinates except coordinate I are identically distributed, we can write $\mathbb{E}_{\eta_0}[\sum_{i\neq I}(L_i)^p] = (n-1)\cdot\mathbb{E}[(L_t)^p] + \mathbb{E}[(C+L_t)^p]$, where

$$L_t := \sum_{j\in[k]}\sum_{\ell\in[t]} 10^{\ell-1}\mathbf{X}_j^\ell \quad , \quad \text{and } \mathbf{X}_j^\ell\text{'s are i.i.d } B(1/k).$$

We first prove the following lemma, which upper bounds the p'th moment of a single coordinate ($i \neq I$) in a "NO" instance. Though it is a spacial case of Lemma 5.1.14, for completeness we present an elementary (yet slightly weaker) proof that will be sufficient in our applications.

Lemma 5.5.3. *For every $p \geq 1$,* $\quad \mathbb{E}[(L_t)^p] \leq (2ep)^p \cdot 10^{tp}$.

Proof. We shall show by induction on t, that there exists a function $f(p) \leq (2ep)^p$ for which

$$\mathbb{E}[(L_t)^p] \leq f(p)\cdot 10^{tp}. \tag{5.15}$$

Indeed, define the function $f(p)$ recursively by the formula: $f(p+1) := (ep)^p + f(p)$. It follows that

$$f(p) = (ep)^p + (e(p-1))^{p-1} + (e(p-2))^{p-2} + \ldots + 1 \leq (2ep)^p,$$

as desired. The proof for the base case ($t = 1$) is very similar to the general case, so we postpone it to the end of the proof. Suppose (5.15) is true for for all integers up to t. We shall show that

$$\mathbb{E}[(L_{t+1})^p] \leq f(p+1) \cdot 10^{(t+1)p}. \tag{5.16}$$

To this end, we may write $L_{t+1} := \Delta_{t+1} + L_t$, where $\Delta_{t+1} := 10^t \cdot \sum_{j \in [k]} \mathbf{X}_j^{t+1}$. We first bound $\mathbb{E}[(\Delta_{t+1})^p]$:

$$\begin{aligned}
\mathbb{E}[(\Delta_{t+1})^p] &= (10^{tp}) \cdot \mathbb{E}\left[\left(\sum_{j \in [k]} \mathbf{X}_j^{t+1}\right)^p\right] \\
&\leq 10^{tp} \cdot \sum_{r=1}^{p} \binom{k}{r} \cdot p^r \cdot \mathbb{E}\left[\prod_{i=1}^{r} \mathbf{X}_{j_i}^{t+1}\right] \quad \text{(By the multinomial formula)} \\
&\leq 10^{tp} \cdot \sum_{r=1}^{p} \left(\frac{ek}{r}\right)^r \cdot p^r \cdot \left(\frac{1}{k}\right)^r \quad \text{(By Stirling's approximation and in dependence of } \mathbf{X}_j^{t+1}\text{'s)} \\
&\leq 10^{tp} \cdot \sum_{r=1}^{p} \left(\frac{ep}{r}\right)^r \leq (ep)^p \cdot 10^{tp}.
\end{aligned} \tag{5.17}$$

We therefore have

$$\begin{aligned}
&\mathbb{E}[(L_{t+1})^p] = \mathbb{E}[(\Delta_{t+1} + L_t)^p] \\
&\leq 2^{p-1} \cdot (\mathbb{E}[(\Delta_{t+1})^p] + \mathbb{E}[(L_t)^p]) \\
&\leq 2^{p-1} \cdot \left(\mathbb{E}[(\Delta_{t+1})^p] + f(p) \cdot 10^{tp}\right) \qquad \text{(by the inductive hypothesis)} \\
&\leq 2^{p-1} \cdot 10^{tp} \left[(ep)^p + f(p)\right] \qquad \text{(by (5.17))} \\
&\leq 10^{(t+1)p} \cdot \left[(2ep)^p + f(p)\right] \\
&= 10^{(t+1)p} \cdot f(p+1) \ ,
\end{aligned}$$

which finishes the proof of (5.16). For the base case ($t = 1$), we need to show that $\mathbb{E}[(L_1)^p] := \mathbb{E}[(\sum_{j\in[k]} \mathbf{X}_j^1)^p] \leq 10^p \cdot f(p)$. Indeed, repeating essentially the same calculation as in (5.17), one obtains

$$\begin{aligned}
&\mathbb{E}[(L_1)^p] \leq (2ep)^p = 2^p \cdot (ep)^p \leq 10^p \cdot (e(p-1))^{p-1} \\
&\leq 10^p \cdot \left[(e(p-1))^{p-1} + f(p-1)\right] = 10^p \cdot f(p).
\end{aligned}$$

This finishes the proof of (5.15), and therefore concludes the proof of Claim 5.5.3. □

Substituting the value of $C = 10^t \cdot \gamma \cdot n^{1/p}$, we conclude by Lemma 5.5.3 and (5.18) that

$$\begin{aligned}
&\mathbb{E}_{\eta_0}\left[\sum_{i\neq I}(L_i)^p\right] = (n-1) \cdot \mathbb{E}[(L_t)^p] + \mathbb{E}[(C + L_t)^p] \\
&\leq (n-1) \cdot (2ep)^p \cdot 10^{tp} + 2C^p \\
&= (n-1) \cdot (2ep)^p \cdot 10^{tp} + 2n \cdot \gamma^p \cdot 10^{tp} \\
&\leq n \cdot 10^{tp} \left(2\gamma^p + (2ep)^p\right).
\end{aligned}$$

Second proposition: To upper bound the m-th moment of $\sum_{i\neq I}(L_i)^p$, we note that $\sum_{i\neq I}(L_i)^p$ is a sum of independent random variables, and thus Lemma 5.1.14 implies that

$$\mathbb{E}\left[\left|\sum_{i\neq I}(L_i)^p\right|^m\right] \leq \left(\frac{3em}{\log m}\right)^m \cdot \sum_{i\neq I}\mathbb{E}[(L_i)^{mp}]$$
$$\leq (n-1)\cdot(3em)^m\cdot(2emp)^{mp}\cdot 10^{tmp}$$
$$\leq n\cdot(4emp)^{mp}\cdot 10^{tmp},$$

where the second inequality follows again from Lemma 5.5.3, taken with $p := mp$. The second proposition of Lemma 5.5.2 now follows by raising both sides of the above inequality to the $1/m$ power.

Third proposition: We first upper bound the expected contribution of the I'th coordinate under η_0:

$$\mathbb{E}_{\eta_0}\left[(L_I)^p\right] = \mathbb{E}[(C+L_t)^p] = C^p\cdot\sum_{r=0}^{p}\binom{p}{r}\cdot\mathbb{E}\left[\left(\frac{L_t}{C}\right)^r\right] \leq C^p\cdot\sum_{r=0}^{p}\left(\frac{ep}{r}\right)^r\cdot\frac{(2er)^r\cdot 10^{tr}}{C^r}$$
$$= C^p\cdot\sum_{r=0}^{p}\left(\frac{2e^2\cdot p\cdot 10^t}{C}\right)^r \leq C^p\sum_{r=0}^{\infty}\varepsilon^{-r} \leq \frac{1}{1-\varepsilon}\cdot C^p \leq (1+2\varepsilon)C^p, \tag{5.18}$$

where the third transition follows from Lemma 5.5.3 (applied p times with $p = r$), and the second before last transition follows since $\frac{2e^2\cdot p\cdot 10^t}{C} = \frac{2e^2\cdot p\cdot 10^t}{10^t\cdot\gamma\cdot n^{1/p}} \ll \varepsilon$ for large enough n.

Next, we upper bound the m-th moment of L_I^p. Similar to the calculations in (5.18), we have

$$\sigma_m((L_I^p))^m := \mathbb{E}_{\eta_0}\left[|(L_t+C)^p - \mathbb{E}[(L_t+C)^p]|^m\right] \leq$$

$$\leq \mathbb{E}_{\eta_0}\left[|(L_t+C)^p - C^p]|^m\right] \leq \mathbb{E}_{\eta_0}\left[C^{pm}\cdot\left(\sum_{r=0}^{p}\left(\frac{ep}{r\cdot C}\right)^r\cdot L_t^r - 1\right)^m\right]$$

$$= \mathbb{E}_{\eta_0}\left[C^{pm}\cdot\left(\sum_{r=1}^{p}\left(\frac{ep}{r\cdot C}\right)^r\cdot L_t^r\right)^m\right] \leq \mathbb{E}_{\eta_0}\left[C^{pm}\cdot p^m\cdot\sum_{r=1}^{p}\left(\frac{ep}{r\cdot C}\right)^{rm}\cdot L_t^{rm}\right] \tag{5.19}$$

$$\leq C^{pm}\cdot\sum_{r=1}^{p}\left(\frac{ep^2}{r\cdot C}\right)^{rm}\cdot\mathbb{E}_{\eta_0}[L_t^{rm}] \leq C^{pm}\cdot\sum_{r=1}^{p}\left(\frac{ep^2}{r\cdot C}\right)^{rm}\cdot(2eprm)^{rm}\cdot 10^{trm} \qquad \text{(By Lemma 5.5.3)}$$

$$= C^{pm}\cdot\sum_{r=1}^{p}\left(\frac{10^t\cdot 2e^2p^3m}{C}\right)^{rm} \leq C^{pm}\cdot\frac{\varepsilon^m}{1-\varepsilon^m} \leq (2\varepsilon C^p)^m, \tag{5.20}$$

where (5.19) follows from Jensen's inequality ($(\sum_{i=1}^n a_i)^m \leq n^m\cdot\sum_{i=1}^n a_i^m$), and the second before last transition again follows since $\frac{10^t\cdot 2e^2p^3m}{C} = \frac{10^t\cdot 2e^2p^3m}{10^t\cdot\gamma\cdot n^{1/p}} \ll \varepsilon$ by the premise $m = o(n^{1/p})$.

Given the assumption $\delta > 2^{-o(n^{1/p})}$, we can now apply Lemma 5.1.15 with $m = \log 1/\delta$ ($\leq o(n^{1/p})$), $\lambda = 2$, to conclude that

$$\Pr_{\eta_0}[(L_I)^p \geq (1+7\varepsilon)C^p] \leq \Pr_{\eta_0}\left[|(L_I)^p - \mathbb{E}[(L_I)^p]| > 4\varepsilon C^p\right]$$

$$\leq \Pr_{\eta_0}\left[\ |(L_I)^p - \mathbb{E}[(L_I)^p]| > 2\cdot\sigma_m((L_I)^p)]\ \right] \leq 2^{-m} = \delta,$$

where the first and second transitions follow from (5.18) and (5.20) respectively. □

□

Chapter 6

Applications to Economics: Welfare Maximization with Limited Interaction

In this chapter we study the tradeoff between the amount of communication and the number of rounds of interaction required to find an (approximately) optimal matching in a bipartite graph. In our model there are n "players" and m "items". Each player initially knows a subset of the items to which it may be matched (i.e. m bits of information). The players communicate in rounds: in each round each player writes a message on a shared blackboard. The message can only depend on what the player knows at that stage: his initial input and all the messages by all other players that were written on the blackboard in previous rounds.

This problem was recently introduced by [62] as a simple market scenario: the players are unit-demand bidders and our goal is to find an (approximately) welfare-maximizing allocation of the items to players. The classic auction of [60] – that may be viewed as a simple Walrasian-like market process for this setting – can be implemented as to find an approximately optimal allocation where each player needs only send $O(\log n)$ bits of communication (on the average). The question considered by [62] was whether such a low communication burden suffices without using multiple rounds of interaction. As

a lower bound, they proved that a non-interactive protocol, i.e. one that uses a *single round of communication*, cannot get a $n^{1/2-\epsilon}$-factor approximation (for any fixed $\epsilon > 0$) with $n^{o(1)}$ bits of communication per player. As upper bounds they exhibited (I) an $O(\log n)$-round protocol, where each player sends $O(\log n)$ bits per round, that gets a $\frac{1}{1-\delta}$-factor approximation (for any fixed $\delta > 0$) and (II) for any fixed $r \geq 1$, a r-round protocol, where each player sends $O(\log n)$ bits per round, that gets a $O(n^{1/(r+1)})$-approximation.

The natural question at this point is whether there are r-round protocols with better approximation factors that still use $n^{o(1)}$ bits of communication per player. This question was left open in [62], where it was pointed out that it was even open whether the exactly optimal matching can be found by 2-round protocols that use $O(\log n)$ bits of communication per player. We answer this open problem by proving lower bounds for any fixed number of rounds.

Theorem: For every $r \geq 1$ there exists $\epsilon(r) = \exp(-r)$, such that every (deterministic or randomized) r-round protocol requires $n^{\epsilon(r)}$ bits of communication per player in order to find a matching whose size is at least $n^{-\epsilon(r)}$ fraction of the optimal matching.

Our proof relies on information theory, and uses a type of multiparty round-reduction argument which requires the analysis of information sent by multiple players in a way that avoids summing the information costs. In contrast to the standard two-party model, round-elimination arguments in the multiparty model are non-trivial, as the number of parties scales with the input size, and calls for a subtle embedding argument (we discuss this further in Section 6.2).

6.0.1 More context and related models

The bipartite matching problem is clearly a very basic one and obviously models a host of situations beyond the simple market scenario that was the direct motivation of [62] and this paper. Despite having been widely studied, even its algorithmic status is not

well understood, and it is not clear whether a nearly-linear time algorithm exists for it. (The best known running time (for the dense case) is the 40-year old $O(n^{2.5})$ algorithm of [83], but for special cases like regular or near-regular graphs nearly linear times are known (e.g. [3, 156]). In parallel computation, a major open problem is whether bipartite matching can be solved in deterministic parallel poly-logarithmic time (with a polynomial amount of processors). (Randomized parallel algorithms for the problem [127, 101] have been known for over 25 years.) It was suggested in [62] that studying the problem in the communication complexity model is an approach that might lead to algorithmic insights as well.

The bipartite matching problem has been studied in various other multi-party models that focus on communication as well. In particular, strong and tight bounds for approximate matching are known in the weaker "message passing" or "private channels" models [85] that have implications to models of parallel and distributed computation . Related work has also been done in networked distributed computing models, e.g., [119]. "One-way" communication models are used to analyze streaming or semi-streaming models and some upper bounds (e.g., [109]) as well as weak lower bounds [73] are known for approximate matchings in these models. For "r-way" protocols, a super-linear communication lower bound was recently shown by [77] for *exact* matchings, in an incomparable model[1]. A somewhat more detailed survey of these related models can be found in the appendix of [62].

It should be noted that the open problems mentioned above remain so even in the standard two-party setting where each of the two players holds all the information of $n/2$ of our players. We do not know any better upper bounds than what is possible in the

[1]Besides of the fact that this lower bound applies only for testing *exact* matching and not approximate matchings, their model consists of p-parties for some *constant or logarithmic* p, who are communicating in some fixed number of *sequential* rounds (not simultaneous). The input itself of each player is therefore *super-linear* in the number of nodes of the input graph (n), and indeed they prove a super-linear communication lower bound for fixed-round protocols. Such a result is obviously impossible in our model. The [77] model does not seem to capture the economic scenario we attempt to model in this paper (i.e., that of private-valuations) and therefore we view these results as tangential, as also evidenced by the distinct proof-techniques.

multi-player model, and certainly, as the model is stronger, no better lower bounds are known. We also do not know whether our lower bound (or the single round one of [62]) applies also in this stronger two-player model.

6.0.2 The blackboard model and approximate matchings

Our framework in this chapter is the *Number-In-Hand* (NIH) multiparty communication complexity model with shared blackboard. In this model, n players receive inputs $(x_1, x_2, \ldots, x_n) \in \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \mathcal{X}_n$ respectively. In our context, each of the n players (bidders) is associated with a node $u \in U = [n]$ of some bipartite graph $G = (U, V, E)$, and her input is the set of incident edges on her node (her demand set of items in $V = [m]$). The players' goal is to compute a maximum set of disjoint connected pairs $(u, v) \in E(G)$, i.e., a maximum matching in G (we define this formally below).

The players communicate in some fixed number of rounds r, where in each communication round, players *simultaneously* write (at most) ℓ bits each on a *shared blackboard* which is viewable to all parties. We sometimes refer to the parameter ℓ as the *bandwidth* of the protocol. In a deterministic protocol, each player's message should be completely determined by the content of the blackboard and her own private input x_i. In a randomized protocol, the message of each player may further depend on both public and private random coins. When player's inputs are distributional ($(x_1, x_2, \ldots, x_n) \sim \mu$) which is the setting in this paper, we may assume without loss of generality that the protocol is *deterministic*, since the averaging principle asserts that there is always some fixing of the randomness that will achieve the same performance with respect to μ. We remark that by Yao's minimax theorem (see e.g. [111]), our main result applies to randomized protocols as well[2].

[2]More formally, if there is a distribution μ on players inputs such that the approximation ratio of any r-round deterministic protocol with respect to μ is at most α in expectation, then Yao's minimax theorem asserts this lower bound applies to randomized r-round protocols as well.

The transcript of a protocol π (namely, the content of the blackboard) when executed on an input graph G is denoted by $\Pi(G)$, or simply Π when clear from context. At the end of the r'th communication round, a *referee* (the "central planner" in our context) computes a matching $\hat{\mathcal{M}}(\Pi)$, which is completely determined by Π. We call this the *output* of the protocol.

We will be interested in protocols that compute *approximate matchings*. To make this more formal, let $\mathcal{G}(n,m)$ denote the family of bipartite graphs on (n,m)-vertex sets respectively, and denote by $\mathcal{F}(n,m)$ the family of all matchings in $\mathcal{G}(n,m)$ (not necessarily maximum matchings). Denote by $|\mathcal{M}(G)|$ the size of a maximum matching in the input graph G. We require that the output of any protocol satisfies $\hat{\mathcal{M}}(\Pi) \in \mathcal{F}(n,m)$. The following definition is central to this work.

Definition 6.0.4 (Approximate Matchings)**.** *We say that a protocol π computes an α-*approximate matching *($\alpha \geq 1$) if $|\hat{\mathcal{M}}(\Pi) \cap E(G)|$ is at least $\frac{1}{\alpha} \cdot |\mathcal{M}(G)|$, i.e., if the number of matched pairs $(u,v) \in E(G)$ is at least a $(1/\alpha)$-fraction of the maximum matching in G. Similarly, when the input graph G is distributed according to some distribution μ (i.e., $(x_1, x_2, \ldots, x_n) \sim \mu$), we say that the* approximation ratio *of π is $\alpha \geq 1$ if*

$$\mathop{\mathbb{E}}_{G \sim \mu}[[]\, |\hat{\mathcal{M}}(\Pi) \cap E(G)|] \geq \frac{1}{\alpha} \cdot \mathop{\mathbb{E}}_{G \sim \mu}[[]\, |\mathcal{M}(G)|].$$

The *expected matching size* of π is $\mathbb{E}_{\mu}[|\hat{\mathcal{M}}(\Pi) \cap E(G)|]$ (we remark that the "hard" distribution we construct in the next section will satisfy $|\mathcal{M}(G)| \equiv n$ for all G in the support of μ, so the quantity $\mathbb{E}_{G \sim \mu}[[]\, |\mathcal{M}(G)|]$ will always be n). Note that these definitions in particular allow the protocol to be *erroneous*, i.e., the referee is allowed to output "illegal" pairs $(u,v) \notin E(G)$, but we only count the correctly matched pairs. Our lower bound holds even with respect to this more permissive model.

6.1 A hard distribution for r-round protocols

We begin by defining a family of hard distributions for protocols with r rounds. Recall that $\mathcal{G}(n, m)$ is the family of bipartite graphs on (n, m) vertex-sets. For any given number of rounds r, we define a hard distribution μ_r on bipartite graphs in $\mathcal{G}(n_r, m_r)$. μ_r is recursively defined in Figure 6.1.

A recursive definition of the hard distribution μ_r

In what follows, ℓ is a parameter to be defined later.

1. For $r = 0$, $G^0 = (U^0, V^0, E^0)$ consists of a set of n_0 bidders $U^0 = \{b_1, \ldots, b_{n_0}\}$ and a set of m_0 items $V^0 = \{j_1 \ldots, j_{m_0}\}$, such that $n_0 = m_0 = \ell^{11}$. E^0 is then obtained by selecting a random permutation $\sigma \in_R S_{\ell^{11}}$ and connecting $(b_i, j_{\sigma(i)})$ by an edge. This specifies μ_0.

2. For any $r \geq 0$, the distribution μ_{r+1} over $G^{r+1} = (U^{r+1}, V^{r+1}, E^{r+1})$ is defined as follows:

   Vertices:

   - The set of bidders is $U^{r+1} := \bigcup_{i=1}^{n_r^{10}} B_i$ where $|B_i| = n_r$. Thus, $n_{r+1} = n_r^{11}$.
   - The set of items is $V^{r+1} := \bigcup_{j=1}^{n_r^{10}+\ell \cdot n_r^8} T_j$ where $|T_j| = m_r$. Thus, $m_{r+1} = (n_r^{10} + \ell \cdot n_r^8) \cdot m_r$.

   Edges: Let d_r be the degree of each vertex in U^r (this is well defined as the degree of any vertex is fixed for every graph in the support of μ_r). The distribution on edges is obtained by first choosing $\ell \cdot n_r^8$ random indices $\{a_1, a_2, \ldots a_{\ell \cdot n_r^8}\}$ from $[n_r^{10}+\ell \cdot n_r^8]$, and a random invertible map $\sigma : [n_r^{10}] \longrightarrow [n_r^{10}+\ell \cdot n_r^8] \backslash \{a_1, a_2, \ldots a_{\ell \cdot n_r^8}\}$. Each vertex $u \in B_i$ is connected to d_r uniformly random vertices *in each one* of the blocks $T_{a_1}, T_{a_2}, \ldots, T_{a_{\ell \cdot n_r^8}}$, using independent randomness for each of the blocks. The entire block B_i is further connected to the entire block $T_{\sigma(i)}$ using an independent copy of the distribution μ_r. Note that this is well defined, as $|B_i| = n_r$, $|T_{a_j}| = |T_{\sigma(i)}| = m_r$ and μ_r is indeed a distribution on bipartite graphs from $\mathcal{G}(n_r, m_r)$.

Figure 6.1: A hard distribution for r-round protocols.

As standard, the input of each of bidder $u \in U^{r+1}$ is the set of incident edges on the vertex u (defined by μ_{r+1}). Note that every graph in the support of μ_{r+1} has a perfect matching ($|\hat{\mathcal{M}}(G^{r+1})| = n_{r+1}$).

Notation To facilitate our analysis, the following notation will be useful. Notice that each block B_i of players is connected to exactly $\ell \cdot n_r^8 + 1$ blocks of items whose indices we denoted by

$$\mathcal{I}_i := \{\sigma(i), a_1, a_2, \dots a_{\ell \cdot n_r^8}\}.$$

For each B_i, let $\tau_i : \mathcal{I}_i \longrightarrow [\ell \cdot n_r^8 + 1]$ be the bijection that maps any index in $\mathcal{I}_i$ to its location in the sorted list of $\mathcal{I}_i$ (i.e., $\tau_i^{-1}(1)$ is the smallest index in $\mathcal{I}_i$, $\tau_i^{-1}(2)$ is the second smallest index in $\mathcal{I}_i$ and so forth). By a slight abuse of notation, let us denote $G_j^i := (B_i, T_{\tau_i^{-1}(j)})$ the (induced) subgraph of G on the sets $(B_i, T_{\tau_i^{-1}(j)})$, for each $j \in [\ell \cdot n_r^8 + 1]$. Similarly, for a bidder $u \in B_i$, let $G_j^i(u) := (u, T_{\tau_i^{-1}(j)})$ denote the (induced) subgraph of G on the sets $(u, T_{\tau_i^{-1}(j)})$. In this notation, the entire input of players in B_i is $\Gamma_i := \{G_1^i, G_2^i, \dots, G_{\ell \cdot n_r^8 + 1}^i\}$. Let

$$J_i := \tau_i(\sigma(i))$$

denote the index of the "hidden graph" $G_{J_i}^i = (B_i, T_{\sigma(i)})$. To avoid confusion (with the other indices j), we henceforth write

$$G(J_i) := G_{J_i}^i.$$

Note that by symmetry of our construction, the index J_i is uniformly distributed in $[\ell \cdot n_r^8 + 1]$. The following fact will be crucial to our analysis:

Fact 6.1.1 (Marginal Indistinguishability)**.** *For any block B_i and any bidder $u \in B_i$, $G_j^i(u) \sim G_k^i(u)$ for any $j \neq k \in [\ell \cdot n_r^8 + 1]$. That is, the* marginal *distribution of the induced subgraphs on the vertex u is the same for any $j \in [\ell \cdot n_r^8 + 1]$.*

Proof. By construction of μ_{r+1}, the marginal distribution of edges in $G^i_{J_i}(u)$ is uniform (over $T_{\tau_i^{-1}(J_i)}$), and the number of edges (degree of u) in $G^i_{J_i}(u)$ is always d_r. But this is precisely the definition of the distribution on edges of $G^i_j(u)$ for all $j \in [\ell \cdot n_r^8 + 1] \setminus \{J_i\}$. We conclude that the distribution of $G^i_j(u)$ is the same for all $j \in [\ell \cdot n_r^8 + 1]$. □

Finally, Let $\mathcal{B}$ denote the partition of bidders in $U := U^{r+1}$ into the blocks B_i, and $\mathcal{T}$ denote the partition of items in $V := V^{r+1}$ into the blocks T_j. Since $\mathcal{T}$ and $\mathcal{B}$ are fixed (publicly known) in the distribution μ_{r+1}, our entire analysis is performed under the implicit conditioning on $\mathcal{T}, \mathcal{B}$. Note that $\mathcal{T}$ does not reveal the identity of the "fooling blocks" T_{a_j}, but only the items belonging to each block.

6.2 Main Result and Overview of the Proof

In this section we state our main result. Recall that the expected matching size of π (with respect to μ) is $\mathbb{E}_\mu[|\hat{\mathcal{M}}(\Pi) \cap E(G)|]$. We shall prove the following theorem.

Theorem 6.2.1 (Main Result)**.** *The expected matching size of any r-round protocol under μ_r is at most $5n_r^{1-1/11^r}$. This holds as long as the number of bits sent by each player at any round is at most $\ell = n_r^{1/11^{r+1}}$. In particular, since μ_r has a perfect matching, the approximation ratio of any r-round protocol is no better than $\Omega\left(n^{1/11^r}\right)$.*

The intuition behind the proof is as follows. Consider some $(r+1)$-round protocol π (with bandwidth ℓ), and let $M_{B_i} = M^1_{B_i} M^2_{B_i}, \ldots, M^{n_r}_{B_i}$ denote the (concatenated) messages sent by all of the bidders in a block B_i in the first round of π. Informally speaking, the distribution μ_{r+1} is designed so that messages of players in B_i (M_{B_i}) convey little information about the "hidden" graph $G(J_i)$. Intuitively, this will be true since the *marginal* distribution of the graphs $G^i_j(u)$ for any bidder $u \in B_i$ is identical for each j (Fact 6.1.1) and therefore a bidder in B_i will not be able to distinguish between vertices (items) in $\bigcup_{j=1}^{\ell \cdot n_r^8} T_{a_j}$ and in $T_{\sigma(i)}$. By simultaneity of the protocol, we will show that the latter condi-

tion also implies that the *total* information conveyed by M_{B_i} on $G(J_i)$ is small. In order to make this information $\ll 1$ bit, the parameters are chosen so that n_r grows doubly-exponentially in r ($n_r \approx \ell^{11^r}$), and this choice is the cause for the approximation ratio we eventually obtain. Intuitively, the fact the little information is conveyed by each block on the "hidden graph" implies that the distribution of edges in the graph $G(J_i)$ is still close to μ_r even *conditioned* on the first message of the i'th block M_{B_i}. Now suppose an $(r+1)$-round protocol finds a large matching with respect to the original distribution μ_{r+1} (in expectation). Then the expected matching size on each of the $G(J_i)$ must be large as well. Hence, "ignoring" the first round of the protocol, the original protocol essentially induces an r-round protocol for finding a large matching with respect to the distribution μ_r, up to some error term (indeed, *some* edges of $G(J_i)$ may have already been discovered in the first round of the protocol, but the argument above ensures that not too many are revealed). Since we have now reduced the problem to finding a large matching under μ_r using only r rounds, we may use an inductive hypothesis to upper bound this expected matching size.

Making the above intuition precise is complicated by the fact that, unlike standard "round-elimination" arguments in the two-party setting, in our setup one cannot simply "project" an r-round n_{r+1}-party protocol (with inputs $\sim \mu_{r+1}$) directly to the distribution μ_r, since a protocol for the latter distribution has only n_r players (inputs). To remedy this, we use the conditional independence properties of our construction together with an embedding argument to obtain the desired lower bound. The embedding part of the proof is subtle, since in general, conditioning on the first message M_1 correlates the inputs of the players, so it is not clear how to sample the "missing" inputs of the "higher-dimensional" protocol. Luckily and crucially, the edges to the "fooling blocks" T_{a_j} in μ_{r+1} were chosen independently for each bidder $u \in U$ (unlike the hidden graphs $G(J_i)$ in which players have correlated edges). This independence is what allows to embed a lower-dimensional

graph correctly according to the *conditioned on the first message* M_1 using no communication.

The formal proof of Theorem 6.2.1 is subtle and technical. We refer the reader to the full version of the paper for the detailed proof.

Chapter 7

Applications to Privacy and Secure Computation

7.1 Interactive computation between two untrusting parties: From honest-but-curious to malicious

In this Chapter we consider an application of the interactive information odometer presented in Section 3.4, to the setting where Alice and Bob *do not trust each other* and wish to compute a function $f(X, Y)$ of their inputs while revealing as little information to each other as possible. This setting has been extensively studied in the theoretical cryptography literature. In the the case of $3+$ parties with private channels (and honest majority), [23] showed that secure multiparty computation is possible, that is, it is possible to compute any function of the player's inputs while revealing nothing beyond the value of the function to the players. It is known that no such protocol can exist for two parties, even in the case of *honest-but-curious* participants. In this model, Chor and Kushilevitz [54] characterized the family of two-party Boolean functions computable with perfect privacy. This characterization was extended by Kushilevitz [110] and Beaver [19] to general-valued functions, asserting that most function are not privately computable. Subsequent

papers studied the privacy loss of specific functions, and explored communication trade-offs required to achieve perfect or approximate privacy in the honest model (Bar Yehuda et al [13] [65] [2]).

In the *malicious* model, where one of the parties is assumed to be adversarial, much less was known. When the malicious party is assumed to be computationally bounded, and thus one can use cryptographic primitives, [74] ensure the "best possible" privacy can be preserved, assuming the existence of so called "trapdoor permutations" [1]. Other works define a weaker notion of privacy and obtain privacy-preserving schemes for specific functions under these notions ([131, 122]). None of these works has a pure statistical security guarantee against general, unrestricted adversaries.

As information-theoretically secure two-party computation is impossible for most functions, several approaches for quantifying privacy loss have been proposed over the years in the security and privacy literature [106, 65, 123, 104]. In fact, one way to view the information complexity $\mathsf{IC}(f, \varepsilon)$ is as the smallest (average) amount of information Alice and Bob must reveal to each other to compute f with error ε (here the information revealed by the value of $f(X, Y)$ is included in the information complexity). Thus, information complexity gives the precise answer to the two-party private computation in the information-theoretic *honest-but-curious* model: Alice and Bob will try to learn about Y and X respectively from the protocol, while adhering to its prescribed execution.

Therefore, in the honest-but-curious case, a protocol π whose information cost is close to the information complexity of f will achieve a near-optimal performance in terms of privacy, revealing only $\approx I := \mathsf{IC}(f, \varepsilon)$ information to Alice and Bob. That is, assuming Alice and Bob adhere to the execution of π[2].

[1] The authors show that a malicious player cannot learn anything more that the value of $f(X', Y)$ for any X' of her choice.

[2] In the secure computation literature information loss is typically measured as the difference between what the parties learn (the information cost) and what they were supposed to learn (the mutual information between the output and the other partie's input). To keep notation simple, we ignore the latter term here, since it does not substantially affect any of the result. To be specific, one may assume that Alice and Bob are trying to compute only a few bits of output, and thus this term is negligible.

What happens when either Alice or Bob is malicious? There are easy examples where a cheating Bob can extract $\omega(I)$ bits of information on Alice's input, if the protocol is executed naively (an instructive example is the standard hashing protocol for the Equality function). Is there a way to compile π into a protocol π' such that (1) if Alice and Bob are honest is close to π in terms of computing f; (2) even if Alice or Bob are dishonest, reveals at most $O(I)$ information to the dishonest party (that is, a dishonest Bob cannot "phish" more than $O(I)$ bits of information out of Alice)? If information complexity was known to be equal to communication complexity, we could just compress π into a protocol π' with $O(I)$ bits of communication. Even if Alice or Bob are dishonest, they cannot cause the protocol π' to run for more than $O(I)$ rounds, and thus they cannot make it reveal more than $O(I)$ bits of information. Unfortunately, the recent result of [71] asserts that there are I-bit information protocols which cannot be simulated by less than than $2^{\Omega(I)}$ bits of communication, and therefore this approach does not work.

We adapt our odometer construction to get a generic (black-box) conversion from a low-information protocol in the honest-but-curious model to a low-information protocol for the *adversarial* model. The basic premise is simple: we would like to maintain an estimate on the amount of information revealed so far, and abort if this number exceeds, say, $10I$. This plan is complicated by the fact that the dishonest party (say Bob) may try to attack this process in various ways. Firstly, he can try to fool the odometer into thinking that he learns less information than he actually does. Secondly, and perhaps more importantly, Bob can try to use the odometer itself to learn additional information about X. In particular, if it is Bob's turn to select the variable Z discussed above, Bob may cheat and select Z adversarially to elicit information from Alice. We modify the odometer protocol so that such cheating *can only hasten the termination of the simulation* (and cause Bob learn less information). We note that in our simulation Alice does not try to enforce Bob's compliance; rather, we just guarantee that the odometer has a proper estimate on what Bob learned so far, and thus it allows us to terminate once too much information has

been revealed. Our conversion result postulates that Alice and Bob share the knowledge of a prior distribution μ of their inputs (information-theoretic quantities are meaningless without an underlying prior). We believe that these results can be generalized to the *prior-free* setting using techniques similar to the ones used to define prior-free information complexity in [28].

We show how Alice can use the information odometer, in a black-box fashion, to achieve (the best possible) information-theoretic security against an *arbitrarily* (computationally unbounded) malicious Bob. More specifically, we prove:

Theorem 7.1.1 (Privacy-preserving simulation, informally stated)**.** *Let θ be a two-party communication protocol such that $\mathsf{IC}(\theta) = I$. Then for any $\delta > 0$, there is a communication protocol $\tilde{\pi}$ using "live" randomness, with the following properties:*

- *If both parties are honest, then $\tilde{\pi}$ 2δ-simulates θ.*
- $\mathsf{IC}(\tilde{\pi}) \leq O(I + \log(\|\theta\|))$.
- *There is a global constant $\lambda > 0$ such that for any protocol $\tilde{\pi}'$ where at least one party is honest (follows $\tilde{\pi}$), the following holds: $\forall\ k \in \mathcal{N}$,*

$$\Pr[\textit{ Honest party reveals more than } \lambda k(I/\delta + \log(\|\theta\| + 1)) \textit{ bits of information }] \leq 2^{-\Omega(k)}.$$

That is, an honest player never reveals to the other party much more than the essential amount of information required to solve f. We stress that *the protocol does not assume any prior knowledge about the honesty of any player*.

Due to space constraints, we defer the formal proof of Theorem 7.1.1 to the full version of the paper [40].

7.2 Conclusion and Open Problems

The full resolution of many of the problems addressed in this monograph are beyond our reach. We henceforth list some of these interesting and important open questions.

The computability of information complexity and its rate of convergence in the number of rounds. An unsatisfactory state of affairs is that despite the characterization and understanding of the information complexity measure, it is still not known whether this measure is even computable. More precisely, we do not have an algorithm that given the truth table of a function $F(X,Y)$ calculates the (zero-error) information complexity of this function. Note we *can* compute the communication complexity of F^n for any n, and we have $\frac{\mathsf{CC}(F^n)}{n} \searrow \mathsf{IC}^{\mathsf{ext}}(F)$, which gives us a sequence which decreases down to $\mathsf{IC}^{\mathsf{ext}}(F)$, but we do not have a similar sequence of lower bound. Figuring out the rate of convergence of the bounded-round information complexity $\mathsf{IC}^{\mathsf{ext}}{}^r_\mu(f)$ to $\mathsf{IC}^{\mathsf{ext}}{}_\mu(f)$, or at least an upper bound on it, would give a stopping criteria and therefore is sufficient for the computability of $\mathsf{IC}^{\mathsf{ext}}(F)$.

The rate of convergence of $\mathsf{IC}^{\mathsf{ext}}{}_r(F) \searrow \mathsf{IC}^{\mathsf{ext}}(F)$ is a very interesting question in its own right. The question is about the usefulness of additional rounds in giving an information-theoretically efficient protocol for F, and equivalently whether extra rounds of communication are useful for computing n copies of F for large n. We showed that in the case of $F = AND$, the rate of convergence is $1/r^2$. We conjecture that this is always the right rate, except when full convergence happens within a fixed number of rounds.

Conjecture 7.2.1. *For all $F(X,Y)$ one of the two scenarios hold: (1) $\mathsf{IC}^{\mathsf{ext}}{}_r(F) = \mathsf{IC}^{\mathsf{ext}}(F)$ for some $r = r(F)$; or (2) $\mathsf{IC}^{\mathsf{ext}}{}_r(F) - \mathsf{IC}^{\mathsf{ext}}(F) = \Theta_F(1/r^2)$.*

As we've seen in Chapter 2, the AND function exhibits the second behavior. An example of the first behavior is the single-bit transmission function $F(X,Y) = X$. Its

information cost is $\mathsf{IC}^{\mathsf{ext}}{}_r(F) = \mathsf{IC}^{\mathsf{ext}}(F) = 1$ for all $r \geq 1$.

Multi-party information complexity. An extremely interesting and potentially gratifying direction is developing the "right" notions of information complexity for the Number-On-Forehead multi-party communication complexity model. There are examples where information-theoretic methods were successfully applied to multiparty number-in-hand communication [47]. However, it is not clear whether (and how) similar techniques can apply to the number-on-the-forehead model. One obstacle here is the existence of *private multi-party protocols* that allow three or more parties to evaluate a function of their inputs while only learning the value of the function [22].

A XOR lemma for communication complexity. In Chapter 3 we proved direct product theorems which assert a lower bound on computing n independent copies of f in terms of the cost of a single copy. When n is very large, such theorems can be superseded by trivial arguments, since f^n must require at least n bits of communication just to describe the output. One could hope to achieve hardness amplification without blowing up the output size – a classical example is YAO's XOR lemma in circuit complexity. In light of the state-of-the-art direct product result, we state the following conjecture:

Conjecture 7.2.2 (XOR Lemma for communication complexity)**.**

$$\mathsf{D}_{\mu^n}(f^{\oplus n}, 1/2 + e^{-\Omega(n)}) = \tilde{\Omega}(\sqrt{n}) \cdot \mathsf{D}_{\mu}(f, 2/3)$$

where $f^{\oplus n}((x_1, y_1), \ldots, (x_n, y_n)) := f(x_1, y_1) \oplus .... \oplus f(x_n, y_n)$.

We remark that the analogues "direct-sum" of this conjecture is true: [16] proved that their direct sum result for f^n can be easily extended to the computation of $f^{\oplus n}$, showing (roughly) that $\mathsf{D}_{\mu^n}(f^{\oplus n}, 3/4) = \tilde{\Omega}(\sqrt{n}) \cdot \mathsf{D}_{\mu}(f, 2/3)$. However, this conversion technique

does not apply to the direct product setting.

Approximate matchings and round-Communication tradeoffs in welfare maximization. There are many open problems related to our result from Chapter 6. Let us mention a few of the most natural ones. Our first open problem is closing the gap between our lower bound (Theorem 6.2.1) and the upper bound of [62]: We show that $r = \Omega(\log \log n)$ rounds of communication are required to achieve constant approximation ratio using poly-logarithmic bits per player, while the upper bound is $r = O(\log n)$. We believe that the upper bound is in fact tight, and improving the lower bound is left as our first and direct open problem.

Another interesting direction is trying to extend our lower bound technique to obtain similar-in-spirit round-communication tradeoffs for the more general setup of combinatorial auctions, also studied by [62]. From a communication complexity perspective, lower bounds in this setup are more compelling, since player valuations require *exponentially* many bits to encode, hence interaction has the potential to reduce the overall communication (required to obtain efficient allocations) from exponential to polynomial. Indeed, it is shown in [62] that, in the case of *sub-additive bidders*, there is an r-round randomized protocol that obtains an $\tilde{O}(r \cdot m^{1/(r+1)})$-approximation to the optimal social welfare, where in each round each player sends $poly(m, n)$ bits. Once again, an (exponential in m) lower bound was given only for the case of simultaneous protocols ($r = 1$) and the natural question is to extend it to multiple rounds as well.

A more general open problem advocated by [62] is to analyze the communication complexity of finding an *e*xact optimal matching. One may naturally conjecture that $n^{\Omega(1)}$ rounds of interaction are required for this if each player only sends $n^{o(1)}$ bits each round, but no super-logarithmic bound is known. The communication complexity of the problem without any limitation on the number of rounds is also open: no significantly super linear, $\omega(n \log n)$, bound is known, while the best upper bound known is $\tilde{O}(n^{3/2})$.

Information complexity has proved to be a powerful tool for proving strong bounds in many computational models, through the communication complexity lens. We believe that there are many more potential applications to be explored. A partial list includes lower bounds for the Private Information Retrieval problem (PIR) and for Secret-Sharing, applications to Differential Privacy and to mechanism design.

Bibliography

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

[2] Anil Ada, Arkadev Chattopadhyay, Stephen A. Cook, Lila Fontes, Michal Kouck?, and Toniann Pitassi. The hardness of being private. In *IEEE Conference on Computational Complexity*, pages 192–202. IEEE, 2012.

[3] Noga Alon. A simple algorithm for edge-coloring bipartite multigraphs. *Inf. Process. Lett.*, 85(6):301–302, March 2003.

[4] Noga Alon, Irit Dinur, Ehud Friedgut, and Benny Sudakov. Graph products, fourier analysis and spectral techniques. *Geometric and Functional Analysis GAFA*, 14(5):913–940, 2004.

[5] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

[6] Alexandr Andoni. High frequency moment via max stability. Available at http://web.mit.edu/andoni/www/papers/fkStable.pdf.

[7] Alexandr Andoni, Robert Krauthgamer, and Krzysztof Onak. Streaming algorithms from precision sampling. *CoRR*, abs/1011.1263, 2010.

[8] Alexandr Andoni, Huy L. Nguyên, Yury Polyanskiy, and Yihong Wu. Tight lower bound for linear sketches of moments. In *ICALP (1)*, pages 25–32, 2013.

[9] Alexandr Andoni, Huy Le Nguyen, Yury Polyanskiy, and Yihong Wu. Tight lower bound for linear sketches of moments. In *ICALP*, 2013.

[10] Gilad Asharov and Yehuda Lindell. A full proof of the bgw protocol for perfectly-secure multiparty computation. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:36, 2011.

[11] Moshe Babaioff, Robert Kleinberg, and Renato Paes Leme. Optimal mechanisms for selling information. In *EC*, pages 92–109, 2012.

[12] R. Bar-Yehuda, B. Chor, E. Kushilevitz, and A. Orlitsky. Privacy, additional information and communication. *Information Theory, IEEE Transactions on*, 39(6):1930–1943, 1993.

[13] Reuven Bar-Yehuda, Benny Chor, Eyal Kushilevitz, and Alon Orlitsky. Privacy, additional information, and communication. *IEEE Transactions on Information Theory*, 39:55–65, 1993.

[14] 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.

[15] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci.*, 68(4):702–732, 2004.

[16] 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.

[17] Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. How to compress interactive communication. In *STOC*, pages 67–76, 2010.

[18] Balthazar Bauer, Shay Moran, and Amir Yehudayoff. Internal compression of protocols to entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 21:101, 2014.

[19] Donald Beaver. Perfect privacy for two-party protocols. In J. Feigenbaum and M. Merritt, editors, *Proceedings of DIMACS Workshop on Distributed Computing and Cryptology*, volume 2, pages 65–77. American Mathematical Society, 1989.

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

[21] M. Ben-Or, S. Goldwasser, and A. Wigderson. Completeness theorems for non-cryptographic fault-tolerant distributed computation. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, pages 1–10. ACM, 1988.

[22] Michael Ben-Or, Shafi Goldwasser, Joe Kilian, and Avi Wigderson. Multi-prover interactive proofs: How to remove intractability assumptions. In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing*, pages 113–131, 1988.

[23] Michael Ben-Or, Shafi Goldwasser, and Avi Wigderson. Completeness theorems for non-cryptographic fault-tolerant distributed computation (extended abstract). In *STOC*, pages 1–10, 1988.

[24] Lakshminath Bhuvanagiri, Sumit Ganguly, Deepanjan Kesh, and Chandan Saha. Simpler algorithm for estimating frequency moments of data streams. In *SODA*, pages 708–713, 2006.

[25] Eric Blais, Joshua Brody, and Kevin Matulef. Property testing lower bounds via communication complexity. *Computational Complexity*, 21(2):311–358, 2012.

[26] Maria Luisa Bonet and Samuel R. Buss. Size-depth tradeoffs for boolean fomulae. *Inf. Process. Lett.*, 49(3):151–155, 1994.

[27] Mark Braverman. Interactive information complexity. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:123, 2011.

[28] 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.

[29] 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.

[30] Mark Braverman, Ankit Garg, Denis Pankratov, and Omri Weinstein. From information to exact communication. *Electronic Colloquium on Computational Complexity (ECCC)*, 19(171), 2012.

[31] Mark Braverman, Ankit Garg, Denis Pankratov, and Omri Weinstein. From information to exact communication. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 151–160, New York, NY, USA, 2013. ACM.

[32] Mark Braverman, Ankit Garg, Denis Pankratov, and Omri Weinstein. Information lower bounds via self-reducibility. In *CSR*, pages 183–194, 2013.

[33] Mark Braverman and Ankur Moitra. An information complexity approach to extended formulations. In *STOC*, pages 161–170, 2013.

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

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

[36] Mark Braverman, Anup Rao, Omri Weinstein, and Amir Yehudayoff. Direct products in communication complexity. *Electronic Colloquium on Computational Complexity (ECCC)*, 19:143, 2012.

[37] Mark Braverman, Anup Rao, Omri Weinstein, and Amir Yehudayoff. Direct product via round-preserving compression. *Electronic Colloquium on Computational Complexity (ECCC)*, 20:35, 2013.

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

[39] Mark Braverman and Omri Weinstein. A discrepancy lower bound for information complexity. In *APPROX-RANDOM*, pages 459–470, 2012.

[40] Mark Braverman and Omri Weinstein. An interactive information odometer with applications. *Electronic Colloquium on Computational Complexity (ECCC)*, 21:47, 2014.

[41] Vladimir Braverman, Jonathan Katzman, Charles Seidell, and Gregory Vorsanger. An optimal algorithm for large frequency moments using o(nˆ(1-2/k)) bits. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2014, September 4-6, 2014, Barcelona, Spain*, pages 531–544, 2014.

[42] Vladimir Braverman and Rafail Ostrovsky. Recursive sketching for frequency moments. *CoRR*, abs/1011.2571, 2010.

[43] Vladimir Braverman and Rafail Ostrovsky. Approximating large frequency moments with pick-and-drop sampling. *CoRR*, abs/1212.0202, 2012.

[44] Richard P. Brent. The parallel evaluation of general arithmetic expressions. *J. ACM*, 21(2):201–206, 1974.

[45] Joshua Brody, Harry Buhrman, Michal Koucký, Bruno Loff, Florian Speelman, and Nikolay K. Vereshchagin. Towards a reverse newman's theorem in interactive information complexity. In *Proceedings of the 28th Conference on Computational Complexity, CCC 2013, K.lo Alto, California, USA, 5-7 June, 2013*, pages 24–33, 2013.

[46] Joshua Brody, Amit Chakrabarti, Ranganath Kondapally, David P. Woodruff, and Grigory Yaroslavtsev. Certifying equality with limited interaction. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2014, September 4-6, 2014, Barcelona, Spain*, pages 545–581, 2014.

[47] A. Chakrabarti, S. Khot, and X. Sun. Near-optimal lower bounds on the multi-party communication complexity of set disjointness. In *Computational Complexity, 2003. Proceedings. 18th IEEE Annual Conference on*, pages 107–117. IEEE, 2003.

[48] A. Chakrabarti and O. Regev. An optimal lower bound on the communication complexity of gap-hamming-distance. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, pages 51–60. ACM, 2011.

[49] Amit Chakrabarti, Subhash Khot, and Xiaodong Sun. Near-optimal lower bounds on the multi-party communication complexity of set disjointness. In *IEEE Conference on Computational Complexity*, pages 107–117, 2003.

[50] Amit Chakrabarti, Ranganath Kondapally, and Zhenghui Wang. Information complexity versus corruption and applications to orthogonality and gap-hamming. *CoRR*, abs/1205.0968, 2012.

[51] Amit Chakrabarti, Ranganath Kondapally, and Zhenghui Wang. Information complexity versus corruption and applications to orthogonality and gap-hamming. In *APPROX-RANDOM*, pages 483–494, 2012.

[52] 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.

[53] B. Chor and E. Kushilevitz. A zero-one law for boolean privacy. In *Proceedings of the twenty-first annual ACM symposium on Theory of computing*, pages 62–72. ACM, 1989.

[54] Benny Chor and Eyal Kushilevitz. A zero-one law for boolean privacy. *STOC 89 and SIAM J. Disc. Math*, 4:36–47, 1991.

[55] Kenneth L. Clarkson, Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, Xiangrui Meng, and David P. Woodruff. The fast cauchy transform and faster robust linear regression. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pages 466–477, 2013.

[56] Kenneth L Clarkson and David P Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, pages 205–214. ACM, 2009.

[57] Don Coppersmith and Ravi Kumar. An improved data stream algorithm for frequency moments. In *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 151–156, 2004.

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

[59] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, 1991.

[60] Gabrielle Demange, David Gale, and Marilda Sotomayor. Multi-item auctions. *The Journal of Political Economy*, pages 863–872, 1986.

[61] Shahar Dobzinski and Noam Nisan. Limitations of vcg-based mechanisms. *Combinatorica*, 31(4):379–396, 2011.

[62] Shahar Dobzinski, Noam Nisan, and Sigal Oren. Economic efficiency requires interaction. In *STOC*, pages 233–242, 2014.

[63] Jeff Edmonds, Russell Impagliazzo, Steven Rudich, and Jiri Sgall. Communication complexity towards lower bounds on circuit depth. *Computational Complexity*, 10(3):210–246, 2001.

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

[65] Joan Feigenbaum, Aaron D. Jaggard, and Michael Schapira. Approximate privacy: Foundations and quantification (extended abstract). In *Proceedings of the 11th ACM Conference on Electronic Commerce*, EC '10, pages 167–178, New York, NY, USA, 2010. ACM.

[66] Sumit Ganguly. Estimating frequency moments of data streams using random linear combinations. In *Proceedings of the 8th International Workshop on Randomization and Computation (RANDOM)*, pages 369–380, 2004.

[67] Sumit Ganguly. A hybrid algorithm for estimating frequency moments of data streams, 2004. Manuscript.

[68] Sumit Ganguly. Lower bounds on frequency estimation of data streams (extended abstract). In *CSR*, pages 204–215, 2008.

[69] Sumit Ganguly. Polynomial estimators for high frequency moments. *CoRR*, abs/1104.4552, 2011.

[70] Sumit Ganguly. A lower bound for estimating high moments of a data stream. *CoRR*, abs/1201.0253, 2012.

[71] Anat Ganor, Gillat Kol, and Ran Raz. Exponential separation of information and communication. *Electronic Colloquium on Computational Complexity (ECCC)*, 21:49, 2014.

[72] Dmitry Gavinsky, Or Meir, Omri Weinstein, and Avi Wigderson. Toward better formula lower bounds: An information complexity approach to the krw composition conjecture. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, pages 213–222, New York, NY, USA, 2014. ACM.

[73] Ashish Goel, Michael Kapralov, and Sanjeev Khanna. On the communication and streaming complexity of maximum bipartite matching. In *SODA*, pages 468–485, 2012.

[74] O. Goldreich, S. Micali, and A. Wigderson. How to play any mental game or a completeness theorem for protocols with honest majority. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, pages 218–229, New York, NY, USA, 1987. ACM.

[75] D. Greenwell and L. Lovász. Applications of product colouring. *Acta Mathematica Hungarica*, 25(3):335–340, 1974.

[76] Michelangelo Grigni and Michael Sipser. Monotone separation of logspace from nc. In *Structure in Complexity Theory Conference'91*, pages 294–298, 1991.

[77] Venkatesan Guruswami and Krzysztof Onak. Superlinear lower bounds for multipass graph processing. In *IEEE Conference on Computational Complexity*, pages 287–298, 2013.

[78] Johan Håstad. The shrinkage exponent of de morgan formulas is 2. *SIAM J. Comput.*, 27(1):48–64, 1998.

[79] Johan Håstad, Stasys Jukna, and Pavel Pudlák. Top-down lower bounds for depth-three circuits. *Computational Complexity*, 5(2):99–112, 1995.

[80] Johan Håstad and Avi Wigderson. The randomized communication complexity of set disjointness. *Theory Of Computing*, 3:211–219, 2007.

[81] Johan Håstad and Avi Wigderson. The randomized communication complexity of set disjointness. *Theory of Computing*, 3(1):211–219, 2007.

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

[83] John E Hopcroft and Richard M Karp. An nˆ5/2 algorithm for maximum matchings in bipartite graphs. *SIAM Journal on computing*, 2(4):225–231, 1973.

[84] Zengfeng Huang, Bozidar Radunovic, Milan Vojnovic, and Qin Zhang. Communication complexity of approximate maximum matching in distributed graph data. Technical Report MSR-TR-2013-35, April 2013.

[85] Zengfeng Huang, Bozidar Radunovic, Milan Vojnovic, and Qin Zhang. Communication complexity of approximate maximum matching in distributed graph data. *Microsoft Technical Report, MSR-TR-2013-35*, 2013.

[86] D.A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

[87] D.A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

[88] Johan Hstad and Avi Wigderson. Composition of the universal relation. In *ADVANCES IN COMPUTATIONAL COMPLEXITY THEORY, AMS-DIMACS*, 1993.

[89] P. Indyk and D. Woodruff. Optimal approximations of the frequency moments of data streams. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 202–208. ACM, 2005.

[90] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM*, 53(3):307–323, 2006.

[91] Piotr Indyk and David Woodruff. Tight lower bounds for the distinct elements problem. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '03, pages 283–, Washington, DC, USA, 2003. IEEE Computer Society.

[92] Rahul Jain. New strong direct product results in communication complexity. 2011.

[93] Rahul Jain and Hartmut Klauck. The partition bound for classical communication complexity and query complexity. *CoRR*, abs/0910.4266, 2009.

[94] Rahul Jain, Attila Pereszlenyi, and Penghui Yao. A direct product theorem for the two-party bounded-round public-coin communication complexity. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 167–176. IEEE, 2012.

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

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

[97] T. S. Jayram and David P. Woodruff. Optimal bounds for johnson-lindenstrauss transforms and streaming problems with subconstant error. *ACM Transactions on Algorithms*, 9(3):26, 2013.

[98] Daniel M. Kane, Jelani Nelson, and David P. Woodruff. On the exact space complexity of sketching and streaming small norms. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pages 1161–1178, 2010.

[99] Mauricio Karchmer, Ran Raz, and Avi Wigderson. Super-logarithmic depth lower bounds via the direct sum in communication complexity. *Computational Complexity*, 5(3/4):191–204, 1995. Prelim version CCC 1991.

[100] Mauricio Karchmer and Avi Wigderson. Monotone circuits for connectivity require super-logarithmic depth. In *STOC*, pages 539–550, 1988.

[101] Richard M Karp, Eli Upfal, and Avi Wigderson. Constructing a perfect matching is in random nc. In *Proceedings of the seventeenth annual ACM symposium on Theory of computing*, pages 22–32. ACM, 1985.

[102] 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.

[103] Iordanis Kerenidis, Sophie Laplante, Virginie Lerays, Jérémie Roland, and David Xiao. Lower bounds on information complexity via zero-communication protocols and applications. *CoRR*, abs/1204.1505, 2012.

[104] Iordanis Kerenidis, Mathieu Laurière, and David Xiao. New lower bounds for privacy in communication protocols. In *ICITS*, pages 69–89, 2013.

[105] V. M. Khrapchenko. A method of obtaining lower bounds for the complexity of π-schemes. *Mathematical Notes Academy of Sciences USSR*, 10:474–479, 1972.

[106] Hartmut Klauck. On quantum and approximate privacy. In *STACS*, pages 335–346, 2002.

[107] Hartmut Klauck. Quantum and approximate privacy. *Theory Comput. Syst.*, 37(1):221–246, 2004.

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

[109] Swastik Kopparty. List-decoding multiplicity codes. *Electronic Colloquium on Computational Complexity (ECCC)*, 19:44, 2012.

[110] Eyal Kushilevitz. Privacy and communication complexity. *SIAM J. Discrete Math.*, 5(2):273–284, 1992.

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

[112] Eyal Kushilevitz and Noam Nisan. *Communication complexity*. Cambridge University Press, New York, 1997. 96012840 96012840 Eyal Kushilevitz, Noam Nisan.

[113] Rafal Latala. Estimation of moments of sums of independent real random variables. *The Annals of Probability*, 25(3):1502–1513, 07 1997.

[114] T. Lee, A. Shraibman, and R. Spalek. A direct product theorem for discrepancy. In *Computational Complexity, 2008. CCC'08. 23rd Annual IEEE Conference on*, pages 71–80. IEEE, 2008.

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

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

[117] Yi Li, Huy L. Nguyen, and David P. Woodruff. Turnstile streaming algorithms might as well be linear sketches. In *STOC*, pages 174–183, 2014.

[118] Yi Li and David P. Woodruff. A tight lower bound for high frequency moment estimation with small error. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 16th International Workshop, APPROX 2013, and 17th International Workshop, RANDOM 2013, Berkeley, CA, USA, August 21-23, 2013. Proceedings*, pages 623–638, 2013.

[119] Zvi Lotker, Boaz Patt-Shamir, and Seth Pettie. Improved distributed approximate matching. In *Proceedings of the twentieth annual symposium on Parallelism in algorithms and architectures*, pages 129–136. ACM, 2008.

[120] N. Ma and P. Ishwar. Two-terminal distributed source coding with alternating messages for function computation. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 51–55. IEEE, 2008.

[121] Nan Ma and Prakash Ishwar. Infinite-message distributed source coding for two-terminal interactive computing. In *Proc. of the 47th annual Allerton Conf. on Comm., Control, and Comp.*, Allerton'09, pages 1510–1517, Piscataway, NJ, USA, 2009. IEEE Press.

[122] D. Malkhi, N. Nisan, B. Pinkas, and Y. Sella. Fairplay - a secure two-party computation system. In *Proceedings of the 13th USENIX Security Symposium*, pages 287–302, Berkeley, CA, USA. USENIX Association.

[123] Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil P. Vadhan. The limits of two-party differential privacy. In *FOCS*, pages 81–90, 2010.

[124] P.B. Miltersen, N. Nisan, S. Safra, and A. Wigderson. On data structures and asymmetric communication complexity. In *Proceedings of the twenty-seventh annual ACM symposium on Theory of computing*, pages 103–111. ACM, 1995.

[125] 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.

[126] Morteza Monemizadeh and David P. Woodruff. 1-pass relative-error l_p-sampling with applications. In *SODA*, 2010.

[127] Ketan Mulmuley, Umesh V Vazirani, and Vijay V Vazirani. Matching is as easy as matrix inversion. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, pages 345–354. ACM, 1987.

[128] 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.

[129] Mihai Patrascu. Towards polynomial lower bounds for dynamic problems. In *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, STOC '10, pages 603–610, New York, NY, USA, 2010. ACM.

[130] Mihai Patrascu and Ryan Williams. On the possibility of faster sat algorithms. In Moses Charikar, editor, *SODA*, pages 1065–1075. SIAM, 2010.

[131] Benny Pinkas. Fair secure two-party computation. In *EUROCRYPT*, pages 87–105, 2003.

[132] Eric Price and David P. Woodruff. Applications of the shannon-hartley theorem to data streams and sparse recovery.

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

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

[135] Ran Raz and Avi Wigderson. Monotone circuits for matching require linear depth. *J. ACM*, 39(3):736–744, 1992.

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

[137] Alexander A. Razborov. On the distributional complexity of disjointness. *Theor. Comput. Sci.*, 106(2):385–390, 1992.

[138] Mert Saglam and Gábor Tardos. On the communication complexity of sparse set disjointness and exists-equal problems. *CoRR*, abs/1304.1217, 2013.

[139] Michael E. Saks and Xiaodong Sun. Space lower bounds for distance approximation in the data stream model. In *STOC*, pages 360–369. ACM, 2002.

[140] R. Shaltiel. Towards proving strong direct product theorems. *Computational Complexity*, 12(1):1–22, 2003.

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

[142] Adi Shamir. Ip=pspace. In *FOCS*, pages 11–15, 1990.

[143] A.A. Sherstov. The communication complexity of gap hamming distance. *Theory of Computing*, 8:197–208, 2012.

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

[145] D V Smirnov. Shannon's information methods for lower bounds for probabilistic communication complexity. Master's thesis, Moscow State University, 1988.

[146] Christian Sohler and David P. Woodruff. Subspace embeddings for the l_1-norm with applications. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 755–764, 2011.

[147] Philip M. Spira. On time-hardware complexity tradeoffs for boolean functions. In *Proceedings of the Fourth Hawaii International Symposium on System Sciences*, pages 525–527, 1971.

[148] Emanuele Viola. The communication complexity of addition. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:152, 2011.

[149] Juraj Waczulk. Area time squared and area complexity of vlsi computations is strongly unclosed under union and intersection. In Jrgen Dassow and Jozef Kelemen, editors, *Aspects and Prospects of Theoretical Computer Science*, volume 464 of *Lecture Notes in Computer Science*, pages 278–287. Springer Berlin Heidelberg, 1990.

[150] Ingo Wegener. *The complexity of Boolean functions*. Wiley-Teubner, 1987.

[151] Virginia Vassilevska Williams. Multiplying matrices faster than coppersmith-winograd. In *Proceedings of the Forty-fourth Annual ACM Symposium on Theory of Computing*, STOC '12, pages 887–898, New York, NY, USA, 2012. ACM.

[152] David P. Woodruff. Optimal space lower bounds for all frequency moments. In *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 167–175, 2004.

[153] David P. Woodruff and Qin Zhang. Tight bounds for distributed functional monitoring. In *Proceedings of the 44th symposium on Theory of Computing*, STOC '12, pages 941–960, 2012.

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

[155] Andrew Chi-Chih Yao. Theory and applications of trapdoor functions (extended abstract). In *FOCS*, pages 80–91. IEEE, 1982.

[156] Raphael Yuster. Maximum matching in regular and almost regular graphs. *Algorithmica*, 66(1):87–92, 2013.