INFORMATION THEORETIC RELAXATIONS IN COMPLEXITY THEORY

ANKIT GARG

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

SEPTEMBER, 2016

Abstract

Since Shannon's "A Mathematical Theory of Communication" [Sha48], information theory has found applicability in a wide range of scientific disciplines. Over the past two decades, information theory has reemerged in theoretical computer science as a mathematical tool with applications to streaming algorithms, data structures, communication complexity etc. Properties of mutual information such as additivity and chain rule play an important role in these applications. In this thesis, we apply information theoretic tools to study various problems in complexity theory. These include the study of information complexity and communication complexity [BGPW13a, BGPW13c, BG14], hardness amplification of 2-prover games [BG15], applications of quantum information complexity to the study of quantum communication complexity of disjointness [BGK$^+$15] and the use of strong data processing inequalities to analyze communication complexity of distributed statistical estimation [GMN14, BGM$^+$16]. Along the way, we also develop several information theoretic tools such as correlated sampling theorems, subadditivity properties of information and quantum information cost etc. which could be of independent interest.

Acknowledgements

I have been extremely fortunate to be advised by Mark Braverman. He has been a huge inspiration and has had a profound effect on my way of thinking and approaching research. It is very interesting to see his deeply intuitive approach on the one hand and the technical "hands-on" approach on the other. I will always cherish the numerous interactions we have had over the years.

I would like to thank my collaborators, joint collaborations with whom constitute this thesis: Mark Braverman, Young Kun Ko, Tengyu Ma, Jieming Mao, Huy Nguyen, Denis Pankratov, Dave Touchette, Omri Weinstein and David Woodruff. It has been a real pleasure working with them and this thesis wouldn't be possible without them. I would also like to thank Bernard Chazelle, Zeev Dvir, Ran Raz and Avi Wigderson for devoting their time and agreeing to be on my committee. Special thanks to Avi, who is an inspiration for the whole theoretical computer science community. I have had the privilege of interacting and collaborating with Avi over the past few years and it has immensely helped me as a researcher.

I would also like to thank my friends over the years: Akshay, Chinmay, Dan, Debajit, Huy, Jon, Jieming, Mahim, Rafael, Ravi, Sharvanath, Sivakanth (usually called by his last name "Gopi"), Yogesh, Yonatan, Young (usually called by his last name "Ko") and several others, who made Princeton a fun place.

Thanks to all the Princeton University Staff, especially Nicki Gotsis, Mitra Kelly and

Melissa Lawson for their full support in various bureaucratic tasks over the years. I am also thankful to the Simons Foundation and Siebel Foundation for their generous scholarships. My research has been funded by Mark's grants from Simons Collaboration on Algorithms and Geometry, The David and Lucile Packard Foundation, The Alfred P. Sloan Foundation, The John Templeton Foundation and NSF grants CCF-1525342 and CCF-1149888.

Last, but not the least, I would thank my parents, my sister and my girlfriend for their support throughout these years and letting me go across the world while they patiently wait for my return. I cannot thank my Guruji enough, for his constant support and guidance. It is to him that I dedicate this thesis.

Organization and Main Results

We describe here the main results in this thesis and links to the chapters where they appear.

Background and Preliminaries

We start by describing some background and preliminaries about communication complexity, information theory and information complexity (both classical and quantum) in Chapter 1.

Exact Communication Bounds for Disjointness

In Chapter 2, 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: This leads to a tight (upper and lower bound) characterization of the communication complexity of the set intersection and set disjointness problems on subsets of $\{1, \ldots, n\}$.

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

Only preliminary results without proofs will be presented in this chapter. Full proofs can be found in [BGPW13b].

Information Lower Bounds via Self-reducibility

In Chapter 3, we use self-reduction methods to prove strong information lower bounds on two of the most studied functions in the communication complexity literature: Gap Hamming Distance (GHD) and Inner Product (IP). In our first result we affirm the conjecture that the information cost of GHD is linear even under the *uniform* distribution, which strengthens the $\Omega(n)$ bound recently shown by [KLL$^+$12a], and answers an open problem from [CKW12]. In our second result we prove that the information cost of IP_n is arbitrarily close to the trivial upper bound n as the permitted error tends to zero, again strengthening the $\Omega(n)$ lower bound recently proved by [BW12].

Our proofs demonstrate that self-reducibility makes the connection between information complexity and communication complexity lower bounds a two-way connection. Whereas numerous results in the past [CSWY01, BYJKS04, BBCR10] used information complexity techniques to derive new communication complexity lower bounds, we explore a generic way in which communication complexity lower bounds imply information complexity lower bounds *in a black-box manner*.

Public vs Private Coins in Information Complexity

In Chapter 4, we precisely characterize the role of private randomness in the ability of Alice to send a message to Bob while minimizing the amount of information revealed to him. We show that if using private randomness a message can be transmitted while revealing I bits of information, the transmission can be simulated without private coins using $I + \log I + O(1)$ bits of information. Moreover, we give an example where this bound is tight: at least $I + \log I - O(1)$ bits are necessary in some cases. Our example also shows that the one-round compression construction of Harsha et al. [HJMR07] cannot be improved.

Small Value Parallel Repetition

In Chapter 5, we prove a parallel repetition theorem for general games with value tending to 0. Previously Dinur and Steurer [DS14] proved such a theorem for the special case of projection games. We use information theoretic techniques in our proof. Our proofs also extend to the high value regime (value close to 1) and provide alternate proofs for the parallel repetition theorems of Holenstein [Hol07] and Rao [Rao08] for general and projection games respectively. We also extend the example of Feige and Verbitsky [FV02] to show that the small-value parallel repetition bound we obtain is tight. Our techniques are elementary in that we only need to employ basic information theory and discrete probability in the small-value parallel repetition proof.

Bounded-round Quantum Communication Complexity Lower Bounds for Disjointness

In Chapter 6, we prove a near optimal round-communication tradeoff for the two-party quantum communication complexity of disjointness. For protocols with r rounds, we prove a lower bound of $\tilde{\Omega}(n/r + r)$ on the communication required for computing disjointness of input size n, which is optimal up to logarithmic factors. The previous best lower bound was $\Omega(n/r^2 + r)$ due to Jain, Radhakrishnan and Sen [JRS03]. Along the way, we develop several tools for quantum information complexity, one of which is a lower bound for quantum information complexity in terms of the generalized discrepancy method. As a corollary, we get that the quantum communication complexity of any boolean function f is at most $2^{O(QIC(f))}$, where $QIC(f)$ is the prior-free quantum information complexity of f (with error 1/3).

Communication Complexity Lower Bounds for Statistical Estimation

In Chapter 7, we study the tradeoff between the statistical error and communication cost of distributed statistical estimation problems in high dimensions. In the distributed sparse Gaussian mean estimation problem, each of the m machines receives n data points from a d-dimensional Gaussian distribution with unknown mean θ which is promised to be k-sparse. The machines communicate by message passing and aim to estimate the mean θ. We provide a tight (up to logarithmic factors) tradeoff between the estimation error and the number of bits communicated between the machines. This directly leads to a lower bound for the distributed *sparse linear regression* problem: to achieve the statistical minimax error, the total communication is at least $\Omega(\min\{n, d\}m)$, where n is the number of observations that each machine receives and d is the ambient dimension. These lower results improve upon [Sha14, SD15] by allowing multi-round iterative communication model. We also give the first optimal simultaneous protocol in the dense case for mean estimation.

As our main technique, we prove a *distributed strong data processing inequality*, as a generalization of strong data processing inequalities, which might be of independent interest and useful for other problems.

Contents

Chapter 1

Background and Preliminaries

1.1 Communication Complexity

The two-party communication model was introduced by Yao [Yao79] in 1979. Communication complexity [Yao79] can be viewed as the generalization of transmission problems to general tasks performed by two (or more) parties over a communication channel. Communication complexity is much more general than one-way transmission, but unlike circuit complexity, it is still amenable to lower bounds proofs by a broad range of techniques [KN97]. Furthermore, communication complexity lower bounds have found many applications, for example in obtaining tight bounds on streaming algorithms and data structures. In addition, some of the most promising approaches for strong circuit lower bounds that appear viable, such as Karchmer-Wigderson games and ACC lower bounds [KRW95, BT91] involve communication complexity lower bounds. Thus, at the moment, developing tools in communication complexity is one of the most promising approaches for making progress within computational complexity. In this model, two parties, traditionally called Alice and Bob, are trying to collaboratively compute a known boolean function $f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$. Each party is computationally unbounded; however, Alice is only given input $x \in \mathcal{X}$ and Bob is only

given $y \in \mathcal{Y}$. In order to compute $f(x, y)$, Alice and Bob communicate in accordance with an agreed-upon communication protocol π. Protocol π specifies as a function of transmitted bits only whether the communication is over and, if not, who sends the next bit. Moreover, π specifies as a function of the transmitted bits and x the value of the next bit to be sent by Alice and similarly for Bob. The communication is over when *both parties* know the value of $f(x, y)$. The cost of the protocol π is the number of bits exchanged on the worst input. *The transcript* of a protocol is a concatenation of all the bits exchanged during the execution of the protocol.

There are several ways in which the deterministic communication model can be extended to include randomization. In the *public-coin model*, Alice and Bob have access to a shared random string r chosen according to some probability distribution. The only difference in the definition of a protocol is that now the protocol π specifies the next bit to be sent by Alice as a function of x, the already transmitted bits, and a random string r. Similarly for Bob. This process can also be viewed as the two players having an agreed-upon distribution on deterministic protocols. Then the players jointly sample a protocol from this distribution. In the *private-coin model*, Alice has access to a random string r_A hidden from Bob, and Bob has access to a random string r_B hidden from Alice. Public coins are considered to be part of the transcript of the protocol while private coins are not part of the transcript.

Definition 1.1.1 (Randomized Communication Complexity)**.** For a function $f : \mathcal{X} \times \mathcal{Y} \to Z$ and a parameter $\epsilon > 0$, $R_\epsilon(f)$ denotes the cost of the best randomized public coin protocol for computing f with error at most ϵ on *every* input. Sometimes, we will denote by $R_\epsilon^{\text{priv}}(f)$ and $R_\epsilon^{\text{pub}}(f)$, the randomized private-coin and public-coin communication complexities of f, respectively.

Definition 1.1.2 (Distributional Communication Complexity)**.** The *distributional communication complexity* of $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ with respect to a distribution μ on $\mathcal{X} \times \mathcal{Y}$ and

error tolerance $\epsilon > 0$ is the least cost of a deterministic protocol computing f with error probability at most ϵ when the inputs are sampled according to μ. It is denoted by $\mathcal{D}_\mu(f, \epsilon)$.

Observe that for the purpose of the communication complexity, once we allow public randomness, it makes no difference whether we permit the players to have private random strings or not. This is because the private random strings can be simulated by parts of the public random string. Thus for a protocol π we use $\Pi(x, y)$ to denote the concatenation of the transcript of π and the public randomness when the protocol runs on inputs (x, y). The worst-case number of bits transmitted in π is denoted by $\mathrm{CC}(\pi)$. For $i \in [\mathrm{CC}(\pi)]$ we write $\Pi_i(x, y)$ to denote the ith bit transmitted in Π on input (x, y) if it exists.

For the pre-1997 results on communication complexity see the excellent book by Kushilevitz and Nisan [KN97].

1.2 Information Theory

Information theory as the primary mathematical tool for analyzing communication was first discovered by Shannon in the late 1940's [Sha48]. In particular, Shannon introduced his entropy function $H(X)$ to measure the amount of information contained in a random variable X. Shannon's "source coding theorem" also known as the "noiseless coding theorem" postulates that in the limit the per-message cost of transmitting a stream of messages $x_1, x_2, \ldots$ independently distributed according to X is exactly $H(X)$.

In this section we briefly provide the essential information-theoretic concepts used throughout this thesis. For a thorough introduction to the area of information theory, the reader should consult a wonderful book by Cover and Thomas [CT91]. Unless stated otherwise, all logarithms will be base-2.

Definition 1.2.1. Let μ be a (discrete) probability distribution on sample space Ω. *Shannon*

entropy (or just *entropy*) of μ, denoted by $H(\mu)$, is defined as

$$H(\mu) := \sum_{\omega \in \Omega : \mu(\omega) > 0} \mu(\omega) \log \frac{1}{\mu(\omega)}$$

For a random variable A we shall write $H(A)$ to denote the entropy of the induced distribution on the range of A. The same also holds for other information-theoretic quantities appearing later in this section.

We will denote $H(B_p)$ by just $H(p)$, where B_p is the Bernoulli distribution with success probability p.

Definition 1.2.2. *Conditional entropy* of a random variable A conditioned on B is defined as

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

Conditional entropy measures the amount of uncertainty in the random variable A from the point of view of an observer who already knows B.

Fact 1.2.3. $H(A|B) = \mathbb{E}_{b \sim P_B} H(A|B = b)$.

Here P_B denotes the distribution of the random variable B.

Definition 1.2.4. The *mutual information* between two random variable A and B, denoted by $I(A; B)$ is defined as

$$I(A; B) := H(A) - H(A|B) = H(B) - H(B|A).$$

The *conditional mutual information* between A and B given C, denoted by $I(A; B|C)$, is defined as

$$I(A; B|C) := H(A|C) - H(A|BC) = H(B|C) - H(B|AC).$$

The mutual information term $I(A;B)$ measures the amount of information the random variable A carries about B (or vice versa). The conditional mutual information term $I(A;B|C)$ measure the amount of additional information that A carries about B for an observer who already knows C.

Fact 1.2.5. $I(A;B|C) = \mathbb{E}_{c\sim P_C} I(A;B|C=c)$.

One of the most important and useful properties of mutual information is the chain rule. It says that the amount of information two random variables A_1 and A_2 carry about B (from the point of view of C) can be broken into two parts: the amount of information A_1 has about B (from the point of view of C) + the additional information that A_2 carries about B (from the point of view of A_1 and C).

Fact 1.2.6 (Chain Rule)**.** *Let A_1, A_2, B, C be random variables. Then*

$$I(A_1A_2;B|C) = I(A_1;B|C) + I(A_2;B|A_1C).$$

Definition 1.2.7. Given two probability distributions μ_1 and μ_2 on the same sample space Ω such that $(\forall\omega\in\Omega)(\mu_2(\omega)=0 \Rightarrow \mu_1(\omega)=0)$, the *Kullback-Leibler Divergence* between is defined as

$$D(\mu_1||\mu_2) = \sum_{\omega\in\Omega} \mu_1(\omega)\log\frac{\mu_1(\omega)}{\mu_2(\omega)}.$$

The connection between the mutual information and the Kullback-Leibler divergence is provided by the following fact.

Fact 1.2.8. *For random variables $A, B,$ and C we have*

$$I(A;B|C) = \mathbb{E}_{b,c\sim P_{B,C}} D(P_{A|B=b,C=c}||P_{A|C=c}).$$

Definition 1.2.9. Let μ_1 and μ_2 be two probability distributions on the same sample space

Ω. *Total variation distance* (or statistical distance) is defined as

$$\|\mu_1 - \mu_2\| := \frac{1}{2}\sum_{\omega\in\Omega} |\mu_1(\omega) - \mu_2(\omega)|.$$

Observe that $\|\mu_1 - \mu_2\| = \max_{\mathcal{S}\subseteq\Omega} |\mu_1(\mathcal{S}) - \mu_2(\mathcal{S})|$.

Fact 1.2.10 (Data Processing Inequality)**.** *Let A, B, C be random variables on the same sample space, and let D be a probabilistic function of B only. Then we have*

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

Fact 1.2.11. *Let X, Y be two random variables whose joint distribution is $P_{X,Y}$. Then for any distribution μ, we have*

$$\mathbb{E}_{x\sim P_X}[D(P_{Y|X=x}||P_Y)] \leq \mathbb{E}_{x\sim X}[D(P_{Y|X=x}||\mu)]$$

Proof.

$$\begin{aligned}\mathbb{E}_{x\sim X}[D(P_{Y|X=x}||\mu)] - \mathbb{E}_{x\sim P_X}[D(P_{Y|X=x}||P_Y)] &= \mathbb{E}_{x\sim P_X}\left[\sum_y P_{Y|X=x}(y)\log\left(\frac{P_Y(y)}{\mu(y)}\right)\right] \\ &= D(P_Y||\mu) \\ &\geq 0\end{aligned}$$

□

Fact 1.2.12. *Let A, B, C, D be four random variables such that $I(D; B|AC) = 0$. Then $I(A; B|C) \geq I(A; B|CD)$*

Proof. Expand $I(A, D; B|C)$ via chain rule in two ways. □

Fact 1.2.13. *Let A, B, C, D be four random variables such that $I(A; C|BD) = 0$. Then $I(A; B|D) \geq I(A; C|D)$*

Proof. Expand $I(A; B, C|D)$ via chain rule in two ways. □

Fact 1.2.14. *Let A, B, C, D be random variables s.t. $I(A; D|C) = 0$. Then*

$$I(A; B|C, D) \geq I(A; B|C)$$

Proof. Expand $I(A; B, D|C)$ via chain rule in two ways. □

Fact 1.2.15. *Let A, B, C, D, E be random variables.If C, D determine E and $D \to CE \to AB$ is Markov chain, then*

$$I(A; B|CE) = I(A; B|CD)$$

Proof. $I(A; B|CD) = I(A; B|CDE)$, since C, D determine E. Now consider $I(A; BD|CE)$

$$I(A; BD|CE) = I(A; B|CE) + I(A; D|BCE) = I(A; B|CE)$$

Also

$$I(A; BD|CE) = I(A; D|CE) + I(A; B|CDE) = I(A; B|CDE)$$

which completes the proof. □

Fact 1.2.16 (Chain Rule for relative entropy)**.** *Let P_{V_1,V_2} and P_{U_1,U_2} be two bivariate distributions. Then*

$$D\left(P_{V_1,V_2}||P_{U_1,U_2}\right) = D\left(P_{V_1}||P_{U_1}\right) + \mathbb{E}_{v_1 \sim P_{V_1}} D\left(P_{V_2|V_1=v_1}||P_{U_2|U_1=v_1}\right)$$

Fact 1.2.17 (Convexity of relative entropy)**.** *Let P_1, P_2, Q_1, Q_2 be distributions and $\lambda \in [0, 1]$*

be a number. Then

$$D\left(\lambda P_1 + (1-\lambda)P_2 || \lambda Q_1 + (1-\lambda)Q_2\right) \leq \lambda D\left(P_1||Q_1\right) + (1-\lambda)D\left(P_2||Q_2\right)$$

Fact 1.2.18 (Pinsker's inequality)**.** *Let P, Q be two distributions. Then*

$$D(P||Q) \geq \frac{||P-Q||_1^2}{2\ln 2}$$

Here $||P-Q||_1$ is the l_1 distance between the distributions P and Q.

The following lemma is well known and is used often in proofs of parallel repetition theorems.

Lemma 1.2.19. *Let $P_{V_1,\ldots,V_n}$ and $P_{U_1,\ldots,U_n}$ be two distributions over some space $\mathcal{U}^n$. Also suppose that $P_{U_1,\ldots,U_n}$ is a product distribution i.e. $P_{U_1,\ldots,U_n}(u_1,\ldots,u_n) = P_{U_1}(u_1)\cdots P_{U_n}(u_n)$. Then*

$$\sum_{i=1}^{n} D\left(P_{V_i}||P_{U_i}\right) \leq D\left(P_{V_1,\ldots,V_n}||P_{U_1,\ldots,U_n}\right)$$

Proof. By the chain rule for relative entropy, we get that:

$$\begin{aligned}
D\left(P_{V_1,\ldots,V_n}||P_{U_1,\ldots,U_n}\right) &= \sum_{i=1}^{n} \mathbb{E}_{v_1,\ldots,v_{i-1}\sim P_{V_1,\ldots,V_{i-1}}} D\left(P_{V_i|V_1=v_1,\ldots,V_{i-1}=v_{i-1}}||P_{U_i|U_1=v_1,\ldots,U_{i-1}=v_{i-1}}\right) \\
&= \sum_{i=1}^{n} \mathbb{E}_{v_1,\ldots,v_{i-1}\sim P_{V_1,\ldots,V_{i-1}}} D\left(P_{V_i|V_1=v_1,\ldots,V_{i-1}=v_{i-1}}||P_{U_i}\right) \\
&\geq \sum_{i=1}^{n} D\left(P_{V_i}||P_{U_i}\right)
\end{aligned}$$

The second equality is because $P_{U_1,\ldots,U_n}$ is a product distribution. The inequality follows by convexity of relative entropy. □

Fact 1.2.20. *Let P_U be the distribution of some random variable U and let W be an arbitrary event. Then*

$$D\left(P_{U|W}||P_U\right) \leq \log(1/\Pr[W])$$

Proof.

$$\begin{aligned} D\left(P_{U|W}||P_U\right) &= \sum_u P_{U|W}(u)\log(P_{U|W}(u)/P_U(u)) \\ &\leq \sum_u P_{U|W}(u)\log(1/\Pr[W]) = \log(1/\Pr[W]) \end{aligned}$$

□

The following lemma is taken from [BRWY13b].

Fact 1.2.21 ([BRWY13b], Lemma 19)**.** *Suppose A, B, C are random variables s.t. $I(A;B|C) = 0$ and W be an arbitrary event. Then*

$$I(A;B|C,W) \leq \log(1/\Pr[W])$$

Proof.

$$\begin{aligned} I(A;B|C,W) &= \mathbb{E}_{b,c\sim P_{B,C|W}} D(P_{A|B=b,C=c,W}||P_{A|C=c,W}) \\ &\leq \mathbb{E}_{b,c\sim P_{B,C|W}} D(P_{A|B=b,C=c,W}||P_{A|C=c}) \\ &= \mathbb{E}_{b,c\sim P_{B,C|W}} D(P_{A|B=b,C=c,W}||P_{A|B=b,C=c}) \\ &\leq \log(1/\Pr[W]) \end{aligned}$$

First inequality is by Fact 1.2.11. Second equality is because of $I(A;B|C) = 0$. Second inequality is by Fact 1.2.20. □

The following lemma is used in a lot of information complexity papers.

Lemma 1.2.22. *Let P and Q be distributions over a universe $\mathcal{U}$. Let $\mathcal{B} = \{u : \frac{P(u)}{Q(u)} \geq 2^t\}$. Then*

$$P(\mathcal{B}) \leq \frac{D\left(P||Q\right)+1}{t}$$

Proof.

$$D(P||Q) = \sum_{u\in\mathcal{B}} P(u)\cdot\log(P(u)/Q(u)) + \sum_{u\notin\mathcal{B}} P(u)\cdot\log(P(u)/Q(u))$$
$$\geq P(\mathcal{B})\cdot t + \sum_{u\notin\mathcal{B}} P(u)\cdot\log(P(u)/Q(u)) \tag{1.1}$$

Denote the complement of $\mathcal{B}$ by $\bar{\mathcal{B}}$. Then

$$\sum_{u\notin\mathcal{B}} P(u)\cdot\log(P(u)/Q(u)) \geq P(\bar{\mathcal{B}})\log(P(\bar{\mathcal{B}})/Q(\bar{\mathcal{B}}))$$
$$\geq P(\bar{\mathcal{B}})\log(P(\bar{\mathcal{B}}))$$
$$> -1 \tag{1.2}$$

The first inequality follows from log-sum inequality. The second inequality is true because $Q(\bar{\mathcal{B}}) \leq 1$. The third inequality follows from the fact that $x\log(x) > -1$ for all $x \geq 0$. Now combining equations (1.1) and (1.2) completes the proof of the lemma. □

Fact 1.2.23. *Let P and Q be distributions over a universe $\mathcal{U}$. Suppose $\mathcal{V} \subseteq \mathcal{U}$ is such that $P(\mathcal{V}) = 1$. Then $Q(\mathcal{V}) \geq 2^{-D(P||Q)}$.*

Proof. It directly follows from the log-sum inequality. Denote the complement of $\mathcal{V}$ by $\bar{\mathcal{V}}$.

$$D(P||Q) = \sum_{u\in\mathcal{U}} P(u)\cdot\log(P(u)/Q(u)) \geq P(\mathcal{V})\cdot\log(P(\mathcal{V})/Q(\mathcal{V})) + P(\bar{\mathcal{V}})\cdot\log(P(\bar{\mathcal{V}})/Q(\bar{\mathcal{V}}))$$
$$= \log(1/Q(\mathcal{V}))$$

□

Lemma 1.2.24. *Let P and Q are two distributions s.t. $P \leq c\cdot Q$. Let $P_{Y|X}$ be a channel.*

Let $P_{X,Y} = P_X P_{Y|X}$ and $Q_{X,Y} = Q_X P_{Y|X}$ be two bivariate distributions. Then

$$I(X;Y)_P \leq c \cdot I(X;Y)_Q$$

Proof.

$$\begin{aligned} I(X;Y)_P &= \mathbb{E}_{x\sim P_X} D(P_{Y|X=x}||P_Y) \\ &\leq \mathbb{E}_{x\sim P_X} D(P_{Y|X=x}||Q_Y) \\ &\leq c \cdot \mathbb{E}_{x\sim Q_X} D(P_{Y|X=x}||Q_Y) \\ &= c \cdot I(X;Y)_Q \end{aligned}$$

First inequality is by Fact 1.2.11. Second inequality is because of $P \leq c \cdot Q$. □

Lemma 1.2.25. *Let $A, B_1, \ldots, B_n$ be random variables s.t. $B_1, \ldots, B_n$ are independent. Then*

$$I(A; B_1, \ldots, B_n) \geq \sum_{i=1}^{n} I(A; B_i)$$

Proof.

$$\begin{aligned} I(A; B_1, \ldots, B_n) &= \sum_{i=1}^{n} I(A; B_i | B_1, \ldots, B_{i-1}) \\ &\geq \sum_{i=1}^{n} I(A; B_i) \end{aligned}$$

The inequality follows from the independence of $B_1, \ldots, B_n$ and Fact 1.2.14. □

1.3 Information Complexity

One of the first applications of information theory to 2-party communication complexity appear in [Abl93] and [CSWY01]. Since then, the theory of information complexity has

been developed, which is a continuous relaxation of communication complexity and is more nicely behaved than communication complexity (e.g. information complexity is additive over computing multiple copies of a function). Information complexity has led to progress in several important problems in (randomized) communication complexity, such as the direct sum [BR11, BBCR10] and direct product questions [BRWY13b].

Despite information theory being so successful in reasoning about one-way communication, it took a while until information theory has been adopted into the communication complexity toolbox. Indeed, the first applications of information in communication complexity [Abl93, CSWY01] were in the context of one-way and simultaneous message communication complexity, which is most directly related to the classical transmission setting. It was not until the work of Bar-Yossef et al. [BYJKS04] that these techniques were extended to the two-way setting. Further developments [BBCR10, BR11, Bra12] showed that information-theoretic notions generalize nicely, at least to two-party communication complexity. One can define the information complexity of a task as the two-party analogue of Shannon's entropy. Shannon's entropy of a random variable X captures the amount of information contained in one sample – the least amount of information that needs to be conveyed to transmit an $x \sim X$. The *information complexity* of an interactive task T is the least amount of information about their inputs that Alice and Bob need to disclose to each other in order to perform T. Information complexity is similar to Shannon's entropy in that it captures exactly the amortized communication complexity of computing n independent copies of T over a noiseless binary channel as $n \to \infty$ [BR11]. Also, like Shannon's entropy, information complexity satisfies the direct sum property, i.e. it is additive: the information complexity of performing two independent tasks (T_1, T_2) is equal to the sum of the information complexities of T_1 and T_2 [BR11, Bra12].

Definition 1.3.1. The *internal information cost* of a protocol π with respect to a distribu-

tion μ on inputs from $\mathcal{X} \times \mathcal{Y}$ is defined as

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

Here $(X,Y) \sim \mu$ and $\Pi(X,Y)$ denotes the random variable for the transcript of the protocol π when run on inputs X and Y. While public randomness is part of the transcript Π, private randomness of Alice and Bob doesn't appear explicitly in the definition of information cost, it implicitly governs the conditional distribution of $\Pi(X,Y)|X,Y$. One can also write an expression where the private random strings R_A and R_B appear explicitly and that in fact is equal to the above expression.

Fact 1.3.2.

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

The external information cost of a protocol measures the amount of information that an external observer learns about the parties inputs from the protocol transcript.

Definition 1.3.3. The *external information cost* of π with respect to μ is

$$\mathrm{IC}^{\mathrm{ext}}_\mu(\pi) := I(\Pi(X,Y); XY).$$

It is essential to allow the players to have private randomness (as it can be used to hide information). Public randomness on the other hand can be easily simulated by private randomness without changing the information cost (one party can sample the public random string from his/her private randomness and send it across).

Lemma 1.3.4 ([BR11])**.** *For any distribution μ, $\mathsf{IC}_\mu(\pi) \leq CC(\pi)$.*

The *information complexity* of f with respect to μ is

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

where the infimum ranges over all (randomized) protocols π solving f with error at most ϵ when inputs are sampled according to μ. Note that we cannot replace the above quantifier with a min, since the information complexity of a function may not be achievable by any fixed (finite-round) protocol[1].

Similarly, the *external information complexity* of f with respect to μ is defined as

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

Braverman and Rao [BR11] gave an operational meaning to information complexity via their "information equals amortized communication" theorem. Let $D_{\epsilon}^{\mu^n}(f^n)$ denote the distributional communication complexity (under μ^n) of computing n copies of the function f, where the marginal error on each copy should be at most ϵ. Then the following holds:

Theorem 1.3.5 ([BR11])**.** *For all $\epsilon > 0$,*

$$\lim_{n\to\infty} \frac{D_{\epsilon}^{\mu^n}(f^n)}{n} = \mathrm{IC}_{\mu}(f,\epsilon)$$

1.4 Quantum Information Theory

We use the following notation for quantum theory; see [Wat13, Wil13] for more details. We associate a quantum register A with a corresponding vector space, also denoted by A. We only consider finite-dimensional vector spaces. A state of quantum register A is represented by a density operator $\rho \in \mathcal{D}(A)$, with $\mathcal{D}(A)$ the set of all unit trace, positive semi-definite linear operators mapping A into itself. We say that a state ρ is pure if it is a projection operator, i.e. $(\rho^A)^2 = \rho^A$. For a pure state ρ, we might use the pure state formalism, and represent ρ by the vector $|\rho\rangle$ it projects upon, i.e. $\rho = |\rho\rangle\langle\rho|$; this is well-defined up to an

[1]In fact, we shall see that this is the case for the AND function whose information complexity will be analyzed in Chapter 2

irrelevant phase factor.

A quantum channel from quantum register A into quantum register B is represented by an operator $\mathbb{N}^{A\to B} \in \mathcal{C}(A,B)$, with $\mathcal{C}(A,B)$ the set of all completely positive, trace-preserving linear operators from $\mathcal{D}(A)$ into $\mathcal{D}(B)$. If $A = B$, we might simply write $\mathbb{N}^A$, and when systems are clear from context, we might drop the superscripts. For channels $\mathbb{N}_1 \in \mathcal{C}(A,B), \mathbb{N}_2 \in \mathcal{C}(B,C)$, we denote their composition as $\mathbb{N}_2 \circ \mathbb{N}_1 \in \mathcal{C}(A,C)$, with action $(\mathbb{N}_2 \circ \mathbb{N}_1)(\rho) = \mathbb{N}_2(\mathbb{N}_1(\rho))$ on any state $\rho \in \mathcal{D}(A)$. We might drop the $\circ$ symbol if the composition is clear from context. For A and B isomorphic, we denote the identity mapping as $I^{A\to B}$, with some implicit choice for the change of basis. For $\mathbb{N}^{A_1\to B_1} \otimes I^{A_2\to B_2} \in \mathcal{C}(A_1 \otimes A_2, B_1 \otimes B_2)$, we might abbreviate this as $\mathbb{N}$ and leave the identity channel implicit when the meaning is clear from context.

An important subset of $\mathcal{C}(A,B)$ when A and B are isomorphic spaces is the set of unitary channels $\mathcal{U}(A,B)$, the set of all maps $U \in \mathcal{C}(A,B)$ with an adjoint map $U^\dagger \in \mathcal{C}(B,A)$ such that $U^\dagger \circ U = I^A$ and $U \circ U^\dagger = I^B$. More generally, if $\dim(B) \geq \dim(A)$, we denote by $\mathcal{U}(A,B)$ the set of isometric channels, i.e. the set of all maps $V \in \mathcal{C}(A,B)$ with an adjoint map $V^\dagger \in \mathcal{C}(B,A)$ such that $V^\dagger \circ V = I^A$. Another important example of channel that we use is the partial trace $\mathrm{Tr}_B(\cdot) \in C(A \otimes B, A)$ which effectively gets rid of the B subsystem to obtain the marginal state on subsystem A. Fixing an orthonormal basis $\{|b\rangle\}$ for B, we can write the action of Tr_B on any $\rho^{AB} \in \mathcal{D}(A \otimes B)$ as $\mathrm{Tr}_B(\rho^{AB}) = \sum_b \langle b | \rho^{AB} | b \rangle$. Note that the action of Tr_B is independent of the choice of basis chosen to represent it, so we unambiguously write $\rho^A = \mathrm{Tr}_B(\rho^{AB})$. We also use the notation $\mathrm{Tr}_{\neg A} = \mathrm{Tr}_B$ to express that we want to keep only the A register.

Fixing a basis also allows us to talk about classical states and joint states: $\rho \in \mathcal{D}(B)$ is classical (with respect to this basis) if it is diagonal in basis $\{|b\rangle\}$, i.e.

$$\rho = \sum_b p_B(b) \cdot |b\rangle\langle b|$$

for some probability distribution p_B. More generally, subsystem B of ρ^{AB} is said to be classical if we can write

$$\rho^{AB} = \sum_b p_B(b) \cdot |b\rangle\langle b|^B \otimes \rho_b^A$$

for some $\rho_b^A \in \mathcal{D}(A)$. An important example of a channel mapping a quantum system to a classical one is the measurement channel Δ_B, defined as

$$\Delta_B(\rho) = \sum_b \langle b \,|\, \rho \,|\, b\rangle \cdot |b\rangle\langle b|^B$$

for any $\rho \in \mathcal{D}(B)$. Note that for any state $\rho \in \mathcal{D}(B_1 \otimes B_2 \otimes C \otimes R)$ of the form

$$|\rho\rangle^{B_1 B_2 CR} = \sum_b \sqrt{p_B(b)} \cdot |b\rangle^{B_1} |b\rangle^{B_2} |\rho_b\rangle^{CR},$$

we have

$$\mathrm{Tr}_{B_2}(\rho^{B_1 B_2 CR}) = \sum_b p_B(b) \cdot |b\rangle\langle b|^{B_1} \otimes \rho_b^{CR}$$

and

$$\mathrm{Tr}_{B_2 R}(\rho^{B_1 B_2 CR}) = \sum_b p_B(b) \cdot |b\rangle\langle b|^{B_1} \otimes \rho_b^{C}$$

with the state on B_1 classical in both cases. Often, $A, B, C, \cdots$ will be used to discuss general systems, while $X, Y, Z, \cdots$ will be reserved for classical systems, or quantum systems like B_1 and B_2 above that are classical once one of them is traced out, and can be thought of as containing a quantum copy of the classical content of one another.

For a state $\rho^A \in \mathcal{D}(A)$, a purification is a pure state $\rho^{AR} \in \mathcal{D}(A \otimes R)$ satisfying $\mathrm{Tr}_R(\rho^{AR}) = \rho^A$. If R has dimension at least that of A, then such a purification always exists. For a given R, all purifications are equivalent up to a unitary on R, and more generally, if $\dim(R') \geq \dim(R)$ and $\rho_1^{AR}, \rho_2^{AR'}$ are two purifications of ρ^A, then there exists an isometry $V_\rho^{R \to R'}$ such that $\rho_2^{AR'} = V_\rho(\rho_1^{AR})$. For a channel $\mathbb{N} \in C(A, B)$, an isometric exten-

sion is a unitary $U_{\mathbb{N}} \in \mathcal{U}(A, A' \otimes B)$ with $\mathrm{Tr}_{A'}(U_{\mathbb{N}}(\rho^A)) = \mathbb{N}(\rho^A)$ for all ρ^A. Such an extension always exists provided A' is of dimension at least $\dim(A)^2$. For the measurement channel Δ_B, an isometric extension is given by $U_\Delta = \sum_b |b\rangle^{B'} |b\rangle^B \langle b|^B$.

The notion of distance we use is the trace distance, defined for two states $\rho_1, \rho_2 \in \mathcal{D}(A)$ as the sum of the absolute values of the eigenvalues of their difference:

$$\|\rho_1 - \rho_2\|_A = \mathrm{Tr}(|\rho_1 - \rho_2|).$$

It has an operational interpretation as four times the best bias possible in a state discrimination test between ρ_1 and ρ_2. The subscript tells on which subsystems the trace distance is evaluated, and remaining subsystems might need to be traced out. We use the following results about trace distance. For proofs of these and other standard results in quantum information theory that we use, see [Wil13]. The trace distance is monotone under noisy channels: for any $\rho_1, \rho_2 \in \mathcal{D}(A)$ and $\mathbb{N} \in \mathcal{C}(A, B)$,

$$\|\mathbb{N}(\rho_1) - \mathbb{N}(\rho_2)\|_B \leq \|\rho_1 - \rho_2\|_A. \tag{1.3}$$

For isometries, the inequality becomes an equality, a property called isometric invariance of the trace distance. Hence, for any $\rho_1, \rho_2 \in \mathcal{D}(A)$ and any $U \in \mathcal{U}(A, B)$, we have

$$\|U(\rho_1) - U(\rho_2)\|_B = \|\rho_1 - \rho_2\|_A. \tag{1.4}$$

Also, the trace distance cannot be increased by adjoining an uncorrelated system: for any $\rho_1, \rho_2 \in \mathcal{D}(A), \sigma \in \mathcal{D}(B)$

$$\|\rho_1 \otimes \sigma - \rho_2 \otimes \sigma\|_{AB} = \|\rho_1 - \rho_2\|_A. \tag{1.5}$$

The trace distance obeys a property that we call joint linearity: for a classical system X and two states $\rho_1^{XA} = p_X(x) \cdot |x\rangle\langle x|^X \otimes \rho_{1,x}^A$ and $\rho_2^{XA} = p_X(x) \cdot |x\rangle\langle x|^X \otimes \rho_{2,x}^A$,

$$\|\rho_1 - \rho_2\|_{XA} = \sum_x p_X(x) \|\rho_{1,x} - \rho_{2,x}\|_A. \tag{1.6}$$

The measure of information that we use is the von Neumann entropy, defined for any state $\rho \in \mathcal{D}(A)$ as

$$H(A)_\rho = -\operatorname{Tr}(\rho \log \rho),$$

in which we take the convention that $0 \log 0 = 0$, justified by a continuity argument. The logarithm log is taken in base 2. Note that H is invariant under isometries applied on ρ. If the state to be evaluated is clear from context, we might drop the subscript. von Neumann entropy of state ρ measures how much uncertainty there is in a system A whose state is ρ. Asymptotically, von Neumann entropy measures the amount of quantum communication needed to send the system A across a noiseless quantum channel. This was proven by Schumacher [Sch95].

Conditional entropy of state A conditioned on B for a state $\rho^{AB} \in \mathcal{D}(A \otimes B)$ is then defined as

$$H(A|B)_\rho = H(A, B)_\rho - H(B)_\rho.$$

While for classical random variables, conditional entropy is always non-negative, for quantum states, conditional entropy could be negative. This happens, for example, when ρ^{AB} is a pure entangled state. Then $H(A, B) = 0$ but $H(A) = H(B) > 0$. It thus appears that it would be hard to make sense of conditional entropy. However, an operational interpretation of conditional entropy has been given in [HOW07] in terms of a task called "quantum state

merging". Asymptotically, conditional entropy measures the entanglement cost of Alice transmitting the system A to Bob (Alice holds system A, Bob holds B and the joint state is ρ). If conditional entropy is negative, then that means this state merging can be achieved by classical communication from Alice to Bob alone and $-H(A|B)$ ebits are produced on the side (so entanglement is gained and not lost in the process).

Mutual information between systems A and B in a joint state $\rho^{AB} \in \mathcal{D}(A \otimes B)$ is defined as

$$I(A;B)_\rho = H(A)_\rho - H(A|B)_\rho.$$

Intuitively, mutual information measures how much correlation is there between systems A and B. Mutual information appears in the characterization of entanglement assisted classical capacity of a noisy quantum channel [BSST99]. Mutual information between systems A and B can also be interpreted as the amount of work required to destroy the correlations between A and B [GPW05]. Furthermore, quantum mutual information (between a purifying register R and A) also measures the amount of classical communication required in quantum state merging [HOW07].

Conditional mutual information between systems A and B from the point of a system C (when the joint state is $\rho^{ABC} \in \mathcal{D}(A \otimes B \otimes C)$) is defined as

$$I(A;B|C)_\rho = I(A;B,C)_\rho - I(A;C)_\rho$$

In the classical case, conditional mutual information can also be written as an expectation (over C) of mutual information terms. This does not exist in the quantum case. While conditional entropy could be negative, conditional mutual information is always non-negative. This was proven in [LR73] via a deep theorem in operator analysis, called Lieb's concavity theorem. Since then, several proofs (many operational) have been given for this fundamental

fact. For example, [HOW07] gives one such proof. The only operational interpretation for conditional mutual information that exists right now is via a task called "quantum state redistribution". In this there is a pure state $|\psi\rangle$ on four systems R, A, B, C. Systems A, C are with Alice, B is with Bob and R is a purifying system which no one has excess to. Alice wants to send system C to Bob while maintaining the correlations with R (and A and B as well). It was proven in [YD09] that the amount of entanglement-assisted classical communication required for this task is $I(R;C|B)_\psi$. Intuitively, the amount of correlations Bob has with R before is $I(R;B)$ and after is $I(R;C,B)$, so their difference $I(R;C|B)$ is the amount of communication that needs to be sent from Alice to Bob. The "quantum state redistribution" task played an important role in the definition of quantum information complexity, as we will see later.

Now we list some simple facts about quantum information theoretic quantities, which will be useful later. Note that mutual information and conditional mutual information are symmetric in interchange of A, B, and invariant under a local isometry applied to A, B or C. For any pure bipartite state $\rho^{AB} \in \mathcal{D}(A \otimes B)$, the entropy on each subsystem is the same:

$$H(A) = H(B). \tag{1.7}$$

Since all purifications are equivalent up to an isometry on the purification registers, we get that for any two pure states $|\phi\rangle^{ABCR'}$ and $|\psi\rangle^{ABCR}$ such that $\phi^{ABC} = \psi^{ABC}$,

$$I(C;R'|B)_\phi = I(C;R|B)_\psi. \tag{1.8}$$

For isomorphic A, A', a maximally entangled state $\psi \in \mathcal{D}(A \otimes A')$ is a pure state satisfying $H(A) = H(A') = \log \dim(A) = \log \dim(A')$. For a system A of dimension $\dim(A)$ and any

$\rho \in \mathcal{D}(A \otimes B \otimes C)$, we have the bounds

$$0 \leq H(A) \leq \log \dim(A), \tag{1.9}$$

$$-H(A) \leq H(A|B) \leq H(A), \tag{1.10}$$

$$0 \leq I(A;B) \leq 2H(A), \tag{1.11}$$

$$0 \leq I(A;B|C) \leq 2H(A). \tag{1.12}$$

If A or B is a classical system, we get the tighter bounds

$$0 \leq H(A|B), \tag{1.13}$$

$$I(A;B) \leq H(A), \tag{1.14}$$

$$I(A;B|C) \leq H(A). \tag{1.15}$$

The conditional mutual information satisfies a chain rule: for any $\rho \in \mathcal{D}(A \otimes B \otimes C \otimes D)$,

$$I(AB;C|D) = I(A;C|D) + I(B;C|AD). \tag{1.16}$$

For product states $\rho^{A_1B_1C_1A_2B_2C_2} = \rho_1^{A_1B_1C_1} \otimes \rho_2^{A_2B_2C_2}$, entropy is additive,

$$H(A_1A_2) = H(A_1) + H(A_2), \tag{1.17}$$

and so there is no conditional mutual information between product system,

$$I(A_1;A_2|B_1B_2) = 0, \tag{1.18}$$

and conditioning on a product system is useless,

$$I(A_1; B_1|C_1A_2) = I(A_1; B_1|C_1). \tag{1.19}$$

More generally,

$$I(A_1A_2; B_1B_2|C_1C_2) = I(A_1; B_1|C_1) + I(A_2; B_2|C_2). \tag{1.20}$$

Two important properties of the conditional mutual information are non-negativity, equivalent to strong subadditivity, and the data processing inequality. For any $\rho \in \mathcal{D}(A \otimes B \otimes C)$ and $\mathbb{N} \in \mathcal{C}(B, B')$, with $\sigma = \mathbb{N}(\rho)$,

$$I(A; B|C)_\rho \geq 0, \tag{1.21}$$

$$I(A; B|C)_\rho \geq I(A; B'|C)_\sigma. \tag{1.22}$$

For classical systems, conditioning is equivalent to taking an average: for any $\rho^{ABCX} = \sum_x p_X(x) \cdot |x\rangle\langle x|^X \otimes \rho_x^{ABC}$, for a classical system X and some appropriate $\rho_x \in \mathcal{D}(A \otimes B \otimes C)$,

$$H(A|BX)_\rho = \sum_x p_X(x) \cdot H(A|B)_{\rho_x}, \tag{1.23}$$

$$I(A; B|CX)_\rho = \sum_x p_X(x) \cdot I(A; B|C)_{\rho_x}. \tag{1.24}$$

1.5 Quantum Communication Complexity

Quantum communication complexity, introduced by Yao [Yao93], studies the amount of quantum communication that two parties, Alice and Bob, need to exchange in order to compute a function (usually boolean) of their private inputs. It is the natural quantum extension of classical communication complexity [Yao79]. While the inputs are classical and the end

result is classical, the players are allowed to use quantum resources while communicating. The model for quantum communication that we consider is the following. For a given bipartite relation $T \subset X \times Y \times Z_A \times Z_B$ and input distribution μ on $X \times Y$, Alice and Bob are given input registers A_{in}, B_{in} containing their classical input $x \in X, y \in Y$ at the outset of the protocol, respectively, and they output registers A_{out}, B_{out} containing their classical output $z_A \in Z_A, z_B \in Z_B$ at the end of the protocol, respectively, which should satisfy the relation T. We generally allow for some small error ϵ in the output, which will be formalized below. In this distributional communication complexity setting, the input is a classical state

$$\rho = \sum_{x \in X, y \in Y} \mu(x,y) \cdot |x\rangle\langle x|^{A_{in}} \otimes |y\rangle\langle y|^{B_{in}}$$

Similarly, the output

$$\Pi(\rho) = \sum_{z_A \in Z_A, z_B \in Z_B} p_{Z_A Z_B}(z_A, z_B) \cdot |z_A\rangle\langle z_A|^{A_{out}} \otimes |z_B\rangle\langle z_B|^{B_{out}}$$

of the protocol Π implementing the relation, and the error parameter corresponds to the average probability of failure $\sum_{x,y} \mu(x,y) \cdot [(x, y, \Pi(x,y)) \notin R] \leq \epsilon$.

A r-round protocol Π for implementing relation T on input $\rho^{A_{in}B_{in}}$ is defined by a sequence of isometries $U_1, \cdots, U_{r+1}$ along with a pure state $\psi \in \mathcal{D}(T_A \otimes T_B)$ shared between Alice and Bob, for arbitrary finite dimensional registers T_A, T_B. For appropriate finite dimensional memory registers $A_1, A_3, \cdots A_{r-1}, A'$ held by Alice, $B_2, B_4, \cdots B_{r-2}, B'$ held by Bob, and communication registers $C_1, C_2, C_3, \cdots C_r$ exchanged by Alice and Bob, we have $U_1 \in \mathcal{U}(A_{in} \otimes T_A, A_1 \otimes C_1), U_2 \in \mathcal{U}(B_{in} \otimes T_B \otimes C_1, B_2 \otimes C_2), U_3 \in \mathcal{U}(A_1 \otimes C_2, A_3 \otimes C_3), U_4 \in \mathcal{U}(B_2 \otimes C_3, B_4 \otimes C_4), \cdots, U_r \in \mathcal{U}(B_{r-2} \otimes C_{r-1}, B_{out} \otimes B' \otimes C_r), U_{r+1} \in \mathcal{U}(A_{r-1} \otimes C_r, A_{out} \otimes A')$. We adopt the convention that, in the first round, $B_1 = B_0 = B_{in} \otimes T_B$, in even rounds $B_i = B_{i-1}$, and in odd rounds $A_i = A_{i-1}$. In this way, in round i, after application of U_i, Alice holds register

A_i, Bob holds register B_i and the communication register is C_i. We slightly abuse notation and also write Π to denote the channel implemented by the protocol, i.e.

$$\Pi(\rho) = \mathrm{Tr}_{A'B'}(U_{r+1}U_r \cdots U_2 U_1(\rho \otimes \psi)). \tag{1.25}$$

To formally define the error, we introduce a purification register R. For a classical input

$$\rho^{A_{in}B_{in}} = \sum_{x \in X, y \in Y} \mu(x,y) \cdot |x\rangle\langle x|^{A_{in}} \otimes |y\rangle\langle y|^{B_{in}}$$

like we consider here, we can always take this purification to be of the form

$$|\rho\rangle^{A_{in}B_{in}R} = \sum_{x \in X, y \in Y} \sqrt{\mu(x,y)}\, |x\rangle^{A_{in}} |y\rangle^{B_{in}} |xy\rangle^{R_1} |xy\rangle^{R_2}$$

for an appropriately chosen partition of R into R_1, R_2. Note that if we trace out the the R_2 register, then we are left with a classical state such that R_1 contains a copy of the joint input. Then we say that a protocol Π for implementing relation T on input $\rho^{A_{in}B_{in}}$, with purification $\rho^{A_{in}B_{in}R}$, has average error $\epsilon \in [0,1]$ if $P_e^{\mu} = \Pr_{\mu,\Pi}[\Pi(\rho^{A_{in}B_{in}R_1}) \notin T] \leq \epsilon$. We denote the set of all such protocols as $\mathcal{T}(T,\mu,\epsilon)$. If we want to restrict this set to bounded round protocols with r rounds, we write $\mathcal{T}^r(T,\mu,\epsilon)$. The worst case error of a protocol is $P_e^w = \max_\mu P_e^\mu$, in which it is sufficient to optimize over all atomic distributions μ. We denote by $\mathcal{T}(T,\epsilon)$ the set of all protocols implementing relation T with worst case error at most ϵ, and by $\mathcal{T}^r(T,\epsilon)$ if we restrict this set to r-round protocols.

Let us formally define the different quantities that we work with.

Definition 1.5.1. For a protocol Π as defined above, we define the *quantum communication*

cost of Π as

$$QCC(\Pi) = \sum_i \log \dim(C_i).$$

Note that we do not require that $\dim(C_i) = 2^k$ for some $k \in \mathbb{N}$, as is usually done. This will not affect our definition on information cost and complexity, but might affect the quantum communication complexity by at most a factor of two, without affecting the round complexity. The corresponding notions of quantum communication complexity of a relation are:

Definition 1.5.2. For a relation $T \subset X \times Y \times Z_A \times Z_B$, an input distribution μ on $X \times Y$ and an error parameter $\epsilon \in [0, 1]$, we define the ϵ-error *quantum communication complexity* of T on input μ as

$$QCC(T, \mu, \epsilon) = \min_{\Pi \in \mathcal{T}(T,\mu,\epsilon)} QCC(\Pi),$$

and the worst-case ϵ-error *quantum communication complexity* of T as

$$QCC(T, \epsilon) = \min_{\Pi \in \mathcal{T}(T,\epsilon)} QCC(\Pi),$$

Remark 1.5.3. *For any $T, \mu, 0 \leq \epsilon_1 \leq \epsilon_2 \leq 1$, the following holds:*

$$QCC(T, \mu, \epsilon_2) \leq QCC(T, \mu, \epsilon_1),$$
$$QCC(T, \epsilon_2) \leq QCC(T, \epsilon_1).$$

We have the following definitions for bounded round quantum communication complexity, and a similar remark holds.

Definition 1.5.4. For a relation $T \subset X \times Y \times Z_A \times Z_B$, an input distribution μ on $X \times Y$,

an error parameter $\epsilon \in [0,1]$ and a bound $r \in \mathbb{N}$ on the number of rounds, we define the r-round, ϵ-error *quantum communication complexity* of T on input μ as

$$QCC^r(T,\mu,\epsilon) = \min_{\Pi \in \mathcal{T}^r(T,\mu,\epsilon)} QCC(\Pi),$$

and r-round, worst-case ϵ-error *quantum communication complexity* of T as

$$QCC^r(T,\epsilon) = \min_{\Pi \in \mathcal{T}^r(T,\epsilon)} QCC(\Pi),$$

1.5.1 Generalized Discrepancy Method

Generalized discrepancy method, also known as smooth discrepancy method, is one of the strongest methods for proving lower bounds for quantum communication.

Definition 1.5.5. Let $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ be a boolean function. The δ-generalized discrepancy bound of f, denoted by $GDM_\delta(f)$, is defined as:

$$GDM_\delta(f) = \max\{GDM_\delta^\mu(f):\ \mu \text{ a distribution over } \mathcal{X} \times \mathcal{Y}\}$$
$$GDM_\delta^\mu(f) = \max\{\log\left(\frac{1}{\mathrm{disc}^\mu(g)}\right),\ g : \mathcal{X} \times \mathcal{Y} \to \{0,1\},\ \Pr_{(x,y)\sim\mu}[f(x,y) \neq g(x,y)] \leq \delta\}$$
$$\mathrm{disc}^\mu(g) = \max\left\{ \left| \sum_{(x,y)\in R} (-1)^{g(x,y)} \cdot \mu(x,y) \right| :\ R \in \mathcal{R} \right\}$$

Here $\mathcal{R}$ is the set of combinatorial rectangles $\mathcal{A} \times \mathcal{B}$, $\mathcal{A} \subseteq \mathcal{X}, \mathcal{B} \subseteq \mathcal{Y}$. We state two results on the generalized discrepancy method, both due to Sherstov [She07, She12], which we will use to lower bound the quantum information complexity of disjointness. The first is a threshold direct product result that will be useful to prove that the generalized discrepancy method is a lower bound on the quantum information complexity of boolean functions, and the second is a lower bound on the generalized discrepancy for the disjointness function.

Theorem 1.5.6 ([She12]). *Let $\epsilon_{sh} > 0$ be a small enough absolute constant. Then for any boolean function f , the following communication problem requires $\Omega(nGDM_{1/5}(f))$ qubits of communication (with arbitrary entanglement): Solving with probability $2^{-\epsilon_{sh}n}$, at least $(1-\epsilon_{sh})n$ among n instances of f.*

The disjointness function is defined as follows: for $x, y \in \{0,1\}^n \times \{0,1\}^n$, $DISJ_n(x,y) = 1$ if for all $i \in [n]$, $x_i \wedge y_i = 0$, and 0 otherwise. We will need the following theorem.

Theorem 1.5.7 ([She07]). $GDM_{1/5}(DISJ_n) \geq \Omega(\sqrt{n})$

1.6 Quantum Information Complexity

We use the notion of quantum information complexity as defined in [Tou15]. The register R is the purification register, invariant throughout the protocol since we consider local isometric processing. Note that, as noted before when considering a R_1R_2 partition for R, for classical input distributions, the purification register can be thought of as containing a (quantum) copy of the classical input. The definition is however invariant under the choice of R and corresponding purification.

Definition 1.6.1. For a protocol Π and a state ρ with purification held in system R, we define the *quantum information cost* of Π on input ρ as

$$QIC(\Pi, \rho) = \sum_{i>0, odd} \frac{1}{2} I(C_i; R|B_i) + \sum_{i>0, even} \frac{1}{2} I(C_i; R|A_i).$$

Note that the above definition is obtained by summing the asymptotic quantum communication costs of the "quantum state redistribution" tasks in various steps of the protocol Π. The above definition has a very nice interpretation as sums of information learnt and forgotten in various rounds of the protocol [TL].

Proposition 1.6.2 ([TL])**.**

$$\begin{aligned} QIC(\Pi,\rho) = &\sum_{i>0,odd} \frac{1}{2} \cdot \textit{Information-forgotten-by-Alice in round } i \\ &+ \sum_{i>0,odd} \frac{1}{2} \cdot \textit{Information-learnt-by-Bob in round } i \\ &+ \sum_{i>0,even} \frac{1}{2} \cdot \textit{Information-forgotten-by-Bob in round } i \\ &+ \sum_{i>0,even} \frac{1}{2} \cdot \textit{Information-learnt-by-Alice in round } i \end{aligned}$$

Here

$$\textit{Information-forgotten-by-Alice in round } i = I(Y;C_i|X,A_i)_\rho$$

$$\textit{Information-learnt-by-Bob in round } i = I(X;C_i|Y,B_i)_\rho$$

$$\textit{Information-forgotten-by-Bob in round } i = I(X;C_i|Y,B_i)_\rho$$

$$\textit{Information-learnt-by-Alice in round } i = I(Y;C_i|X,A_i)_\rho$$

Definition 1.6.3. For a relation $T \subset X \times Y \times Z_A \times Z_B$, an input distribution μ on $X \times Y$, an error parameter $\epsilon \in [0,1]$ and a number of round r, we define the ϵ-error *quantum information complexity* of T on input μ as

$$QIC(T,\mu,\epsilon) = \inf_{\Pi \in \mathcal{T}(T,\mu,\epsilon)} QIC(\Pi,\mu),$$

and the r-round, ϵ-error *quantum information complexity* of T on input μ as

$$QIC^r(T,\mu,\epsilon) = \inf_{\Pi \in \mathcal{T}^r(T,\mu,\epsilon)} QIC(\Pi,\mu),$$

The following properties of quantum information cost and complexity were proved in

Ref. [Tou15].

Lemma 1.6.4. *For any protocol Π and input distribution μ, the following holds:*

$$0 \leq QIC(\Pi, \mu) \leq QCC(\Pi).$$

Lemma 1.6.5. *For a relation $T \subset X \times Y \times Z_A \times Z_B$, an input distribution μ on $X \times Y$, an error parameter $\epsilon \in [0, 1]$ and a number of round r, the following holds:*

$$0 \leq QIC(T, \mu, \epsilon) \leq QCC(T, \mu, \epsilon),$$
$$0 \leq QIC^r(T, \mu, \epsilon) \leq QCC^r(T, \mu, \epsilon).$$

Lemma 1.6.6. *For any two protocols Π^1 and Π^2 with r_1 and r_2 rounds, respectively, there exists a r-round protocol Π_2, satisfying $\Pi_2 = \Pi^1 \otimes \Pi^2, r = \max(r_1, r_2)$, such that the following holds for any corresponding input states ρ^1, ρ^2:*

$$QIC(\Pi_2, \rho^1 \otimes \rho^2) = QIC(\Pi^1, \rho^1) + QIC(\Pi^2, \rho^2).$$

Lemma 1.6.7. *For any r-round protocol Π_2 and any input states $\rho^1 \in \mathcal{D}(A^1_{in} \otimes B^1_{in}), \rho_2 \in \mathcal{D}(A^2_{in} \otimes B^2_{in})$, there exist r-round protocols Π^1, Π^2 satisfying $\Pi^1(\cdot) = \mathrm{Tr}_{A^2_{out}B^2_{out}} \circ \Pi_2(\cdot \otimes \rho^2)$, $\Pi^2(\cdot) = \mathrm{Tr}_{A^1_{out}B^1_{out}} \circ \Pi_2(\rho^1 \otimes \cdot)$, and the following holds:*

$$QIC(\Pi^1, \rho^1) + QIC(\Pi^2, \rho^2) = QIC(\Pi_2, \rho^1 \otimes \rho^2).$$

Lemma 1.6.8. *For any $p \in [0, 1]$, any two protocols Π^1, Π^2 with r_1, r_2 rounds, respectively, there exists a r-round protocol Π satisfying $\Pi = p\Pi^1 + (1-p)\Pi^2, r = \max(r_1, r_2)$, such that*

the following holds for any state ρ:

$$QIC(\Pi, \rho) = pQIC(\Pi^1, \rho) + (1-p)QIC(\Pi^2, \rho).$$

Corollary 1.6.9. *For any $p \in [0,1], T$ and $\epsilon, \epsilon_1, \epsilon_2 \in [0,1]$ satisfying $\epsilon = p\epsilon_1 + (1-p)\epsilon_2$, for any bound $r = \max(r_1, r_2), r_1, r_2 \in \mathbb{N}$ on the number of rounds and for any input distribution μ on $X \times Y$, the following holds:*

$$QIC(T, \mu, \epsilon) \leq pQIC(T, \mu, \epsilon_1) + (1-p)QIC(T, \mu, \epsilon_2),$$
$$QIC^r(T, \mu, \epsilon) \leq pQIC^{r_1}(T, \mu, \epsilon_1) + (1-p)QIC^{r_2}(T, \mu, \epsilon_2).$$

Lemma 1.6.10. *Let ν be a distribution over input states ρ and denote $\overline{\rho} := \mathbb{E}_{\rho\sim\nu}\rho$. Then for any protocol π,*

$$\mathbb{E}_{\rho\sim\nu}[QIC(\pi, \rho)] \leq QIC(\pi, \overline{\rho})$$

Lemma 1.6.11. *For any r-round protocol Π, any input distribution μ with copies of x, y in R_1, and any $\epsilon \in (0,2], \delta > 0$, there exists a large enough $n_0(\Pi, \rho, \epsilon, \delta)$ such that for any $n \geq n_0$, there exists a r-round protocol Π_n satisfying*

$$\|\Pi_n((\rho^{A_{in}B_{in}R_1})^{\otimes n}) - \Pi^{\otimes n}((\rho^{A_{in}B_{in}R_1})^{\otimes n})\|_{(A_{out}B_{out}R_1)^{\otimes n}} \leq \epsilon,$$
$$\frac{1}{n}QCC(\Pi_n) \leq QIC(\Pi, \rho) + \delta.$$

Chapter 2

Exact Communication Bounds for Disjointness

The results in this chapter are based on joint work with Mark Braverman, Denis Pankratov and Omri Weinstein [BGPW13a]. Only preliminary results will be presented in this chapter. All the results and full proofs can be found in the full version of the paper [BGPW13b].

2.1 Introduction

The set disjointness problem is one of the oldest and most studied problems in communication complexity [KN97]. In the two party setting, Alice and Bob are given subsets $X, Y \subset [n]$, respectively, and need to output 1 if $X \cap Y = \emptyset$, and 0 otherwise. Thus the disjointness function $Disj_n$ can be written as

$$Disj_n(X, Y) = \bigwedge_{i=1}^{n} (\neg X_i \vee \neg Y_i)$$

In the deterministic communication complexity model, it is easy to show that $Disj_n$ has communication complexity $n + 1$. In the randomized communication complexity model –

which is the focus of this paper – an $\Omega(n)$ lower bound was first proven by Kalyanasundaram and Schnitger [KS92]. The proof was combinatorial in nature. A much simpler combinatorial proof was given by Razborov a few years later [Raz92]. In terms of upper bounds on the communication complexity of disjointness, an $n+1$ bound is trivial. No better bound was known prior to this work, although by examining the problem, one can directly convince oneself that there is a randomized protocol for $Disj_n$, with tiny error, that uses only $(1-\varepsilon)n$ communication for some small $\varepsilon > 0$ – so that the deterministic algorithm is suboptimal. Another set of techniques which were successfully applied to versions of disjointness, especially in the quantum and multiparty settings [Raz02, CA08, She14] are analytic techniques. Analytic techniques such as the pattern matrix method [She07], allow one to further extend the reach of combinatorial techniques.

The first information-theoretic proof of the lower bound for disjointness was given by Bar-Yossef et al. [BYJKS04]. The information-theoretic approach was extended to the multi-party number-in-hand setting [CKS03, Gro09, Jay09a] with applications to tight lower bounds on streaming algorithms. At the core of the proof is a direct-sum reduction of proving an $\Omega(n)$ bound on $Disj_n$ to proving an $\Omega(1)$ bound on the information complexity of AND. The direct sum in this and other proofs follows from an application of the chain rule for mutual information – one of the primary information-theoretic tools. More recently, an information complexity view of disjointness lead to tight bounds on the ability of extended formulations by linear programs to approximate the CLIQUE problem [BM13]. This suggests that information complexity and a better understanding of the disjointness problem may have other interesting implications within computational complexity.

A problem related to disjointness is *Set Intersection* Int_n: now Alice and Bob do not want to just determine whether X and Y intersect, but both want to learn the intersection set $X \cap Y$. For this problem, even in the randomized setting, a lower bound of n bits on the communication is trivial: by fixing $X = [n]$ we see that in this special case the problem

will amount to Bob sending his input to Alice – which clearly requires $\geq n$ bits. Thus the randomized communication complexity of this problems lies somewhere between n and $2n$ – the trivial upper and lower bounds. Note that the intersection problem is nothing but n copies of the two-bit AND function. Therefore, determining the communication complexity of Int_n is equivalent to determining the information complexity of the two-bit AND function by the "information equals amortized communication" connection.

Essentially independently of the communication complexity line of work described above, a study of the AND/intersection problem has recently originated in the information theory community. A series of papers by Ma and Ishwar [MI08, MI11] develops techniques which allow one to calculate tight bounds on the communication complexity of Int_n and other amortized functions on the condition that one only considers protocols restricted to r rounds of communication. These techniques allow one to numerically (and sometimes analytically) compute the information complexity of the two-bit AND function – although the numerical computation is not provably correct for the most general unbounded round case since the rate of convergence of r-round information complexity down to the true information complexity is unknown. Furthermore, their results are non-constructive in the sense that they do not exhibit a protocol achieving their bounds. Nonetheless, numerical calculations produced by Ma and Ishwar do point at convergence to 1.4923 bits for the AND function [MI]. As discussed below, our tight upper and lower bounds are consistent with this evidence.

The main result of this chapter is giving tight bounds on the information and communication complexity of the AND, Int_n, and $Disj_n$ functions. We give a (provably) information-theoretic optimal protocol for the two-bit AND function. Combined with prior results – and new additional technical work – this optimality immediately gives a tight optimal randomized protocol for Int_n that uses $C_\wedge \cdot n \pm o(n)$ bits of communication and fails with a vanishing probability. Here $C_\wedge \approx 1.4923$ is an explicit constant given as a maximum of a concave analytic function. We then apply the same optimal result to obtain the optimal

protocol for set disjointness, showing that the best vanishing error randomized protocol for $Disj_n$ will take $C_{DISJ} \cdot n \pm o(n)$ bits of communication, where $C_{DISJ} \approx 0.4827$ is another explicit constant (which we found to be surprisingly low). The fact that we need the bounds to be exact throughout requires us to develop some new technical tools for dealing with information complexity in this context. For example, we show that the randomized ε-error information complexity converges to the 0-error information complexity as $\varepsilon \to 0$.

Applying what we've learned about the AND function to the *sparse sets* regime, we are able to determine the precise communication complexity of disjointness $Disj_n^k$ where the sets are restricted to be of size at most k. Håstad and Wigderson [HW07] showed that the randomized communication complexity of this problem is $\Theta(k)$. We sharpen this result by showing that for vanishing error the communication complexity of $Disj_n^k$ is $\frac{2}{\ln 2}k \pm o(k) \approx 2.885k \pm o(k)$.

Interestingly the optimal protocol we obtain for AND is not an actual protocol in the strict sense of the definition of a communication protocol. One way to visualize it is as a game show where Alice and Bob both have access to a "buzzer" and the game stops when one of them "buzzes in". The exact time of the "buzz in" matters. If we wanted to simulate this process with a conventional protocol, we'd need the time to be infinitely quantized, with Alice and Bob exchanging messages of the form "no buzz in yet", until the buzz in finally happens. Thus the optimal information complexity of AND is obtained by an infimum of a sequence of conventional protocols rather than by a single protocol.

It turns out that the unlimited number of rounds is necessary, both for the AND function and for $DISJ_n$. Our understanding of information complexity in the context of the AND function allows us to calculate the asymptotics of the amount of communication needed if we restrict the number of rounds of interaction between players to r. $R(Disj_n, 0^+, r) = (C_{DISJ} + \Theta(1/r^2)) \cdot n$. In particular, any constant bound on the number of rounds means a linear loss in communication complexity. There are well-known examples in communication

complexity where adding even a single round causes an exponential reduction in the amount of communication needed [NW93a]. There are also examples of very simple transmission problems where it can be shown that two rounds are much better than one, and more than two are better yet [Orl90, Orl91]. However, to our knowledge, this is the first example of a "natural" function where an arbitrary number of additional rounds is provably helpful.

2.2 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 – characterization was given in the information theory literature by Ma and Ishwar [MI08]. Let $\Delta(\mathcal{X} \times \mathcal{Y})$ denote the set of distributions over $\mathcal{X} \times \mathcal{Y}$.

Lemma 2.2.1. *For any function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ there exists a family $\mathfrak{C}(f)$ of functions $C : \Delta(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}_{\geq 0}$ 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*

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

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 2.2.2.

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

Theorem 2.2.3.

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

We also analyze the rate of convergence to the optimal information cost, as the number r of permitted rounds increases. Recently, the authors in [BS16] obtained a rate of convergence analysis (not tight) for arbitrary functions, thus proving that information complexity is computable.

Theorem 2.2.4. *For all $\mu \in \Delta(\{0,1\} \times \{0,1\})$ with full support we have*

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

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

$$(C_{DISJ} - \delta) \cdot n \leq \mathcal{R}_\epsilon(DISJ_n) \leq C_{DISJ} \cdot n + o(n).$$

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

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

Theorem 2.2.6. *Let* n, k *be such that* $k = \omega(1)$ *and* $n/k = \omega(1)$. *Then for all constant* $\epsilon > 0$,

$$\left(\frac{2}{\ln 2} - O(\sqrt{\epsilon})\right) \cdot k - o(k) \leq \mathcal{R}_\epsilon(DISJ_n^k) \leq \frac{2}{\ln 2} \cdot k + o(k).$$

We also observe that Theorem 2.2.2 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 $\epsilon = 0$:

Theorem 2.2.7. *For all* $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ *and* $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$ *we have*

$$\lim_{\epsilon \to 0} \mathrm{IC}_\mu(f, \epsilon) = \mathrm{IC}_\mu(f, 0), \tag{2.1}$$

$$\lim_{\epsilon \to 0} \mathrm{IC}_\mu^{ext}(f, \epsilon) = \mathrm{IC}_\mu^{ext}(f, 0). \tag{2.2}$$

2.3 Preliminaries

2.3.1 Notation

Capital letters are reserved for random variables (e. g., A, B, C), calligraphic letters for sets (e. g., $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \ldots$), and small letters for elements of sets (e. g., $a, b, c, \ldots$). For typographical purposes we shall write $A_1A_2 \cdots A_n$ to denote the random variable $(A_1, A_2, \ldots, A_n)$ and *not* the random variable that is the product of the A_i, unless otherwise specified. We use $[n]$ to denote the set $\{1, \ldots, n\}$.

For random variables A and B_i ($i \in [n]$) and elements $b_i \in \text{range}\, B_i$ ($i \in [n]$) we write $A_{b_1 b_2 \cdots b_n}$ to denote the random variable A conditioned on the event "$B_1 = b_1, B_2 = b_2, \ldots, B_n = b_n$". Whenever convenient we shall 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 $\mu(1,1) = \delta$. We use $\Delta(\mathcal{X})$ to denote *the family of all probability distributions* on $\mathcal{X}$. 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.

2.3.2 Information Complexity

We refer the reader to Section 1.3 for definitions of $IC_\mu(\pi)$: information cost of a protocol π w.r.t. the distribution μ and $\text{IC}^{\text{ext}}_\mu(\pi)$: external information cost of a protocol π w.r.t. the distribution μ. We repeat here some definitions of the information complexity of a function f.

The *information complexity* of f with respect to μ is

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

where the infimum ranges over all (randomized) protocols π solving f with error at most ϵ when inputs are sampled according to μ. Note that we cannot replace the above quantifier with a min, since the information cost of a function may not be achievable by any fixed (finite-round) protocol[1].

[1] In fact, we shall see that this is the case for the AND function whose information complexity is analyzed in this chapter.

We also define an absolute 0-error information complexity of f w.r.t μ

$$\overline{\mathrm{IC}}_\mu(f,0) := \inf_\pi \mathrm{IC}_\mu(\pi),$$

where the infimum ranges over all (randomized) protocols π solving f correctly on *all* inputs.

Similarly, the *external information complexity* of f with respect to μ is defined as

$$\mathrm{IC}^{\mathrm{ext}}_\mu(f,\epsilon) := \inf_\pi \mathrm{IC}^{\mathrm{ext}}_\mu(\pi).$$

The *prior-free* information complexity of a function f (or simply, the *information cost* of f) with error ϵ is defined as

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

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 $\mathrm{IC}(f,0)$ 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 2.3.1 ([Bra12])**.**

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

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

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

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.

2.4 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 2.7. 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 complexity 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 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

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

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 $I(\mu) := \mathrm{IC}_\mu(\pi)$ in terms of $\alpha, \beta, \gamma, \delta$. In addition, we conclude that $I(\mu) \geq \mathrm{IC}_\mu(\mathrm{AND}, 0)$. 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 2.5. 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 $I(\mu) = \mathrm{IC}_\mu(\mathrm{AND}, 0)$ 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 can be found in the full version of the 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 2.4.1. *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 $\mathrm{IC}_{\mu}(\mathrm{AND}, 0) = \mathrm{IC}_{\mu^T}(\mathrm{AND}, 0)$, therefore it suffices to compute the information cost for the AND function only for distributions with $\beta \leq \gamma$.

Theorem 2.4.2. *For a symmetric distribution* $\nu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \beta & \delta \\ \hline \end{array}$ *we have*

$$\mathrm{IC}_{\nu}(AND, 0) = \frac{\beta}{\ln 2} + 2\delta \log \frac{\beta+\delta}{\delta} + 2\beta \log \frac{\beta+\delta}{\beta} + \frac{\beta^2}{\alpha} \log \frac{\beta}{\beta+\alpha} + \alpha \log \frac{\alpha+\beta}{\alpha} \qquad (2.3)$$

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

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

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*

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

Theorem 2.4.3. *(Theorem 2.2.2 restated)*

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

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 2.4.4. *Observe that the maximum of* $\mathrm{IC}(AND, 0)$ *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* $\mathrm{IC}_{\mu}(f, 0) = \mathrm{IC}_{\mu^T}(f, 0)$ *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* $\mathrm{IC}_{\mu'}(f, 0) \geq \mathrm{IC}_{\mu}(f, 0)/2 + \mathrm{IC}_{\mu^T}(f, 0)/2 = \mathrm{IC}_{\mu}(f, 0)$.

Theorem 2.4.5. *(Theorem 2.2.3 restated)*

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

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 2.5 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 2.4.6. *For symmetric distributions* $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \beta & 0 \\ \hline \end{array}$ *we have*

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

Theorem 2.4.7. *For distributions* $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & 0 \\ \hline \end{array}$ *we have*

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

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 2.4.8.

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

The distribution that achieves this maximum is

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

2.5 Characterization of Information Cost

In this section we prove Lemma 2.2.1, a local characterization of the zero-error information complexity 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 $I(\mu) := \mathrm{IC}_\mu(f, 0)$ 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* $\mathrm{IC}_\mu(\mathrm{AND}, 0)$.

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 observer 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 can be found in the full version of the paper. Now we proceed to define the family $\mathfrak{C}(f)$.

Definition 2.5.1. Let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a given function. Define a family $\mathfrak{C}(f)$ of all functions $C : \Delta(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}_{\geq 0}$ satisfying the following constraints:

- $(\forall \mu \in \Delta(\mathcal{X} \times \mathcal{Y}))(f|_{\mathrm{supp}(\mu)} \text{ is constant} \Rightarrow C(\mu) = 0)$,
- $\forall \mu, \mu_0^A, \mu_1^A \in \Delta(\mathcal{X} \times \mathcal{Y})$ 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 2.5.2. *The notation $f|_{\mathrm{supp}(\mu)} \equiv$ 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 2.5.3. *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 $C(\mu) \leq \mathrm{IC}_\mu(\pi)$.*

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

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

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

□

Corollary 2.5.4. *For all $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ we have*

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

2.6 Applications: Exact Communication Bounds

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" ([BR11, Bra12]), Thus Theorem 2.2.2 (along with continuity of information cost at error $= 0$) implies that

Corollary 2.6.1. *For all $\epsilon > 0$, there exists $\delta = \delta(\epsilon) > 0$ such that $\delta \to 0$ as $\epsilon \to 0$ and*

$$(C_\wedge - \delta) \cdot n \leq \mathcal{R}_\epsilon(Int_n) \leq C_\wedge \cdot n + o(n),$$

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 $f^{-1}(0)$[2]:

Theorem 2.6.2. *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 $\epsilon > 0$, there exists $\delta = \delta(f, \epsilon) > 0$ such that $\delta \to 0$ as $\epsilon \to 0$ and*

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

where $\mathsf{IC}^0(f, 0) := \max_{\mu:\mu(1,1)=0} \overline{\mathrm{IC}}_\mu(f, 0)$.[3]

The formal proof can be found in the full version of the 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 equals 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 2.2.7), we get a contradiction.

Theorem 2.2.5 now follows from Theorem 2.6.2. For convenience, we restate it below

Corollary 2.6.3 (Theorem 2.2.5 restated)**.** *For all $\epsilon > 0$, there exists $\delta = \delta(\epsilon) > 0$ such that $\delta \to 0$ as $\epsilon \to 0$ and*

$$(C_{DISJ} - \delta) \cdot n \leq \mathcal{R}_\epsilon(Disj_n) \leq C_{DISJ} \cdot n + o(n).$$

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

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

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

Proof. Since randomized communication complexity is closed under complementation, $\mathcal{R}_\epsilon(Disj_n) = \mathcal{R}_\epsilon(\bigvee_{i=1}^n (X_i \wedge Y_i))$, and thus Theorem 2.6.2 (with $f = AND$ and $k = 1$) implies that

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

But Theorem 2.4.8 asserts that $\max_{\mu:\mu(1,1)=0} \overline{\mathrm{IC}}_\mu(\mathrm{AND}, 0) = 0.4827\ldots$, which completes the proof. □

The communication complexity of the k-Disjointness problem is known to be $\Theta(k)$ [HW07]. We are able to determine the exact constant in this regime as well.

Theorem 2.6.4. *(Theorem 2.2.6 restated) Let n, k be such that $k = \omega(1)$ and $n/k = \omega(1)$. Then for all constant $\epsilon > 0$,*

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

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 2.4.2 implies that $IC_{\mu_k}(AND, 0) = \frac{2}{\ln 2}\frac{k}{n} \pm o(\frac{k}{n})$. The proof of Theorem 2.6.4 follows the ideas of Theorem 2.6.2, 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 $\epsilon = 0$, using a recursive application of our optimal protocol from section 2.4. For a formal proof see the full version of the paper.

2.7 Rate of Convergence

In this section we consider the rate at which $\mathrm{IC}^r_\mu(\mathrm{AND}, 0)$ converges to $\mathrm{IC}_\mu(\mathrm{AND}, 0)$ 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 [MI09], 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 $\mathrm{IC}_\mu(f)$ for product distributions μ and design an algorithm that computes $\mathrm{IC}^r_\mu(f)$ to within a desired accuracy.

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 2.7.1. *For all* $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & \delta \\ \hline \end{array}$ *with* $\alpha, \beta, \gamma > 0$ *we have*

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

We present the high-level idea of the proof of Theorem 2.7.1. 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 2.4 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 2.7.2.

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

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 2.7.2 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 can be found in the full version of the paper.

From the $\Omega(1/r^2)$-bound on the rate of convergence of r-round information cost of AND function together with results from Section 2.5 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 atleast $\Omega(n/r^2)$ bits. Thus for both problems a protocol that is optimal up to lower-order terms will need to use $\omega(1)$ rounds of communication.

Chapter 3

Information Lower Bounds via Self-reducibility

The results in this chapter are based on joint work with Mark Braverman, Denis Pankratov and Omri Weinstein [BGPW13c].

3.1 Introduction

In this chapter we develop a new self-reducibility technique for deriving information complexity lower bounds from communication complexity lower bounds. The technique works for functions that have 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 starting point is a communication complexity lower bound for f_{nk} (that may be obtained by any means). Assuming self-reducibility, the same bound applies to f_n^k, which through the connection between information complexity and amortized communication complexity [BR11], implies a lower bound on the information complexity of

f_n. In this chapter we develop tools to make this reasoning go through.

Ideas of self-reducibility are central in applications of information complexity to communication complexity lower bounds, starting with the work of Bar-Yossef et al. [BYJKS04]. These arguments start with an information complexity lower bound for a (usually very simple) problem, and derive a communication complexity bound on many copies of the problem. The logic in this chapter is reversed: we start with a communication complexity lower bound, which we use as a black-box, and use self-reducibility to derive an amortized communication complexity bound, which translates into an information complexity lower bound.

3.1.1 Results

We use the self-reducibility technique to prove results about the information complexity of the Gap Hamming Distance and Inner Product problems. 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. [CKW12]. 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 probability of error at most ε under μ. We prove

Theorem 3.1.1. *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, we prove a stronger bound on its information complexity. Formally

Theorem 3.1.2. *For every constant $\delta > 0$, there exists a constant $\epsilon > 0$, and n_0 such that $\forall\ n \geq n_0$, $IC_{\mathcal{U}}(IP_n, \epsilon) \geq (1-\delta)n$. Here $\mathcal{U}$ is the uniform distribution over $\{0,1\}^n \times \{0,1\}^n$.*

Note that $IC_{\mathcal{U}}(IP_n, \epsilon) \leq (1-2\epsilon)(n+1)$, since the parties can always give a random output with probability 2ϵ, and with probability $1-2\epsilon$, have one of the parties send its entire input and the other party send back the output. Also it is known that $IC_{\mathcal{U}}(IP_n, \epsilon) \geq \Omega(n)$, for all $\epsilon \in [0, 1/2)$ [BW12]. 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).

3.1.2 Discussion and open problems

Although in complexity theory we often don't care about the constants (and often it is not necessary), proving theorems with the right constants can often lead to deeper insights into the mathematical structure of the problem [BGPW13a, BM13]. There are few techniques that allow us to find the right constants and there are fewer problems for which we can. We believe that answering the following problem will lead to development of new techniques and also reveal interesting insights into the problem of computing the XOR of n copies of a function.

Open Problem 3.1.3. *Is it true that for small constants ϵ and sufficiently large n,*

$$IC_{\mathcal{U}}(IP_n, \epsilon) \geq (1 - 2\epsilon - o(\epsilon))n$$

As before $\mathcal{U}$ is the uniform distribution. If this is false, is there a different constant $\alpha > 2$ such that as $\epsilon \to 0$ we get $IC_{\mathcal{U}_n}(IP_n, \epsilon) \geq (1 - \alpha \cdot \epsilon)n$?

Solving this problem may require shedding new light on the rate of convergence of the $IC_\mu(\bullet, \epsilon)$ to $IC_\mu(\bullet, 0)$ as $\epsilon \to 0$, and better understanding the role error plays in information

complexity.

It is somewhat difficult to define the exact meaning of the "right" constant for the Gap Hamming Distance problem, since it is a promise problem defined by two parameters (gap and error). Nonetheless, there is a very natural regime in which understanding the exact information complexity of GHD_n is a natural and interesting problem. Namely:

Open Problem 3.1.4. *Is it true that for all $\varepsilon > 0$, there is a $\epsilon > 0$ and a distribution μ such that $\mathsf{IC}_\mu(GHD_{n,n/2,\epsilon\sqrt{n}}, \epsilon) > (1-\varepsilon)n$?*

In other words, does the information complexity of GHD_n tend to the trivial upper bound as we tighten the gap and error parameters? This is related to the same (but weaker) question one can ask about the communication complexity of GHD_n in this regime.

3.2 Information complexity of Gap Hamming Distance

Given two strings $x, y \in \{0,1\}^n$, the hamming distance x and y is defined to be $HAM(x,y) = |\{i \,|\, x_i \neq y_i\}|$. In the Gap Hamming Distance (GHD) problem, Alice gets a string $x \in \{0,1\}^n$ and Bob gets a string $y \in \{0,1\}^n$. They are promised that either $HAM(x,y) \geq n/2 + \sqrt{n}$ or $HAM(x,y) \leq n/2 - \sqrt{n}$, and they have to find which is the case. We can define a general version $GHD_{n,t,g}$, where Alice and Bob have to determine if $HAM(x,y) \geq t+g$ or $HAM(x,y) \leq t-g$, but the parameters $t = n/2$ and $g = \sqrt{n}$ are the most natural as discussed in [CR11]. In a technical tour-de-force, it was proven in [CR11] that the randomized communication complexity of the Gap Hamming Distance problem is linear. Formally,

Theorem 3.2.1. *For all constants $\gamma > 0$, and $\epsilon \in [0, 1/2)$, $R_\epsilon(GHD_{n,n/2,\gamma\sqrt{n}}) \geq \Omega(n)$.*

One can extend the formulation of GHD beyond the promise-problem setting. This particularly makes sense in a distributional-complexity setting. In this setting, we allow f to take the value $\star$, which means that we don't care about the output. The error in this

model is aggregated only over points on which the value of f is not $\star$. Chakrabarti and Regev [CR11] also prove a distributional version of the linear lower bound over the *uniform* distribution $\mathcal{U}$. Specifically, they prove

Theorem 3.2.2 ([CR11])**.** *There exists an absolute constant $\varepsilon > 0$ for which*

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

Kerenidis et al. [KLL$^+$12a] proved that the information complexity of Gap Hamming Distance is also linear, at least with respect to some distribution. The proof of Kerenidis et al. relies on a reduction that shows that a large class of communication complexity lower bound techniques also translate into information complexity lower bounds – including the lower bound for GHD:

Theorem 3.2.3 ([KLL$^+$12a])**.** *There exists a distribution μ on $\{0,1\}^n \times \{0,1\}^n$ and an absolute constant $\varepsilon > 0$ such that*

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

Interestingly, while this approach yields an analogue of Theorem 3.2.1 for information complexity, it does not seem to yield an analogue of the stronger Theorem 3.2.2, i.e. a lower bound on information complexity under the uniform distribution.

We give an alternate proof of the linear information complexity lower bound for GHD using the self-reducibility technique. Unlike the proof in [KLL$^+$12a] we do not need to dive into the details of the proof of the communication complexity lower bound for GHD. Rather, our starting point is Theorem 3.2.2, which we use as a black-box.

In fact, we will prove a slightly weaker lemma, with Theorem 3.1.1 following by a reduction. The reduction is conceptually very simple, but the details are somewhat tedious.

Lemma 3.2.4. *There exists absolute constants $\varepsilon > 0$ and $\gamma > 0$ for which*

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

3.3 Proof of Theorem 3.1.1

3.3.1 Proof Idea

We use the self-reducibility argument. Assume that for some $\epsilon > 0$, $IC_{\mathcal{U}}(GHD_n, \epsilon) = o(n)$. Then using "information equals amortized communication", we can get a protocol τ that solves N copies of GHD_n with $o(nN)$ communication. The heart of the argument is to use this to solve GHD_{nN} with $o(nN)$ *communication*, which is a contradiction. Say that Alice and Bob are given $x, y \in \{0,1\}^{nN}$ respectively. They sample $c \cdot nN$ random coordinates (for some constant c) and then divide these into cN blocks and run GHD_n on them all in parallel using $o(nN)$ communication. If $HAM(x,y) = nN/2 + \sqrt{nN}$, then the expected hamming distance of each block is $n/2 + \sqrt{n/N}$. Although the gain over $n/2$ is small, the hamming distance is still biased towards being $> n/2$. We will see that on each instance the protocol for GHD_n must gain an advantage of $\Omega(1/\sqrt{N})$ over random guessing. This in turn implies that cN copies suffice to get the correct answer with high probability.

3.3.2 Formal Proof of Lemma 3.2.4

Assume that for some ρ sufficiently small (to be specified later), $\mathsf{IC}_{\mathcal{U}}(GHD_{n,n/2,\sqrt{n}}, \rho) = o(n)$. Thus $\forall\ \alpha > 0$, and for sufficiently large n, $\mathsf{IC}_{\mathcal{U}}(GHD_{n,n/2,\sqrt{n}}, \rho) \leq \alpha n$. We will need the following theorem from [Bra12, BR11]:

Theorem 3.3.1 ([Bra12, BR11])**.** *Let $f : X \times Y \to \{0,1\}$ be a (possibly partial) function, let μ be any distribution on $X \times Y$, and let $I = \mathsf{IC}_{\mu}(f, \rho)$, then for each $\delta_1, \delta_2 > 0$, there*

is an $N = N(f, \rho, \mu, \delta_1, \delta_2)$ such that for each $n \geq N$, there is a protocol π_n for computing n instances of f with the following properties: let μ_n be any distribution over $X^n \times Y^n$ s.t. the marginal on each coordinate is μ. The protocol π_n has expected communication cost $< n(1+\delta_1)I$ w.r.t. μ_n. Moreover, if we let π be any protocol for computing f with information cost $\leq (1+\delta_1/3)I$ w.r.t. μ, then we can design π_n so that for each set of inputs, the statistical distance between the output of π_n and π^n is $< \delta_2$, where π^n denotes n independent executions of π.

In other words, Theorem 3.3.1 allows us to take a low-information protocol for f and turn it into a low-communication protocol for (sufficiently) many copies of f.

Step 1: From GHD to a tiny advantage.

In the first step we show that a protocol for GHD over the uniform distribution has a small but detectable advantage in distinguishing inputs from two distributions that are very close to each other. Denote by μ_η the distribution where $X \in \{0,1\}^n$ is chosen uniformly, and Y is chosen so that $X_i \oplus Y_i \sim B_{1/2+\eta}$ are i.i.d. Bernoulli random variables with bias η. Note that in this language the GHD problem is essentially about distinguishing $\mu_{-1/\sqrt{n}}$ from $\mu_{1/\sqrt{n}}$.

Lemma 3.3.2. *There exists absolute constants $\tau > 0$, $\gamma > 0$ and $\rho > 0$ with the following property. Suppose that for all n large enough there is a protocol π_n that solves $GHD_{n,n/2,\gamma\sqrt{n}}$ with error ρ w.r.t the uniform distribution. Then for all n large enough for all $\varepsilon < 1/n^2$ we have*

$$Pr_{(x,y)\sim\mu_\varepsilon}[\pi_n(x,y)=1] - Pr_{(x,y)\sim\mu_0}[\pi_n(x,y)=1] > \tau \cdot \varepsilon \cdot \sqrt{n}, \tag{3.1}$$

and

$$Pr_{(x,y)\sim\mu_{-\varepsilon}}[\pi_n(x,y)=0] - Pr_{(x,y)\sim\mu_0}[\pi_n(x,y)=0] > \tau \cdot \varepsilon \cdot \sqrt{n}. \tag{3.2}$$

Proof. Note that we can assume that the protocol π_n is symmetric w.r.t the hamming distance, i.e. its behavior depends just on the hamming distance between x and y. This is

because Alice and Bob can start with applying a random permutation and a random XOR on their inputs i.e. they sample (using public randomness) a permutation $\pi \in S_n$ and $r \in \{0,1\}^n$ and change their inputs to $\pi(x \oplus r)$ and $\pi(y \oplus r)$. Note that the information cost of the protocol remains the same.

We will establish (3.1), with (3.2) established identically. We first focus on the region where $HAM(x,y) \geq n/2$ and show that its contribution to (3.1) is at least $\Omega(\varepsilon\sqrt{n})$. We break the region into two further regions: (I) (x,y) with $n/2 < H(x,y) < n/2 + \gamma\sqrt{n}$; (II) (x,y) with $n/2 + \gamma\sqrt{n} \leq H(x,y)$ for appropriately chosen γ. We show that the contribution of region (II) is $\Omega(\varepsilon\sqrt{n})$, while the fact that the contribution of region (I) is positive is easy to see.

Denote by p_i the probability that π_n returns 1 on an input of hamming distance $n/2 + i$. The contribution of the region where $H(x,y) = n/2 + i$ is equal to

$$p_i \cdot (Pr_{\mu_\varepsilon}[H(x,y) = n/2 + i] - Pr_{\mu_0}[H(x,y) = n/2 + i]) =$$
$$p_i \cdot Pr_{\mu_0}[H(x,y) = n/2 + i] \cdot \left((1 - 4\varepsilon^2)^{n/2-i}(1 + 2\varepsilon)^{2i} - 1\right)$$

Now $(1 - 4\varepsilon^2)^{n/2-i} \geq 1 - 2\varepsilon/n$ and $(1 + 2\varepsilon)^{2i} \leq e^2$ (since $\varepsilon < 1/n^2$). Thus

$$\sum_{i=0}^{n/2} p_i \cdot Pr_{\mu_0}[H(x,y) = n/2 + i] \cdot (2\varepsilon/n) \cdot (1 + 2\varepsilon)^{2i} = O(\varepsilon/n)$$

and therefore

$$\sum_{i=0}^{n/2} p_i \cdot Pr_{\mu_0}[H(x,y) = n/2 + i] \cdot \left((1-4\varepsilon^2)^{n/2-i}(1+2\varepsilon)^{2i} - 1\right) \geq$$
$$\sum_{i=0}^{n/2} p_i \cdot Pr_{\mu_0}[H(x,y) = n/2 + i] \cdot \left((1-4\varepsilon^2)^{n/2-i}(1+2\varepsilon)^{2i} - 1\right)$$
$$- \sum_{i=0}^{n/2} p_i \cdot Pr_{\mu_0}[H(x,y) = n/2 + i] \cdot \left((1+2\varepsilon)^{2i} - 1\right) \geq -O(\varepsilon/n)$$

Thus the contribution from region (I) is $\geq -O(\varepsilon/n)$.

This leaves us with region (II), where we need to show that we actually get a non-negligible advantage. Let T be an appropriately chosen constant, so that $Pr_{\mu_0}[\gamma\sqrt{n} \leq H(x,y) - n/2 \leq T\sqrt{n}] = \Omega(1)$. The advantage

$$\sum_{i=\gamma\sqrt{n}}^{n/2} p_i \cdot Pr_{\mu_0}[H(x,y) = n/2 + i] \cdot \left((1+2\varepsilon)^{2i} - 1\right)$$
$$\geq \sum_{i=\gamma\sqrt{n}}^{T\sqrt{n}} p_i \cdot Pr_{\mu_0}[H(x,y) = n/2 + i] \cdot 4i\varepsilon$$
$$= \sum_{i=\gamma\sqrt{n}}^{T\sqrt{n}} Pr_{\mu_0}[H(x,y) = n/2 + i] \cdot 4i\varepsilon - \sum_{i=\gamma\sqrt{n}}^{T\sqrt{n}} (1-p_i) \cdot Pr_{\mu_0}[H(x,y) = n/2 + i] \cdot 4i\varepsilon$$
$$\geq \Theta(\varepsilon\sqrt{n}) - \rho \times 4T\varepsilon\sqrt{n}$$

since $(1-p_i)$ is the probability that the protocol errs when the hamming distance is $n/2+i$ and average error is guaranteed to be $\leq \rho$. By making ρ small enough we can get noticeable advantage $\Theta(\varepsilon\sqrt{n})$ in this region.

We now consider the region $HAM(x,y) \leq n/2$ and show that the absolute value of the contribution of this region can be made arbitrarily small w.r.t. $\varepsilon\sqrt{n}$ by appropriate choices of ρ, γ and T which will complete the proof. Let us break this region into three further regions :

(I) (x, y) with $n/2 - \gamma\sqrt{n} < HAM(x, y) \leq n/2$; (II) (x, y) with $n/2 - T\sqrt{n} \leq HAM(x, y) < n/2 - \gamma\sqrt{n}$; (III) (x, y) with $HAM(x, y) < n/2 - T\sqrt{n}$ for appropriately chosen T and γ. Denote by q_i the probability that π_n returns 1 on an input of hamming distance $n/2 - i$. The absolute value of the contribution of the region where $HAM(x, y) = n/2 - i$ is equal to

$$q_i \cdot (Pr_{\mu_0}[HAM(x, y) = n/2 - i] - Pr_{\mu_\varepsilon}[HAM(x, y) = n/2 - i]) =$$
$$q_i \cdot Pr_{\mu_0}[HAM(x, y) = n/2 - i] \cdot (1 - (1 - 4\varepsilon^2)^{n/2-i}(1 - 2\varepsilon)^{2i})$$

As before, $(1 - 4\varepsilon^2)^{n/2-i} \geq 1 - 2\varepsilon/n$. Thus in region (I) the negative contribution is bounded in absolute terms by:

$$\left(1 - \frac{2\varepsilon}{n}\right) \cdot \left(1 - (1 - 2\varepsilon)^{2\gamma\sqrt{n}}\right) < \frac{2\varepsilon}{n} + 4\gamma\varepsilon\sqrt{n}$$

In region (III) the contribution is again bounded by

$$\sum_{i=T\sqrt{n}}^{n/2} Pr_{\mu_0}[HAM(x, y) = n/2 - i] \cdot (1 - (1 - 2\varepsilon)^{2i}) < \sum_{i=T\sqrt{n}}^{n/2} Pr_{\mu_0}[HAM(x, y) = n/2 - i] \cdot 4i\varepsilon$$

By a standard Chernoff bound[1], the probability $Pr_{\mu_0}[HAM(x, y) = n/2 - i]$ is dominated by $e^{-\Omega(i^2/n)}$, and thus the sum can be made into an arbitrarily small multiple of $\varepsilon\sqrt{n}$ by

[1] See e,g [AS92].

choosing T large enough. For region (II) the advantage

$$\begin{aligned}
&\sum_{i=\gamma\sqrt{n}}^{T\sqrt{n}} q_i \cdot Pr_{\mu_0}[HAM(x,y) = n/2 - i] \cdot (1 - (1-2\varepsilon)^{2i}) \\
&\leq \sum_{i=\gamma\sqrt{n}}^{T\sqrt{n}} q_i \cdot Pr_{\mu_0}[HAM(x,y) = n/2 - i] \cdot 4i\varepsilon \\
&\leq 4T\varepsilon\sqrt{n} \sum_{i=\gamma\sqrt{n}}^{T\sqrt{n}} q_i \cdot Pr_{\mu_0}[HAM(x,y) = n/2 - i] \\
&\leq 4T\rho\varepsilon\sqrt{n}.
\end{aligned}$$

By making ρ small enough we can make the absolute contribution of this region small relative to $\varepsilon\sqrt{n}$. This completes the proof. □

Step 2: From tiny advantage to low-communication GHD.

We can now apply Lemma 3.3.2 together with Theorem 3.3.1 to show that a low-information solution to $GHD_{n,n/2,\gamma\sqrt{n}}$ with respect to the uniform distribution contradicts the communication complexity lower bound of Theorem 3.2.2.

Proof. (of Lemma 3.2.4). Assume for the sake of contradiction that for each α there are infinitely many n and a protocol π_n (different for each n) with $\mathsf{IC}_{\mathcal{U}}(\pi_n) < \alpha n$ and which solves $GHD_{n,n/2,\gamma\sqrt{n}}$ with error ρ, where the parameters γ and ρ are from Lemma 3.3.2. Let $N > \max(n^7, N(GHD_{n,n/2,\gamma\sqrt{n}}, \rho, \mathcal{U}, \delta_1, \delta_2))$, where $\delta_1 = 1$ and $\delta_2 = \varepsilon/2$, where ε is the error parameter in Theorem 3.2.2. Denote the protocol obtained from Theorem 3.3.1 (for solving cN copies of $GHD_{n,n/2,\gamma\sqrt{n}}, \rho, \mathcal{U}, \delta_1, \delta_2)$) as π'_{cN}.

Let $t = Pr_{(x,y)\sim\mathcal{U}}[\pi_n(x,y) = 1]$. W.l.o.g. we assume $t = 1/2$ (otherwise we can use a threshold$_{tcN}$ instead of majority in the protocol). Consider the protocol depicted in Figure 2.

Let us first analyze the success probability of the protocol Π_{nN}. We will do this in three steps:

Input: A pair $x, y \in \{0,1\}^{nN}$.
Output: $GHD_{n\cdot N, n\cdot N/2, \sqrt{n\cdot N}}$.

1. Create cN instances of GHD_n by sampling n random coordinates each time (with replacement): $(x_1, y_1), \ldots, (x_{cN}, y_{cN}) \in \{0,1\}^n \times \{0,1\}^n$.
2. Run π'_{cN} on $(x_1, y_1), \ldots, (x_{cN}, y_{cN})$ for $\frac{10\alpha cNn}{\varepsilon}$ steps, otherwise abort. (c and α are constants to be chosen later). π'_{cN} outputs answers $b_1, \ldots, b_{cN}$, one for each coordinate.
3. Return $MAJORITY(b_1, \ldots, b_{cN})$.

Protocol 2: The protocol $\Pi_{nN}(x, y)$

1. First let us analyze the success probability of Π_{nN} if we use π_n^{cN} in the second step i.e. π_n run independently on each coordinate. Suppose that the hamming distance between x and y is $nN/2 + \ell\sqrt{nN}$, where $\ell > 1$. Note that $\ell < n$ except with probability $e^{-\Omega(n^2)}$ (over the uniform distribution). The samples (x_i, y_i) are drawn iid according to the distribution $\mu_{\ell\cdot\sqrt{1/(nN)}}$. Since $N > n^7$ we have $\ell \cdot \sqrt{1/nN} < 1/n^2$. By Lemma 3.3.2, the output of π_n on each copy is thus $\tau \cdot \ell/\sqrt{N}$-biased towards 1. By Chernoff bounds, the probability that the protocol Π_{nN} outputs 1 is at least $1 - e^{-2\tau^2\ell^2 c}$.

2. Now let us analyze the success probability of Π_{nN} if we didn't abort in the second step. For each set of inputs, the statistical distance between the output of π'_{cN} and π_n^{cN} is at most $\varepsilon/2$, therefore, for (x, y) such that the hamming distance between x and y is $nN/2 + \ell\sqrt{nN}$, $1 < \ell < n$, Π_{nN} with no abort outputs 1 w.p. at least $1 - e^{-2\tau^2\ell^2 c} - \varepsilon/2$. The case when the hamming distance between x and y is $nN/2 - \ell\sqrt{nN}$ can be handled similarly.

3. Now let us analyze the success probability of Π_{nN}. Note that for each coordinate i, (x_i, y_i) is distributed according to the uniform distribution. Therefore the expected communication cost of π'_{cN} is less than $2\alpha cNn$. Therefore the probability that it exceeds $\frac{10\alpha cNn}{\varepsilon}$ is at most $\varepsilon/5$. Therefore the overall error of Π_{nN} is at most $e^{-2\tau^2\ell^2 c} +$

$\varepsilon/2 + \varepsilon/5 + 2e^{-\Omega(n^2)}$ which is less than ε for c and n large enough.

Now for α small enough, the communication cost of Π_{nN} can be made arbitrarily small w.r.t. nN which contradicts Theorem 3.2.2 since Π_{nN} solves $GHD_{n\cdot N, n\cdot N/2, \sqrt{n\cdot N}}$ with error $\leq \varepsilon$ w.r.t. the uniform distribution. Note that we got a randomized protocol for solving $GHD_{n\cdot N, n\cdot N/2, \sqrt{n\cdot N}}$ but we can fix the randomness to get a deterministic protocol.

□

3.3.3 The reduction from a small-gap instance to a large-gap instance

Now we complete the proof of Theorem 3.1.1 by providing the details of the reduction. We will start by proving a few technical lemmas.

Lemma 3.3.3. *Let $\alpha > 1$ be an integer. Let $\mathcal{U}_n$ be the uniform distribution over $\{0,1\}^n \times \{0,1\}^n$. Let $X, Y \sim \mathcal{U}_n$. Define a distribution μ over $\{0,1\}^{\alpha n} \times \{0,1\}^{\alpha n}$ by picking αn random coordinates of X, Y (with replacement) and then taking an XOR with a random string $r \in_R \{0,1\}^{\alpha n}$ (let U', V' be the strings obtained by sampling αn random coordinates of X, Y. Then $U = U' \oplus r, V = V' \oplus r$ are the final strings sampled). Then for all $\epsilon > 0$ and n large enough, there exists a constant M_ϵ and a distribution μ_ϵ such that*

1. $|\mu - \mu_\epsilon| \leq \epsilon$
2. $\mu_\epsilon \leq M_\epsilon \cdot \mathcal{U}_{\alpha n}$

Proof. It is easy to see that the distribution μ is symmetric w.r.t the hamming distance i.e. if $x, y \in \{0,1\}^{\alpha n} \times \{0,1\}^{\alpha n}$, and $x', y' \in \{0,1\}^{\alpha n} \times \{0,1\}^{\alpha n}$ such that $HAM(x,y) = HAM(x',y')$, then $\mu(x,y) = \mu(x',y')$. This is because μ is invariant under the application of a random permutation and a random XOR i.e. if $\pi \in_R S_n$ and $r' \in_R \{0,1\}^n$, then $\mu(x,y) = \mu(\pi(x \oplus r'), \pi(y \oplus r'))$. With a slight abuse of notation let $\mu(d)$ denote the probability mass on

strings of hamming distance d, and let $\mathcal{U}_{\alpha n}(d)$ denote the probability mass w.r.t the uniform distribution. Let $N = \alpha n$.

For $\epsilon > 0$, let μ_ϵ be the truncations of the distribution μ to the interval $[N/2 - C_\epsilon\sqrt{N}, N/2 + C_\epsilon\sqrt{N}]$ for a C_ϵ to be chosen later. Note that $|\mu - \mu_\epsilon| = |1 - \mu([N/2 - C_\epsilon\sqrt{N}, N/2+C_\epsilon\sqrt{N}])|$. So we will choose C_ϵ such that $|1-\mu([N/2-C_\epsilon\sqrt{N}, N/2+C_\epsilon\sqrt{N}])| \leq \epsilon$. By Chernoff bounds $Pr[HAM(X,Y) \notin [n/2 - \beta\sqrt{n}, n/2 + \beta\sqrt{n}]] \leq 2e^{-2\beta^2}$. Now if we pick N random coordinates distributed according to $B_{\frac{1}{2}+p}$, where $|p| \leq \beta/\sqrt{n}$, then the expected number of 1's $\in [N/2 - \beta\sqrt{\alpha}\sqrt{N}, N/2 + \beta\sqrt{\alpha}\sqrt{N}]$. Thus by another application of Chernoff bounds, we get that $Pr[HAM(U,V) \notin [N/2 - C_\epsilon\sqrt{N}, N/2 + C_\epsilon\sqrt{N}]] \leq 2e^{-2\beta^2} + 2e^{-2(C_\epsilon - \beta\sqrt{\alpha})^2}$. Now $\beta = \frac{1}{2}\ln(4/\epsilon)$ and $C_\epsilon = \frac{1}{2}\ln(4/\epsilon)(1+\sqrt{\alpha})$ suffices to ensure that $Pr[HAM(U,V) \notin [N/2 - C_\epsilon\sqrt{N}, N/2 + C_\epsilon\sqrt{N}]] \leq \epsilon$.

We will show that there exists a constant M_ϵ such that $\mu_\epsilon \leq M_\epsilon \cdot \mathcal{U}_{\alpha n}$. Note that by the symmetry properties of μ, it suffices to prove that for all d, $\mu_\epsilon(d) \leq M_\epsilon \cdot \mathcal{U}_{\alpha n}(d)$. Now

$$\begin{aligned}
\mu_\epsilon(d)/\mathcal{U}_{\alpha n}(d) &= \frac{1}{\mu([N/2 - C_\epsilon\sqrt{N}, N/2 + C_\epsilon\sqrt{N}])}\mu(d)/\mathcal{U}_{\alpha n}(d) \\
&\leq 2\mu(d)/\mathcal{U}_{\alpha n}(d) \\
&= 2\frac{\sum_{k=0}^{n}\binom{n}{k}\cdot 2^{-n}\cdot\binom{\alpha n}{d}\cdot\left(\frac{k}{n}\right)^d\cdot\left(\frac{n-k}{n}\right)^{N-d}}{\binom{\alpha n}{d}2^{-\alpha n}} \\
&= 2\cdot\sum_{k=0}^{n}\binom{n}{k}\cdot 2^{-n}\cdot\left(\frac{2k}{n}\right)^d\cdot\left(\frac{2(n-k)}{n}\right)^{N-d}
\end{aligned}$$

Let $d = N/2 + T$, where $|T| \leq C_\epsilon\sqrt{N}$. Also we will just concentrate on the sum for $k \geq n/2$. The lower half is analogous. Also it is easy to see that the sum from $k = 3n/4$ to $k = n$ is

small. So we consider

$$\sum_{k=n/2}^{3n/4} \binom{n}{k} \cdot 2^{-n} \cdot \left(\frac{2k}{n}\right)^{d} \cdot \left(\frac{2(n-k)}{n}\right)^{N-d}$$
$$= \sum_{k=n/2}^{3n/4} \binom{n}{k} \cdot 2^{-n} \cdot \left(\frac{2k}{n}\right)^{T} \cdot \left(\frac{2(n-k)}{n}\right)^{-T} \cdot \left(\frac{4k(n-k)}{n^2}\right)^{N/2}$$
$$\leq \sum_{k=n/2}^{3n/4} \binom{n}{k} \cdot 2^{-n} \cdot \left(\frac{k}{n-k}\right)^{T}$$

If $T < 0$, then we are done. So assume $T > 0$. For $n/2 \leq k \leq 3n/4$, $\frac{k}{n-k} = 1 + \frac{2k-n}{n-k} \leq 1 + \frac{8(k-n/2)}{n}$. For $k \leq n/2 + T$, the sum is small as $\frac{k}{n-k}$ is small. Otherwise $(1 + \frac{8(k-n/2)}{n})^T \lesssim (1 + \frac{8T}{n})^{k-n/2}$. Then the sum

$$\leq 2^{-n} \sum_{k=n/2+T}^{3n/4} \binom{n}{k} \cdot \left(1 + \frac{8T}{n}\right)^{k-n/2}$$
$$\leq 2^{-n} \sum_{k=n/2+T}^{3n/4} \binom{n}{k} \cdot \left(1 + \frac{8T}{n}\right)^{k-n/2} \left(1 - \frac{8T}{n}\right)^{n/2-k}$$
$$\leq 2^{-n} \sum_{k=n/2+T}^{3n/4} \binom{n}{k} \cdot \left(1 + \frac{8T}{n}\right)^{k} \left(1 - \frac{8T}{n}\right)^{n-k} \left(1 + \frac{8T}{n}\right)^{-n/2} \left(1 - \frac{8T}{n}\right)^{n/2}$$

Now $\sum_{k=n/2+T}^{3n/4} \binom{n}{k} \cdot \left(1 + \frac{8T}{n}\right)^{k} \left(1 - \frac{8T}{n}\right)^{n-k} \leq 2^n$ by binomial theorem, and

$$\left(1 + \frac{8T}{n}\right)^{-n/2} \left(1 - \frac{8T}{n}\right)^{n/2} = \left(1 - \frac{64T^2}{n^2}\right)^{-n/2}$$

is a constant, since $T \leq C_\epsilon \sqrt{N}$. This completes the proof. □

The next lemma relates the information cost of a protocol w.r.t two distributions that are close in statistical distance. We haven't seen the lemma in this specific form elsewhere.

Nevertheless it is not hard to prove.

Lemma 3.3.4. *Let $\epsilon < 1/2$. Let μ_1 and μ_2 be distributions on $\{0,1\}^N \times \{0,1\}^N$ such that $|\mu_1 - \mu_2| \leq \epsilon$, and fix a protocol π. Then $|IC(\pi,\mu_1) - IC(\pi,\mu_2)| \leq 4N\epsilon + 2H(2\epsilon)$. If ϵ is a constant and N large enough, then $|IC(\pi,\mu_1) - IC(\pi,\mu_2)| \leq 5N\epsilon$. In general for distributions over $\mathcal{X} \times \mathcal{Y}$, we get $|IC(\pi,\mu_1) - IC(\pi,\mu_2)| \leq 2\epsilon \log(|\mathcal{X}| \cdot |\mathcal{Y}|) + 2H(2\epsilon)$.*

Proof. We will design random variables X, Y, E such that $X, Y \in \{0,1\}^N$ and $E \in \{0,1,2\}$, $X, Y|E \in \{0,1\} \sim \mu_1$, $X, Y|E \in \{0,2\} \sim \mu_2$ and $Pr[E=1] = Pr[E=2] \leq \epsilon$. First let us see how this helps. Let Π denote the random variable for the transcript of the protocol when the inputs are X, Y. Let $X_1Y_1 \sim \mu_1$ and $X_2Y_2 \sim \mu_2$. Also let Π_1 and Π_2 denote the random variables for the transcript in these cases respectively.

$$\begin{aligned}
&I(\Pi; X|YE) \\
&= Pr[E=0] \cdot I(\Pi; X|Y, E=0) + Pr[E=1] \cdot I(\Pi; X|Y, E=1) \\
&+ Pr[E=2] \cdot I(\Pi; X|Y, E=2) \\
&= Pr[E \in \{0,1\}] \cdot I(\Pi; X|Y, E_{\{0,1\}}) + Pr[E=2] \cdot I(\Pi; X|Y, E=2)
\end{aligned}$$

Here conditioning on $E_{\{0,1\}}$ means that $E \in \{0,1\}$ and that both Alice and Bob know the value of E i.e. $I(\Pi; X|Y, E_{\{0,1\}}) = I(\Pi; X|Y, E, E \in \{0,1\})$. Now $I(\Pi; X|Y, E \in \{0,1\}) \leq I(\Pi; X|Y, E_{\{0,1\}}) + H(E|E \in \{0,1\}) = I(\Pi; X|Y, E_{\{0,1\}}) + C_1$, where $C_1 \leq H(\epsilon/(1-\epsilon)) \leq H(2\epsilon)$. Also $I(\Pi; X|Y, E=2) \leq N$ and $I(\Pi; X|Y, E \in \{0,1\}) = I(\Pi_1; X_1|Y_1)$. Thus

$$I(\Pi; X|YE) = (1 - Pr[E=2]) \cdot (I(\Pi_1; X_1|Y_1) + C_1) + Pr[E=2] \cdot C_2$$

where $C_1 \leq 1$ and $C_2 \leq N$. Similarly

$$I(\Pi; X|YE) = (1 - Pr[E=1]) \cdot (I(\Pi_2; X_2|Y_2) + C_3) + Pr[E=1] \cdot C_4$$

where $C_3 \leq H(2\epsilon)$ and $C_4 \leq N$. Equating the two we get that

$$(1 - Pr[E = 1]) \cdot (I(\Pi_1; X_1|Y_1) - I(\Pi_2; X_2|Y_2)) =$$
$$Pr[E = 1] \cdot (C_4 - C_3) + (1 - Pr[E = 1]) \cdot (C_2 - C1)$$

Since $Pr[E = 1] \leq \epsilon \leq 1/2$, we get that

$$|I(\Pi_1; X_1|Y_1) - I(\Pi_2; X_2|Y_2)| \leq 2N\epsilon + H(2\epsilon)$$

and hence $|IC(\pi, \mu_1) - IC(\pi, \mu_2)| \leq 4N\epsilon + 2H(2\epsilon)$.

Now let us see how to design random variables X, Y, E satisfying the given conditions. Let U, V, P denote the random variables obtained by sampling uniformly from $\{0,1\}^N \times \{0,1\}^N \times [0,1]$. Let G denote the event that $P < \max(\mu_1(U,V), \mu_2(U,V))$. Let $X, Y = U, V|G$. Also define a random variable $F \in \{0,1,2\}$ as follows :

- $F = 0$, if $P < \min(\mu_1(U,V), \mu_2(U,V))$
- $F = 1$, if $\mu_2(U,V) \leq P < \mu_1(U,V)$
- $F = 2$, if $\mu_1(U,V) \leq P < \mu_2(U,V)$

Now define $E = F|G$. Let us verify that X, Y, E satisfy the conditions.

$$Pr[X = x, Y = y|E \in \{0,1\}] = \frac{Pr[U = x, V = y, F \in \{0,1\}, G]}{Pr[F \in \{0,1\}, G]}$$
$$= \frac{\frac{1}{2^{2N}}\mu_1(x,y)}{\sum_{x,y}\frac{1}{2^{2N}}\mu_1(x,y)} = \mu_1(x,y)$$

Thus $X, Y | E \in \{0,1\} \sim \mu_1$. Similarly $X, Y | E \in \{0,2\} \sim \mu_2$. Also

$$\begin{aligned}
&Pr[E=1] = Pr[F=1|G] \\
&= \sum_{x,y} Pr[U=x, V=y|G] Pr[F=1|G, U=x, V=y] \\
&= \sum_{x,y \text{ s.t. } \mu_1(x,y) > \mu_2(x,y)} \frac{\frac{1}{2^{2N}} \max(\mu_1(x,y), \mu_2(x,y))}{\frac{1}{2^{2N}} \sum_{x,y} \max(\mu_1(x,y), \mu_2(x,y))} \cdot \frac{\mu_1(x,y) - \mu_2(x,y)}{\max(\mu_1(x,y), \mu_2(x,y))} \\
&= \frac{\sum_{x,y \text{ s.t. } \mu_1(x,y) > \mu_2(x,y)} (\mu_1(x,y) - \mu_2(x,y))}{\sum_{x,y} \max(\mu_1(x,y), \mu_2(x,y))}
\end{aligned}$$

Thus

$$Pr[E=1] = \frac{|\mu_1 - \mu_2|}{\sum_{x,y} \max(\mu_1(x,y), \mu_2(x,y))} \leq |\mu_1 - \mu_2| \leq \epsilon$$

Similarly

$$Pr[E=2] = \frac{|\mu_1 - \mu_2|}{\sum_{x,y} \max(\mu_1(x,y), \mu_2(x,y))}$$

Hence $Pr[E=1] = Pr[E=2] \leq \epsilon$. This completes the proof. The general form can be proved in a similar manner.

□

We also need a lemma which relates the information cost of distributions which are not very skewed w.r.t to each other. Formally

Lemma 3.3.5. *Let μ_1 and μ_2 be distributions over $\{0,1\}^N \times \{0,1\}^N$ such that $\mu_1 \leq M \cdot \mu_2$ for some constant M. Let f be a function (possibly partial) with domain $\{0,1\}^N \times \{0,1\}^N$ and let π be a protocol for solving it. Then $IC(\pi, \mu_1) \leq M \cdot IC(\pi, \mu_2)$.*

Proof. Let $X_1, Y_1 \sim \mu_1$ and Π_1 denote the random variable for the transcript when inputs are X_1, Y_1. Let $X_2, Y_2 \sim \mu_2$ and define Π_2 similarly. Now

$$I(\Pi_1; X_1 | Y_1) = \mathbb{E}_{x,y \sim \mu_1} D[\Pi_1|_{x,y} || \Pi_1|_y] = \mathbb{E}_y (\mathbb{E}_x D[\Pi_1|_{x,y} || \Pi_1|_y])$$

By Fact 1.2.11, $\mathbb{E}_x D[\Pi_1|_{x,y}||\Pi_1|_y] \leq \mathbb{E}_x D[\Pi_1|_{x,y}||\Pi_2|_y]$. Also $\Pi_1|_{x,y} = \Pi_2|_{x,y}$. Thus

$$\begin{aligned} I(\Pi_1; X_1|Y_1) \leq \mathbb{E}_{x,y\sim\mu_1} D[\Pi_2|_{x,y}||\Pi_2|_y] &\leq M \cdot \mathbb{E}_{x,y\sim\mu_2} D[\Pi_2|_{x,y}||\Pi_2|_y] \\ &= M \cdot I(\Pi_2; X_2|Y_2) \end{aligned}$$

Hence $IC(\pi, \mu_1) \leq M \cdot IC(\pi, \mu_2)$. □

The next lemma says that if the information cost w.r.t the distribution μ from Lemma 3.3.3 is high, then the information cost w.r.t the uniform distribution is high as well.

Lemma 3.3.6. *Let $f : \{0,1\}^N \times \{0,1\}^N \to \{0,1\}$ be a function (possibly partial). Let μ be a distribution over $\{0,1\}^N \times \{0,1\}^N$, as defined in Lemma 3.3.3. If $IC(f, \mu, \delta) \geq \Omega(N)$, for some $\delta > 0$, then $IC(f, \mathcal{U}_N, \eta) \geq \Omega(N)$, for some $\eta > 0$.*

Proof. Let π be a protocol for computing f with error η w.r.t. the distribution $\mathcal{U}_N$, and information cost $IC(\pi, \mathcal{U}_N) = I$. Let $\epsilon > 0$. Then by Lemma 3.3.3, for N large enough, there exists a distribution μ_ϵ over $\{0,1\}^N \times \{0,1\}^N$ such that $|\mu - \mu_\epsilon| \leq \epsilon$ and $\mu_\epsilon \leq M_\epsilon \cdot \mathcal{U}_N$ for some constant M_ϵ. Then error of the protocol π w.r.t. μ is $\leq M_\epsilon \eta + \epsilon$. Also the information cost of π w.r.t. μ is $\leq M_\epsilon I + 5N\epsilon$ (using Lemmas 3.3.4 and 3.3.5). Now if $M_\epsilon \eta + \epsilon \leq \delta$, then $M_\epsilon I + 5N\epsilon \geq c \cdot N$, for some constant c. Take $\epsilon = \min(\delta/2, c/10)$ and $\eta = (\delta - \epsilon)/M_\epsilon$. Then $I \geq cN/2M_\epsilon$. Thus $IC(f, \mathcal{U}_N, \eta) \geq \Omega(N)$. □

Proof. (of Theorem 3.1.1) Note that because of Lemma 3.3.6, we just need to prove that $IC(GHD_{N,N/2,\sqrt{N}}, \mu, \epsilon) = \Omega(N)$ for some $\epsilon > 0$ for the distribution μ in Lemma 3.3.3. Assume that for all $\epsilon > 0$, $IC(GHD_{N,N/2,\sqrt{N}}, \mu, \epsilon) = o(N)$. That is for all β, ϵ, and for N sufficiently large, $IC(GHD_{N,N/2,\sqrt{N}}, \mu, \epsilon) \leq \beta \cdot N$. By Lemma 3.2.4, there exist constants $\epsilon' > 0$, $\gamma > 0$ and $c > 0$ such that $IC(GHD_{n,n/2,\gamma\sqrt{n}}, \mathcal{U}, \epsilon') \geq c \cdot n$.

Let α be a large integer to be determined later. Set $N = \alpha \cdot n$. Let π_N be a protocol that solves $GHD_{N,N/2,\sqrt{N}}$ with error $\leq \epsilon$ w.r.t μ, and let the information cost of π_N w.r.t

μ be $\leq \beta \cdot N$. Consider the following protocol $\pi_n(x, y)$ for $GHD_{n,n/2,\gamma\sqrt{n}}$: Pick N random coordinates of x, y, call them u', v'. Now pick a random string $r \in_R \{0,1\}^N$ and set $u = u' \oplus r$ and $v = v' \oplus r$. Run π_N on u, v. Let $X, Y \sim \mathcal{U}_n$ be the inputs for π_n. Let U, V denote the random variables denoting the sampled coordinates. Note that $U, V \sim \mu$. Let Π denote the random variable for the transcript of running π_N on U, V. Then the transcript of running π_n on X, Y is ΠR, where R denotes the public randomness involved in sampling u, v from x, y. Now

$$I(\Pi R; X|Y) = I(R; X|Y) + I(\Pi; X|YR) = I(\Pi; X|YR) = I(\Pi; X|VYR)$$

The last equality follows from the fact that V is a deterministic function of YR. Now Π is a probabilistic function of U, V, and the internal randomness of the protocol π_N is independent of X, Y and R. Thus $I(\Pi; XYR|UV) = 0$, as

$$I(\Pi; XYR|UV) = I(\Pi; YR|UV) + I(\Pi; X|UVYR)$$

and $I(\Pi; YR|UV) = 0$, $I(\Pi; X|UVYR) = 0$. Applying Fact 1.2.13, with $A = \Pi$, $B = U$, $C = X$ and $D = VYR$, we get that $I(\Pi; X|VYR) \leq I(\Pi; U|VYR)$. Also $I(\Pi; YR|UV) = 0$. Applying Fact 1.2.12 with $A = U$, $B = \Pi$, $C = V$ and $D = YR$, we get $I(\Pi; U|V) \geq I(\Pi; U|VYR)$. This implies that $I(\Pi R; X|Y) \leq I(\Pi; U|V)$. A similar argument shows that $I(\Pi R; Y|X) \leq I(\Pi; V|U)$ and hence $IC(\pi_n, \mathcal{U}_n) \leq IC(\pi_N, \mu)$.

Now let us calculate the error of the protocol π_n. If $HAM(x, y) \geq n/2 + \gamma\sqrt{n}$, then for a random coordinate I, $Pr[x_I \oplus y_I = 1] \geq 1/2 + \gamma/\sqrt{n}$. Then the expected hamming distance of N random coordinates is $N/2 + \gamma\sqrt{\alpha}\sqrt{N}$. Hence the probability that the hamming distance is $\leq N/2 + \frac{\gamma\sqrt{\alpha}}{2}\sqrt{N}$ is bounded by $e^{-\frac{\alpha\gamma^2}{2}}$. The same holds for the probability that the

hamming distance is $\geq N/2 - \frac{\gamma\sqrt{\alpha}}{2}\sqrt{N}$. Choose α so that $\gamma\sqrt{\alpha} \geq 2$ and $e^{-\frac{\alpha\gamma^2}{2}} \leq \epsilon'/2$. Then

$$\begin{aligned}\text{error}(\pi_n) = &\sum_{x,y \text{ s.t. } HAM(x,y) \geq n/2+\gamma\sqrt{n}} \mathcal{U}_n(x,y) \cdot Pr[\pi_n \text{ outputs 0 on input } x,y] \\ &+ \sum_{x,y \text{ s.t. } HAM(x,y) \leq n/2-\gamma\sqrt{n}} \mathcal{U}_n(x,y) \cdot Pr[\pi_n \text{ outputs 1 on input } x,y]\end{aligned}$$

Now

$$Pr[\pi_n \text{ outputs 0 on input } x,y] = \sum_{u,v} \mu(u,v|x,y) \cdot Pr[\pi_N \text{ outputs 0 on input } u,v]$$

where $\mu(u,v|x,y)$ the probability of getting u,v when coordinates are sampled from x,y. For x,y s.t. $HAM(x,y) \geq n/2 + \gamma\sqrt{n}$,

$$\begin{aligned}&\sum_{u,v} \mu(u,v|x,y) \cdot Pr[\pi_N \text{ outputs 0 on input } u,v] \leq \\ &\sum_{u,v \text{ s.t. } HAM(u,v) \geq N/2+\sqrt{N}} \mu(u,v|x,y) \cdot Pr[\pi_N \text{ outputs 0 on input } u,v] + \epsilon'/2\end{aligned}$$

Doing a similar exercise for the other half, we get that

$$\begin{aligned}\text{error}(\pi_n) \leq &\sum_{u,v \text{ s.t. } HAM(u,v) \geq N/2+\sqrt{N}} \mu(u,v) \cdot Pr[\pi_N \text{ outputs 0 on input } u,v] + \\ &\sum_{u,v \text{ s.t. } HAM(u,v) \leq N/2-\sqrt{N}} \mu(u,v) \cdot Pr[\pi_N \text{ outputs 1 on input } u,v] + \epsilon'/2 \\ &= \text{error}(\pi_N) + \epsilon'/2\end{aligned}$$

Choosing $\epsilon = \epsilon'/2$, and $\beta = c/2\alpha$, we get a protocol π_n with error $\leq \epsilon'$ and information cost $\leq \beta\alpha n \leq cn/2$, which is a contradiction. □

3.4 Information Complexity of Inner Product

The inner product function $IP_n : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ is defined as follows:

$$IP_n(x,y) = \sum_{i=0}^{n} x_i y_i \pmod 2$$

The proof exploits the self-reducible structure of the inner-product function. But since, IP_n is such a sensitive function, we will first prove a statement about the 0-error information cost, and then use continuity of information cost to argue about non-zero errors.

We will need the following lemma from [BR11]. It is essentially the same as Theorem 3.3.1, only when dealing with 0 error, we cannot ensure that error on each copy is 0. We just have an overall error which is the error introduced if compression fails.

Lemma 3.4.1. *Let $f : X \times Y \to \{0,1\}$ be a function, and let μ be a distribution over the inputs. Let π be a protocol computing f with error 0 w.r.t μ, and internal information cost $IC_\mu(\pi) = I$. Then for all $\delta > 0$, $\epsilon > 0$, there is a protocol π_n for computing f^n with error ϵ w.r.t μ^n, with worst case communication cost*

$$\begin{aligned} &= n(I + \delta/4) + O(\sqrt{CC(\pi) \cdot n \cdot (I + \delta/4)}) + O(\log(1/\epsilon)) + O(CC(\pi)) \\ &\leq n(I + \delta) \text{ (for } n \text{ sufficiently large)} \end{aligned}$$

The following lemma from [BBCR10] relates the information cost of computing XOR of n copies of a function f to the information cost of a single copy.

Lemma 3.4.2. *Let f be a function, and let μ be a distribution over the inputs. Then $IC_{\mu^n}(\oplus_n f, \epsilon) \geq n(IC_\mu(f, \epsilon) - 2)$.*

The next lemma says that there is no 0-error protocol for IP_n which conveys slightly less information than the trivial protocol.

Lemma 3.4.3. *For all $n, IC_{\mathcal{U}_n}(IP_n, 0) \geq n$, where $\mathcal{U}_n$ is the uniform distribution over $\{0,1\}^n \times \{0,1\}^n$.*

Proof. It is known that $D_\epsilon^{\mathcal{U}_n}(IP_n) \geq n - c_\epsilon$, for all constant $\epsilon \in (0, 1/2)$, where c_ϵ is a constant depending just on ϵ [KN97, CG88]. Assume that for some n, $IC_{\mathcal{U}_n}(IP_n, 0) \leq n - c$. Then using, Lemma 6.4.2 with $\delta = c/2$ and $\epsilon = 1/3$, we can get a protocol π for solving N copies of IP_n with overall error $1/3$ w.r.t $\mathcal{U}_n^N$, and $CC(\pi) \leq N(n - c + c/2)$. This gives us a protocol π' for solving IP_{Nn} with error $1/3$ w.r.t the uniform distribution, and $CC(\pi') \leq Nn - Nc/2$ (divide the inputs into N chunks, solve the N chunks using π and XOR the answers). But $CC(\pi') \geq Nn - c_{1/3}$, a contradiction. □

Proof. (of Theorem 3.1.2) We use the continuity of (internal) information cost in the error parameter at $\epsilon = 0$:

Theorem 3.4.4 ([BGPW13a])**.** *For all $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ and $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$ we have*

$$\lim_{\epsilon \to 0} \mathrm{IC}_\mu(f, \epsilon) = \mathrm{IC}_\mu(f, 0). \tag{3.3}$$

Given $\delta > 0$, let $l = \lceil \frac{3}{\delta} \rceil$. Then

$$IC_{\mathcal{U}_l}(IP_l, 0) \geq l \geq (1 - \delta)l + 3.$$

Since $\lim_{\epsilon \to 0} IC_{\mathcal{U}_l}(IP_l, \epsilon) = IC_{\mathcal{U}_l}(IP_l, 0)$, there exists $\epsilon(l, \delta) = \epsilon(\delta)$ s.t.

$$IC_{\mathcal{U}_l}(IP_l, \epsilon) \geq (1 - \delta)l + 2.$$

Now using Lemma 3.4.2, we get that $IC_{\mathcal{U}_l^N}(\oplus_N IP_l, \epsilon) \geq (1 - \delta)Nl$. Thus $IC_{\mathcal{U}_{Nl}}(IP_{Nl}, \epsilon) \geq (1 - \delta)Nl$. Thus for sufficiently large n, $IC_{\mathcal{U}_n}(IP_n, \epsilon) \geq (1 - \delta)n$. □

Chapter 4

Public vs Private Coins in Information Complexity

The results in this chapter are based on joint work with Mark Braverman [BG14].

4.1 Introduction

In this chapter we investigate the role of private randomness in the ability of two parties to communicate while revealing as little information as possible to each other – i.e. to communicate at low information cost. More specifically, Alice and Bob are given possibly correlated inputs X and Y and need to perform a task T by means of a communication protocol π. Alice and Bob share a public random string R; in addition they have access to private random strings R_A and R_B, respectively. The *information cost* of π with respect to a distribution $(X, Y) \sim \mu$ is the quantity (it is not hard to see that this the same as the definition of $\mathsf{IC}_\mu(\pi)$ in Section 1.3

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

where $\Pi = \Pi(X, Y, R, R_A, R_B)$ is the random variable representing the transcript of the protocol.

It is not hard to see that if the goal is to solve a task T while minimizing the information cost of the protocol, we can always avoid using the public randomness string R: to simulate public randomness, before the beginning of the protocol's execution, Alice can send a portion of R_A, which will be used as R for the remainder of the protocol. This modification increases the communication cost of the protocol, but it is not hard to see that it does not change its information cost. Therefore, in the context of information complexity, private randomness is at least as good as public randomness. Is the converse true? In other words, can any protocol π that uses private randomness be simulated by a protocol π' which uses only public randomness so that $\mathsf{IC}_\mu(\pi') \leq \mathsf{IC}_\mu(\pi)$? The naïve "solution" to this problem would be to simulate π by using the public randomness to simulate private randomness. The following simple example shows why this approach fails. Consider the protocol π in which $X \in \{0,1\}^n$. Alice samples a uniformly random string $R_A \in_U \{0,1\}^n$, and sends the bitwise XOR $M := X \oplus R_A$ to Bob. This protocol conveys 0 information to Bob about X. However, if the public randomness R were to be used to produce R_A, then Bob would also know R_A, and thus the message M reveals $X = M \oplus R_A$ to Bob – drastically increasing the information cost of the protocol. This, of course, does not mean that a more sophisticated simulation scheme cannot work.

It is instructive to compare this question to the public-vs-private randomness question in randomized *communication complexity*. In the context of communication complexity the situation is somewhat reversed: it is obvious that public randomness can be used to simulate private randomness: the parties can always designate part of their public randomness as "private randomness". This will not affect the communication cost of the protocol (although, as seen above, it may affect its information cost). In the reverse direction, Newman [New91] showed that $R^{\text{priv}}_{\epsilon+\delta}(f) \leq R^{\text{pub}}_{\epsilon}(f) + O(\log(\frac{n}{\delta}))$. Thus, up to an additive $\log n$, pri-

vate randomness replaces public randomness in communication complexity. Does a "reverse Newman theorem" hold for information complexity? Can private randomness be replaced with public randomness at a small cost?

This question has been considered by Brody et al. in [BBK+13], which showed a version of the private-by-public simulation for one-round protocols. In the one-round setting, Alice wishes to send Bob her message – a random variable $M = M(X, R_A)$. Obviously, the information cost of this task is just $I(M; X|Y)$. If Bob receives no input, then it is just $I(M; X)$. In this chapter we prove tight bounds on the one round private-by-public simulation. Specifically, we show that the cost of simulating a message M of information cost I without the use of private randomness is between I and $I + \log I \pm O(1)$, and that the upper bound is in fact tight in some cases. Previously, [BBK+13] showed a weaker translation to information cost of at most $I + O(\log n)$, where $n = \max(\log |\mathcal{X}|, \log |\mathcal{Y}|)$ – the log of the sizes of the domains of X and Y. Note that it is always the case that $I \leq H(X) \leq \log |\mathcal{X}| \leq n$, and therefore $\log I \leq \log n$. Our lower bound example shows that even if dependence on n is allowed, one cannot do with less than $\log n$ additive overhead.

It is interesting to consider the connection between the problem of simulating a protocol without private randomness, and the problem of compressing communication protocols. The general protocol compression problem [BBCR10, Bra12] is the problem of simulating a protocol π with communication cost C and information cost I with a protocol π' of communication cost C' that is as close to I as possible. The problem of compressing interactive communication is essentially equivalent to the direct sum problem for randomized communication complexity [BR11]. The best known general compression results gives $C' = \tilde{O}(\sqrt{I \cdot C})$, and despite the recent breakthrough results of [GKR14a, GKR15], it is still wide open whether $C' = O(I \cdot (\log C)^{O(1)})$ is possible. It has been shown in [BBK+13] (and independently in [Pan12]) that if a protocol π does not use private randomness, then it can be compressed to $O(I \cdot (\log C)^{O(1)})$. Thus a way to replace private randomness with public randomness for

unbounded-round protocols would imply a substantial improvement in the state-of-the-art on protocol compression.

Another interesting connection between removing private randomness and compression is in the context of one-message protocols. In the setting where Bob has no input Y, the information cost of sending a message M is just $I := I(M; X)$. Harsha et al. [HJMR07] showed how to simulate such a transmission using $I + O(\log I)$ bits of (expected) *communication* (with access to public randomness). Their work left open the interesting question of whether the additive $O(\log I)$ is necessary. As noted above, a communication protocol with communication C can always be simulated by a protocol with same communication and only public randomness. As information cost is bounded from above by communication cost, a compression scheme is in particular a private-by-public scheme. Thus our lower bound gives an example showing that the $O(\log I)$ additive overhead in [HJMR07] is necessary.

Results and techniques

Our main result gives an upper and lower bound on simulating private randomness by public randomness for one-message protocols.

Theorem 4.1.1. *Let X, Y be inputs to Alice and Bob respectively distributed according to a distribution μ. Alice and Bob have access to public randomness R', and Alice has access to private randomness R_A. Let π be a protocol where Alice sends a message $M = M(X, R', R_A)$ to Bob, so that the information cost of π is $I := I(X; M|YR')$. Then*

1. *π can be simulated by a one-message public-coin protocol π' such that $\mathsf{IC}_\mu(\pi') \leq I + \log I + O(1)$.*

2. *for each I, there is an example with no Y (i.e. Bob has no "private" knowledge), and no R', such that if $I := I(X; M)$, then any public-coin protocol π' simulating the transmission of M must have information cost of at least $I + \log I - O(1)$.*

Thus, up to an additive constant, our bounds are tight. Note that while the upper bound holds under the most general conditions, for the lower bound it is sufficient to consider protocols without Y (this is the type of protocols considered, for example, in [HJMR07]).

Both the upper and lower bound require some careful analysis. For the upper bound, a natural variant of the one-round compression scheme of Braverman and Rao [BR11] is used. The main challenge is in analyzing the information cost of the resulting public randomness protocol: we need to prove that Bob does not learn too much about X from Alice's message. Suppose that given X and the public randomness R of the simulating protocol, Alice's message in the simulating protocol is $S = S(X, R)$. Observe that in this case

$$I(S; X|YR) = H(S|YR) - H(S|XYR) = H(S|YR).$$

To establish an upper bound on $H(S|YR)$, we show how, someone knowing X, Y and R, can describe S to Bob using a message M' (i.e. $H(S|M'YR) = 0$) such that

$$H(M') \leq I + \log(I) + O(1)$$

Noting that this expression is an upper bound for $H(S|YR)$, completes the proof.

To prove the lower bound, we give a family of specific examples whose information cost necessarily increases by $\log I - O(1)$ when private randomness is replaced with public randomness. Details of the construction are given in Section 4.3, here we only give the high level idea for why the information cost increases in lieu of private randomness. Consider the following example: Alice knows a secret random string $PASS$ of 128 bits (which we can think of as her password). She wants to send Bob a message M such that $M = PASS$ with probability 1/2 and $M = RANDOM$ with probability 1/2 – that is, half of the time she sends her password and half the time she sends a random 128-bit string. The message M reveals approximately 63 bits of information about $PASS$. To see this, note that given M the

posterior distribution of $PASS$ puts mass $1/2$ on M and mass $1/2$ on the remaining $2^{128}-1$ strings. The entropy of this distribution is $\approx \frac{1}{2}\cdot 1+\frac{1}{2}\cdot 129 = 65$, down from the prior entropy of 128. Thus $I(M;PASS) \approx 128-65 = 63$ bits. One might have expected this number to be 64 bits. Indeed, if Alice had told Bob which of the two cases has occurred, M would reveal $\frac{1}{2}\cdot 128+\frac{1}{2}\cdot 0 = 64$ bits of information. However, not knowing whether Alice's message is the password or a random string "saves" one bit in information cost. Now suppose Alice was not allowed to use private randomness. Then, intuitively, the public random string R should reveal to Bob whether $M = PASS$ or $M = RANDOM$. Therefore, the information cost of a public-randomness protocol increases to 64 bits. Generalizing from this example, we construct a situation where Alice sends a binary message M of length n and information cost $I \approx n/2 - \log n$, so that any public randomness simulation of M requires information cost of $\geq n/2 - O(1) = I + \log I - O(1)$ – demonstrating the desired gap.

Let us have a look at another example. Suppose that Alice gets a bit $X \sim B_{\frac{1}{2}}$ and she wants to transmit this bit to Bob with error $\frac{1}{2}-\epsilon$. Consider a private-coin protocol in which Alice samples a $B_{\frac{1}{2}+\epsilon}$ bit R. She sends X if $R = 1$ and a $\neg X$ if $R = 0$. Clearly the protocol performs the task of transmitting the bit with error $\frac{1}{2}-\epsilon$. Let Π denote the random variable for Alice's message. The information cost of this protocol is

$$I(\Pi;X) = \frac{1}{2}D(\Pi_0||\Pi)+\frac{1}{2}D(\Pi_1||\Pi) = \frac{1}{2}D(1/2-\epsilon||1/2)+\frac{1}{2}D(1/2+\epsilon||1/2) = \frac{2}{\ln 2}\epsilon^2 \pm o(\epsilon^2)$$

However if we don't allow private coins, then the information complexity of this task is $\geq 2\epsilon$. To see this consider a public-coin protocol that transmits X with error probability $\leq \frac{1}{2}-\epsilon$. It is basically a function $f : \{0,1\}\times\mathcal{R} \to \{0,1\}$ (in case Alice sends a longer message and then Bob applies a deterministic function to that, f could be the composition of those two functions) such that $\mathbb{E}_{r\sim\mathcal{R}}[f(0,r)] = \frac{1}{2}-\epsilon$ and $\mathbb{E}_{r\sim\mathcal{R}}[f(1,r)] = \frac{1}{2}+\epsilon$. Then

$Pr_{r\sim\mathcal{R}}[f(1,r)=1, f(0,r)=0] \geq 2\epsilon$. Hence

$$I(f(X,R);X|R) = H(f(X,R)|R) = \mathbb{E}_{r\sim\mathcal{R}} H(f(X,r)) \geq 2\epsilon$$

since if $f(1,r)=1, f(0,r)=0$, then $H(f(X,r))=1$. This example, in some sense, highlights the information-cost advantage one gains from having access to private randomness.

Open problems

Our lower bound example is really about the simulation of a protocol and not about solving a boolean function. So it will be nice to get a 1-round gap for a boolean function. Also it would be nice to get a bigger separation between r-round public-coin information complexity and private-coin information complexity, where r is a constant. Note that using the 1-round example, we can also construct a 2-round example by requiring both Alice and Bob to perform the 1-round task. It is quite possible that the example in [GKR15] has an exponential separation between private and public coin information complexities. The compression result of [BMY15] implies this separation for the example in [GKR15] conditioned on some bound on the communication cost of the public coin protocol. It would be nice to get the separation unconditionally.

1. Does there exist a boolean function f for which 0-error private-coin information complexity is I but 0-error public-coin information complexity is $\geq I + \log(I) - O(1)$?

2. Does there exist a (family of) 3-round private-coin protocol(s) π such that information cost of π is I but any 3-round public-coin protocol simulating π has information cost $\geq I + 3\log(I) - O(1)$?

3. Get an exponential separation between private and public coin information complexi-

ties of some boolean function (or even a relation).

4.2 Upper Bound

Definition 4.2.1. We will say that a (randomized) protocol ϕ *simulates* a protocol π if there is a deterministic function g such that $g(\Phi(x, y, R^\phi, R^\phi_A, R^\phi_B))$ is equal in distribution to $\Pi(x, y, R^\pi, R^\pi_A, R^\pi_B)$, $\forall x, y$. Here $R^\phi, R^\phi_A, R^\phi_B$ are the public and private randomness of protocol ϕ and Φ is the random variable for the transcript. Similarly for π.

Theorem 4.2.2. *Let X, Y be inputs to Alice and Bob respectively distributed according to a distribution μ. Alice and Bob have access to public randomness R', and Alice has access to private randomness R_A. Let π be a protocol where Alice sends a message $M = M(X, R', R_A)$ to Bob, so that the information cost of π is $I := I(X; M|YR')$. Then π can be simulated by a one-message public-coin protocol π' such that $\mathsf{IC}_\mu(\pi') \leq I + \log I + O(1)$.*

Proof. We can assume wlog that R' is a part of M, since $I(X; M|YR') = I(X; MR'|Y)$. Let $\mathcal{U}$ be the message space of the message M. Consider the protocol π' defined in Figure 3.

1. Using public randomness, Alice and Bob get samples $\{(u_i, p_i)\}_{i\geq 1}$, where (u_i, p_i) uniformly sampled from $\mathcal{U} \times [0, 1]$.
2. Let P denote the distribution $M_x = M|_{X=x}$ and Q denote the distribution $M_y = M|_{Y=y}$. Alice sends Bob the index of the first sample, s, such that $p_s < P(u_s)$. Bob decodes this message as being u_s

Protocol 3: Protocol π'

It is clear that Bob's decoding of the Alice's message on input x is distributed according to $P = M_x$. What remains is to analyze the information cost of the protocol. Let R denote the random variable for the public randomness and let S denote the random variable for Alice's

message (the index). Then

$$I(S;X|YR) = H(S|YR) - H(S|XYR) = H(S|YR)$$

because S is determined by X and R. It seems difficult to get a handle on $H(S|YR)$, but we can use the following trick : If someone (who knows X, Y, R) can describe to Bob S using a message M' (i.e. S is fixed given M', Y and R), then $H(S|YR) \leq H(M')$. This is because :

$$H(S|YR) + H(M'|SYR) = H(M'S|YR) = H(M'|YR) + H(S|M'YR) = H(M'|YR)$$
$$\leq H(M')$$

Note that since Alice doesn't know Y, she won't be able to compute M' and hence it does not seem possible for Alice to send the message M using a low communication protocol. To achieve low communication, interaction seems necessary, and this problem has been well studied in [BR11] and [BRWY13a]. Now let us describe the message M'. Let P denote the distribution M_x and Q denote the distribution M_y. The message will consist of three parts. The first part would be $k = \lceil \frac{S}{|\mathcal{U}|} \rceil$. The second part would consist of the ceiling of the Q-height of the S^{th} sample i.e. $t = \lceil \frac{p_S}{Q(u_S)} \rceil$. The third part would consist of the index l of the sample Alice wants to send among indices $\{(k-1) \cdot |\mathcal{U}| + 1, \ldots, k \cdot |\mathcal{U}|\}$ that have Q-height between $t-1$ and t.

Now let us look at $\mathbb{E}[|M'||X = x, Y = y]$. We'll analyze the lengths of the three different parts of M' separately.

1. For (u, p) randomly sampled from $\mathcal{U} \times [0, 1]$,

$$Pr[p < P(u)] = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} P(u) = \frac{1}{|\mathcal{U}|}$$

Thus $Pr[S > r \cdot |\mathcal{U}|] = \left(1 - \frac{1}{|\mathcal{U}|}\right)^{r \cdot |\mathcal{U}|} \leq e^{-r}$. Thus $Pr[k > r] \leq e^{-r}$. Thus

$$\mathbb{E}[k] = \sum_{r=0}^{\infty} Pr[k > r] \leq 1 + \frac{1}{e} + \frac{1}{e^2} + \ldots = O(1)$$

Hence $\mathbb{E}[\lceil \log(k) \rceil] = O(1)$.

2. For the S^{th} sample, $p_S < P(u_S)$. Thus $\mathbb{E}[\lceil \log(t) \rceil] \leq \mathbb{E}\left[\log\left(\frac{p}{Q(u)} + 1\right) | p < P(u)\right]$. Since $\log(x+1) - \log(x) \leq \frac{\log(e)}{x}$ (by Lagrange's Mean Value Theorem),

$$\log\left(\frac{p}{Q(u)} + 1\right) \leq \log\left(\frac{p}{Q(u)}\right) + O\left(\frac{Q(u)}{p}\right)$$

Thus

$$\begin{aligned}
&\mathbb{E}\left[\log\left(\frac{p}{Q(u)} + 1\right) | p < P(u)\right] \\
&= \sum_{u \in \mathcal{U}} P(u) \cdot \left(\frac{1}{P(u)} \int_0^{P(u)} \log\left(\frac{p}{Q(u)} + 1\right) \, \mathrm{d}u\right) \\
&\leq \sum_{u \in \mathcal{U}} P(u) \cdot \left(\frac{1}{P(u)} \int_0^{P(u)} \log\left(\frac{P(u)}{Q(u)} + 1\right) \, \mathrm{d}u\right) \\
&\leq \sum_{u \in \mathcal{U}} P(u) \cdot \left(\frac{1}{P(u)} \int_0^{P(u)} \log\left(\frac{P(u)}{Q(u)}\right) + O\left(\frac{Q(u)}{P(u)}\right) \, \mathrm{d}u\right) \\
&= D(P||Q) + O(1)
\end{aligned}$$

Hence $\mathbb{E}[\lceil \log(t) \rceil] \leq D(P||Q) + O(1)$.

3. For (u, p) randomly sampled from $\mathcal{U} \times [0, 1]$,

$$\begin{aligned}
&Pr[(t-1) \cdot Q(u) < p \leq t \cdot Q(u) | p > P(u)] \\
&\leq Pr[(t-1) \cdot Q(u) < p \leq t \cdot Q(u)] / Pr[p > P(u)] \\
&= Pr[(t-1) \cdot Q(u) < p \leq t \cdot Q(u)] / (1 - \frac{1}{|\mathcal{U}|}) \\
&\leq 2Pr[(t-1) \cdot Q(u) < p \leq t \cdot Q(u)] \\
&= \frac{2}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} Q(u) = \frac{2}{|\mathcal{U}|}
\end{aligned}$$

   Thus among the indices $\{(k-1) \cdot |\mathcal{U}| + 1, \ldots, k \cdot |\mathcal{U}|\}$, in expectation, there are a constant number that have Q-height between $t-1$ and t. Thus $\mathbb{E}[\lceil \log(l) \rceil] = O(1)$.

   Note that for the darts appearing before the dart S, the probability of appearing in some region increases slightly, since they are conditioned on not falling under the histogram of P but the probability increases at most by a factor of 2.

Hence $\mathbb{E}[|M'||X = x, Y = y] \leq D(M_x||M_y) + O(1)$. Now

$$\mathbb{E}[|M'|] = \mathbb{E}_{x,y}\left[\mathbb{E}[|M'||X = x, Y = y]\right] \leq \mathbb{E}_{x,y}[D(M_x||M_y)] + O(1) = I(M; X|Y) + O(1)$$

Now we will use the following lemma to bound $H(M')$.

Lemma 4.2.3. *Let P be a distribution on the natural numbers such that $\sum_{n \geq 1} P_n \cdot \lceil \log(n) \rceil = I$. Then $H(P) \leq I + \log(I) + O(1)$.*

The lemma says that if the expected length of the numbers is bounded by I, then the entropy is bounded by $I + \log(I) + O(1)$. A bound of $I + 2\log(I) + O(1)$ or of $I + \log(I) + 2\log(\log(I)) + O(1)$ is easy to get via prefix-free encoding of integers, but the fact that we can bound the entropy by $I + \log(I) + O(1)$ is somewhat surprising.

Using the lemma, we get that $H(M') \leq I(M;X|Y) + \log(I(M;X|Y)) + O(1)$, and thus $I(S;X|YR) \leq I(M;X|Y) + \log(I(M;X|Y)) + O(1)$. It remains to prove the lemma.

Proof. (Of Lemma 4.2.3) Let p_i be the probability mass on the integers between 2^{i-1} and 2^i i.e. $p_i = \sum_{n=2^{i-1}+1}^{2^i} P_n$. Then $\sum_{i=1}^{\infty} i \cdot p_i = \sum_{n\geq 1} P_n \cdot \lceil \log(n) \rceil = I$.

$$H(P) = \sum_{n\geq 1} P_n \log\left(\frac{1}{P_n}\right) \leq P_1 \log\left(\frac{1}{P_1}\right) + \sum_{i\geq 1} p_i \log\left(\frac{2^{i-1}}{p_i}\right) = I \pm O(1) + H(p)$$

The inequality follows from log-sum inequality,

$$\sum_k a_k \log\left(\frac{a_k}{b_k}\right) \geq \left(\sum_k a_k\right) \log\left(\frac{\sum_k a_k}{\sum_k b_k}\right)$$

Then, $\sum_{n=2^{i-1}+1}^{2^i} P_n \log(P_n) \geq p_i \log(p_i/2^{i-1})$. Now let q_j be the probability mass of p_i from $2^{j-1}+1$ to 2^j i.e. $q_j = \sum_{i=2^{j-1}+1}^{2^j} p_i$. Then $\sum_{i\geq 1} i \cdot p_i \geq \sum_{j\geq 1} 2^{j-1} \cdot q_j$. Thus $\sum_{j\geq 1} 2^j \cdot q_j \leq 2I$. Again by the log-sum inequality,

$$H(p) \leq p_1 \log\left(\frac{1}{p_1}\right) + \sum_{j\geq 1} q_j \log\left(\frac{2^{j-1}}{q_j}\right) + O(1) = \sum_{j\geq 1} j \cdot q_j + H(q) \pm O(1)$$

We can assume wlog that I is a power of 2. If $j = \log(2I) + k$, for $k \geq 2$, then $q_j \leq \frac{1}{2^k}$ and hence $q_j \log\left(\frac{1}{q_j}\right) \leq \frac{k}{2^k}$, since $q \log\left(\frac{1}{q}\right)$ is increasing in the interval $(0, \frac{1}{e}]$. Thus $\sum_{j>\log(2I)} q_j \log\left(\frac{1}{q_j}\right) = O(1)$. Let $q = \sum_{j>\log(2I)} q_j$. Since $q_{\log(2I)+k} \leq \frac{1}{2^k}$, $\sum_{j>\log(2I)} j \cdot q_j \leq q \cdot \log(2I) + O(1)$. So all that is needed is to prove that

$$\sum_{j\leq \log(2I)} j \cdot q_j + \sum_{j\leq \log(2I)} q_j \log\left(\frac{1}{q_j}\right) \leq (1-q) \cdot \log(2I) + O(1)$$

Let us look at $j \cdot q_j + \log(2I) \cdot q_{\log(2I)} + q_j \log\left(\frac{1}{q_j}\right) + q_{\log(2I)} \log\left(\frac{1}{q_{\log(2I)}}\right)$. If we decrease q_j and increase $q_{\log(2I)}$ by the same amount, the rate at which $j \cdot q_j + \log(2I) \cdot q_{\log(2I)}$ increases

is $\log(2I) - j$. Also $\left(q \log\left(\frac{1}{q}\right)\right)' = \log(e) \cdot \left(\ln\left(\frac{1}{q}\right) - 1\right)$. The difference in rates for $q_{\log(2I)}$ and q_j is $\log\left(\frac{1}{q_{\log(2I)}}\right) - \log\left(\frac{1}{q_j}\right)$. So as long as

$$\log\left(\frac{1}{q_j}\right) - \log\left(\frac{1}{q_{\log(2I)}}\right) \leq \log(2I) - j$$

increasing $q_{\log(2I)}$ and decreasing q_j (by the same amount) will increase $\sum_{j \leq \log(2I)} j \cdot q_j + \sum_{j \leq \log(2I)} q_j \log\left(\frac{1}{q_j}\right)$. Thus we can assume wlog that, $q_j \leq \frac{q_{\log(2I)}}{2^{\log(2I)-j}}$. Now for these values of q_j, it is easy to check that $\sum_{j \leq \log(2I)} q_j \log\left(\frac{1}{q_j}\right) = O(1)$. Also $\sum_{j \leq \log(2I)} j \cdot q_j \leq (1-q) \cdot \log(2I)$ is trivially true. This completes the proof. Note that it is not always true that $\sum_{j \leq \log(2I)} q_j \log\left(\frac{1}{q_j}\right) = O(1)$ but for the distribution maximizing $\sum_{j \leq \log(2I)} j \cdot q_j + \sum_{j \leq \log(2I)} q_j \log\left(\frac{1}{q_j}\right)$, this is true. □

□

We mention a few easy corollaries :

Corollary 4.2.4. *Let X, Y be inputs to Alice and Bob respectively distributed according to a distribution μ. Suppose that π is a private-coin r-round protocol with information cost $IC_\mu(\pi) = I$. Then π can be simulated by a r-round public-coin protocol π' with information cost $IC_\mu(\pi') \leq I + r\log(I/r) + O(r)$.*

Proof. It follows by applying Theorem 4.2.2 to the messages round by round. Denote the protocol transcript by $\Pi = \Pi_1, \Pi_2, \ldots, \Pi_r$. Assume Alice and Bob send alternate messages with Alice sending Π_1. Then

$$\begin{aligned}
IC_\mu(\pi) &= I(\Pi; X|YR'R_B) + I(\Pi; Y|XR'R_A) \\
&= \sum_{i \leq r} I(\Pi_i; X|Y\Pi_1\Pi_2 \ldots \Pi_{i-1}R'R_B) + \sum_{i \leq r} I(\Pi_i; Y|X\Pi_1\Pi_2 \ldots \Pi_{i-1}R'R_A) \\
&= \sum_{i \text{ odd}, i \leq r} I(\Pi_i; X|Y\Pi_1\Pi_2 \ldots \Pi_{i-1}R'R_B) + \sum_{i \text{ even}, i \leq r} I(\Pi_i; Y|X\Pi_1\Pi_2 \ldots \Pi_{i-1}R'R_A)
\end{aligned}$$

The second equality is chain rule for mutual information and the last equality follows from the fact that for odd i, Π_i is a function of $\Pi_1\Pi_2\ldots\Pi_{i-1}$ and X and for even i, Π_i is a function of $\Pi_1\Pi_2\ldots\Pi_{i-1}$ and Y. Now, after the messages $\Pi_1 = m_1, \Pi_2 = m_2, \ldots, \Pi_{i-1} = m_{i-1}$ have been sent (assume i odd), Alice can send Π_i using public randomness via a message Π_i' and public randomness R such that (apply Theorem 4.2.2 to the inputs $XY|\Pi_1 = m_1, \Pi_2 = m_2\ldots\Pi_{i-1} = m_{i-1}$)

$$I(\Pi_i'; X|Y, \Pi_1 = m_1, \Pi_2 = m_2, \ldots, \Pi_{i-1} = m_{i-1}, R) \leq I(\Pi_i; X|Y\Pi_1 = m_1, \Pi_2 = m_2\ldots\Pi_{i-1} = m_{i-1}R'R_B) + \log(I(\Pi_i; X|Y\Pi_1 = m_1, \Pi_2 = m_2, \ldots\Pi_{i-1} = m_{i-1}R'R_B)) + O(1)$$

This gives by taking expectations and by concavity of log

$$I(\Pi_i'; X|Y\Pi_1\Pi_2\ldots\Pi_{i-1}R) \leq$$
$$I(\Pi_i; X|Y\Pi_1\Pi_2\ldots\Pi_{i-1}R'R_B) + \log(I(\Pi_i; X|Y\Pi_1\Pi_2\ldots\Pi_{i-1}R'R_B)) + O(1)$$

Also by Fact 1.2.15, $I(\Pi_i'; X|Y\Pi_1'\Pi_2'\ldots\Pi_{i-1}'R) = I(\Pi_i'; X|Y\Pi_1\Pi_2\ldots\Pi_{i-1}R)$. This is because $\Pi_1', \ldots, \Pi_{i-1}'$, Y, R determine $\Pi_1, \Pi_2, \ldots\Pi_{i-1}$ and $\Pi_1'\Pi_2'\ldots\Pi_{i-1}' \to YR\Pi_1\Pi_2\ldots\Pi_{i-1} \to \Pi_i'X$ is a Markov chain. Thus

$$\begin{aligned}
&IC_\mu(\pi') \\
&\leq \sum_{i\leq r, i \text{ odd}} I(\Pi_i; X|Y\Pi_1\Pi_2\ldots\Pi_{i-1}R'R_B) + \sum_{i\leq r, i \text{ odd}} \log(I(\Pi_i; X|Y\Pi_1\Pi_2\ldots\Pi_{i-1}R'R_B)) + \\
&\sum_{i\leq r, i \text{ even}} I(\Pi_i; Y|X\Pi_1\Pi_2\ldots\Pi_{i-1}R'R_A) + \sum_{i\leq r, i \text{ odd}} \log(I(\Pi_i; Y|X\Pi_1\Pi_2\ldots\Pi_{i-1}R'R_A)) + O(r) \\
&\leq I + r\log(I/r) + O(r)
\end{aligned}$$

The last inequality follows from concavity of log. □

Our upper bound also improves slightly the bound of Harsha et al. [HJMR07]. In their setting, Alice wants to send a message M with $I(M;X) = I$ to Bob using low communication and public-randomness is allowed. They give protocol with communication cost $I + \log(I) + \log(\log(I)) + \ldots$. We can get a bound of $I + \log(I) + O(1)$ which is tight (even in terms of public-coin information) as shown by the lower bound in next section. The savings essentially come from the surprising Lemma 4.2.3.

Corollary 4.2.5. *Suppose Alice wants to help Bob to sample from the distribution $M|X = x$ and they have access to shared randomness. Let $I(M;X) = I$. Then there exists a public-coin protocol π with expected communication $\leq I + \log(I) + O(1)$, which achieves this task.*

Proof. Note that since Bob has no input, Alice actually knows the message M' in the proof of Theorem 4.2.2 in this case. Huffman encoding of M' gives the desired protocol, since $H(M') \leq I + \log(I) + O(1)$. □

4.3 Lower Bound

Now we give an example where Theorem 4.2.2 is tight. Alice is given a uniformly random string $x \in_R \{0,1\}^n$. Let $M(x,i)$ denote a message distributed according to

$$x_1, \ldots, x_{i-1}, \bar{x_i}, b_{i+1}, \ldots, b_n$$

where b_j's are random bits $\sim B_{1/2}$ and $\bar{x_i}$ denotes the flip of bit x_i.

Given x, Alice's task, T, is to transmit a message distributed according to $M(x,I)$, where $I \in_R \{1,2,\ldots,n\}$. Note that Bob has no input in this task.

First let us bound the private-coin information complexity of this task. Given x, Alice can privately sample I and send $M \sim M(x,I)$. Then the information cost of this protocol is

$I(M;X) = H(M) - H(M|X)$. It is clear that $H(M) = n$.

$$H(M|X) = \mathbb{E}_x[H(M|X = x)]$$

Denote $M|X = x$ by $M|_x$. For strings $x, y \in \{0,1\}^n$ with $x \neq y$, let $j(x,y)$ denote the first index of disagreement between x and y i.e. index j s.t. $x_j \neq y_j$. Then

$$Pr[M|_x = y] = \frac{1}{n} \cdot \frac{1}{2^{n-j(x,y)}}$$

if $x \neq y$ and 0 if $x = y$.

$$\begin{aligned} H(M|_x) &= \sum_y Pr[M|_x = y] \log\left(\frac{1}{Pr[M|_x = y]}\right) \\ &= \sum_{j=1}^{n} 2^{n-j} \cdot \frac{1}{n} \cdot \frac{1}{2^{n-j}} \log(n \cdot 2^{n-j}) + 0 \\ &= \log(n) + \frac{1}{n} \sum_{j=1}^{n} (n-j) \\ &= n/2 + \log(n) - 1/2 \end{aligned}$$

The second equality follows from the fact that there are 2^{n-j} strings y with $j(x,y) = j$, when $j \in \{1, \ldots, n\}$. This gives

$$I(M;X) = n/2 - \log(n) + 1/2$$

The following lemma lower bounds the information complexity of a public round protocol for the task T. Note that the strategy of sampling I publicly would have an information cost $\approx n/2$.

Lemma 4.3.1. *Let* Π *be a one round public-coin protocol* (*using public randomness* R) *such*

that there is a deterministic function g such that $g(\Pi_x, R)$ is distributed according to $M(x, I)$. Then $I(\Pi; X|R) \geq n/2 - O(1)$.

Proof. Since Π is a deterministic function of X and R,

$$I(\Pi; X|R) = H(\Pi|R) - H(\Pi|X, R) = H(\Pi|R)$$

Let J be a random variable that denotes the first index of disagreement between $g(\Pi, R)$ and X (Note that J is well defined because of the distribution of M). Fix a value of $R = r$. Let $p_j = Pr[J = j|R = r]$. Note that the probability is just over random X. Let μ denote the distribution of $\Pi|R = r$ and let μ_j be the distribution of $\Pi|R = r, J = j$. Then

$$\mu = \sum_{j=1}^{n} p_j \cdot \mu_j$$

Let us analyze the distribution μ_j. Let $S_r(j)$ be the set of x's which lead to $J = j$ i.e.

$$S_r(j) = \{x \in \{0,1\}^n : j(x, g(\Pi(x, r), r)) = j\}$$

Note that $|S_r(j)| = p_j \cdot 2^n$. Fixing $\Pi = t$ and $R = r$ fixes $g(\Pi, R) = g(t, r)$. Then

$$Pr[\Pi = t|R = r, J = j] \leq \frac{|\{x \in S_r(j) : j(x, g(t, r) = j)\}|}{|S_r(j)|} \leq \frac{2^{n-j}}{p_j \cdot 2^n} = \frac{1}{p_j \cdot 2^j}$$

The first inequality is because if $R = r, J = j$ are fixed, the event $\Pi = t$ implies that $j(x, g(t, r)) = j$. The second inequality follows from the fact that there are 2^{n-j} x's with $j(x, g(t, r)) = j$.

Claim 4.3.2. $H(\mu) \geq \sum_{j=1}^{n} j \cdot p_j - O(1)$.

Given the claim, we can bound $H(\Pi|R)$ as follows :

$$\begin{aligned} H(\Pi|R) &= \mathbb{E}_{r\sim R}[H(\Pi|R=r)] \\ &\geq \mathbb{E}_{r\sim R}\sum_{j=1}^{n} j\cdot p_j - O(1) \\ &= \sum_{j=1}^{n} j\cdot\frac{1}{n} - O(1) \\ &= n/2 - O(1) \end{aligned}$$

The inequality follows from the claim. The second equality follows from the fact that $\mathbb{E}_{r\sim R}Pr[J=j|R=r] = Pr[J=j] = \frac{1}{n}$. □

Proof. (Of Claim 4.3.2) Increasing a larger probability and decreasing a smaller probability by the same amount always lowers the entropy of a distribution

$$\left(p\log\left(\frac{1}{p}\right)\right)' - \left(q\log\left(\frac{1}{q}\right)\right)' = \log\left(\frac{q}{p}\right) < 0 \text{ if } q < p$$

We are given a μ_j where the mass of every entry $\mu_j(z)$ does not exceed $2^{-j}/p_j$. Therefore, we can replace μ_j with a uniform distribution on a set $\mathcal{L}_j$ of L_j entries, where $L_j = \max(1, \lfloor p_j \cdot 2^j \rfloor)$ (given any z_1, z_2 with $0 < \mu_j(z_1), \mu_j(z_2) < 1/L_j$ we can make sure that one of them becomes 0 or that one of them becomes $1/L_j$ without increasing the entropy). Note that it is always the case that $L_j > p_j \cdot 2^{j-1}$.

Therefore, we can assume wlog that each μ_j is uniform on a set $\mathcal{L}_j$ of size L_j. Consider the process of selecting an index K according to the distribution p_j, and then $Z \sim \mu_K$. Our goal is to show that $H(Z) \geq \sum_{j=1}^{n} j\cdot p_j - O(1)$. We have

$$H(KZ) = H(K) + H(Z|K) = \sum_{j=1}^{n} p_j \log(L_j/p_j) > \sum_{j=1}^{n} p_j \log(p_j\cdot 2^{j-1}/p_j) = \sum_{j=1}^{n} j\cdot p_j - 1,$$

and $H(Z) = H(KZ) - H(K|Z)$. Therefore, it suffices to show that $H(K|Z) = O(1)$. We define a subset S of j's for which p_j is "small":

$$S := \{j : p_j < 2^{-j}\}.$$

Note that for $j \notin S$ we have $p_j \cdot 2^j \geq 1$, and therefore $L_j = \lfloor p_j \cdot 2^j \rfloor$, and $p_j \cdot 2^{j-1} < L_j \leq p_j \cdot 2^j$. Denote by χ_S the indicator random variable for the event $K \in S$. We have

$$\begin{aligned} H(K|Z) \leq H(K, \chi_S|Z) &= H(\chi_S|Z) + H(K|\chi_S Z) \\ &\leq 1 + \Pr[K \in S] H(K|Z, K \in S) + \Pr[K \notin S] H(K|Z, K \notin S) \end{aligned}$$

The second inequality is because χ_S is a boolean random variable. We bound the two terms separately. Assuming $S \neq \emptyset$, denote $p_S := \sum_{j \in S} p_j$.

$$\begin{aligned} \Pr[K \in S] H(K|Z, K \in S) &\leq \Pr[K \in S] H(K|K \in S) \\ &= p_S \cdot \sum_{j \in S} \frac{p_j}{p_S} \log \frac{p_S}{p_j} \\ &\leq \sum_{j \in S} p_j \log \frac{1}{p_j} \\ &< 1 + \sum_{j \geq 2, j \in S} p_j \log \frac{1}{p_j} \\ &\leq 1 + \sum_{j=2}^{n} 2^{-j} \log \frac{1}{2^{-j}} \\ &= O(1) \end{aligned}$$

The last inequality is because the function $x \log 1/x$ is monotone increasing on the interval $(0, 1/e)$, and we have $0 < p_j < 2^{-j} < 1/e$ for $j \in S, j \geq 2$.

Finally, we need to show $\Pr[K \notin S] H(K|Z, K \notin S) = O(1)$. We will in fact show that

$H(K|Z, K \notin S) = O(1)$. We have

$$H(K|Z, K \notin S) = \mathbb{E}_{z \sim Z|_{K \notin S}} H(K|Z = z, K \notin S) \tag{4.1}$$

Fix any value of z such that $\Pr[K \notin S|Z = z] > 0$. We can precisely describe the distribution q of $K|Z = z, K \notin S$. Denote $T_z := \{j : j \notin S, z \in \mathcal{L}_j\}$. Order the elements of T_z in increasing order, and index them: $T_z = \{j_1 < j_2 < \ldots < j_k\}$. Then the distribution q puts weight $q_r := \frac{p_{j_r}/L_{j_r}}{q}$ on j_r, where $q := \sum_{r=1}^{k} p_{j_r}/L_{j_r}$. We have for each r:

$$q_r \leq \frac{p_{j_r}/L_{j_r}}{p_{j_1}/L_{j_1}} < \frac{p_{j_r}/(p_{j_r} \cdot 2^{j_r - 1})}{p_{j_1}/(p_{j_1} \cdot 2^{j_1})} = 2^{j_1 - j_r + 1} \leq 2^{2-r}$$

The second inequality follows from $L_{j_r} > p_{j_r} \cdot 2^{j_r - 1}$ and $L_{j_1} \leq p_{j_1} \cdot 2^{j_1}$(since $j_1 \notin S$) . $q_r \leq 2^{2-r}$ implies that $H(q) = O(1)$. Therefore we have $H(K|Z = z, K \notin S) = O(1)$ for each z, and by (4.1) this implies $H(K|Z, K \notin S) = O(1)$, and completes the proof. □

Chapter 5

Small Value Parallel Repetition

The results in this chapter are based on joint work with Mark Braverman [BG15].

5.1 Introduction

Parallel repetition theorem is one of the cornerstones of complexity theory. It studies hardness amplification of 2-prover 1-round games. In a 2-prover 1-round game $\mathcal{G}$, there are 2 provers, Alice and Bob, and a verifier. The verifier samples a challenge (x, y) from a joint distribution and gives x to Alice and y to Bob. Alice and Bob answer based on x and y, $(a(x), b(y))$, respectively, and they win the game if some predicate of x, y, a, b is satisfied. The central notion of study is the value of game $\text{val}(\mathcal{G})$, which is the maximum probability of winning over all strategies of Alice and Bob. A natural question is what is the value of n independent parallel repetitions of the game, in other words, is it true that $\text{val}(\mathcal{G}^n) \leq \text{val}(\mathcal{G})^n$? The main difficulty in proving such a theorem arises from the ability of the players to correlate their answers across different coordinates. The first bound on $\text{val}(\mathcal{G}^n)$ was proven by Verbitsky [Ver94] who showed that the value must go to zero as n goes to infinity. Later, Raz [Raz98] proved exponential convergence to zero with the convergence rate depending on

the answer length of the game. Feige and Verbitsky [FV02] provided an example to show that the dependence on answer length is necessary. Raz's proof was subsequently simplified and improved by Holenstein [Hol07]. Rao [Rao08] improved Holenstein's proof for the special class of projection games. The techniques of Raz, Holenstein and Rao were information theoretic. Parallel repetition theorem is very useful for gap amplification of PCPs. Rao's theorem for projection games was useful for reducing the Unique Games Conjecture (UGC) to a weaker version.

Parallel repetition for small value: The proofs of Raz, Holenstein and Rao worked only when the value of the game is close to 1. It wasn't known if a version of parallel repetition could be true when $\mathrm{val}(\mathcal{G})$ is $o(1)$. Dinur and Steurer [DS14] recently proved such a theorem for the special case of projection games, introducing linear-algebraic techniques for parallel repetition along the way. In this chapter, we give a proof for a tight parallel repetition theorem in the general small-value case using information theoretic techniques. In the process, we also give an alternative proof for the asymptotically tight bound in the small value projection case, albeit with weaker constants than [DS14].

5.1.1 Proof overview, intuition, and discussion

We start with a somewhat informal proof outline[1]. Here we opt to gloss over some technical details to convey the main ideas of the proof. This brief exposition is followed by a brief technical overview of the innovations in this proof compared to previous attempts, aimed at those familiar with the previous line of work on parallel repetition. We hope that this exposition will help elucidate our techniques and make them reusable in other related settings.

[1]Note that while we formulate our proofs for the low-value case, as this is the case that had been open, our proof easily extends to match existing proofs for $\mathrm{val}(\mathcal{G})$ close to 1.

A high-level overview. All proofs of parallel repetition theorems, including the present one, follow the same high-level strategy: we want to prove that if the value of $\mathcal{G}^n$ is too high, then there is a "too-good-to-be-true" strategy for $\mathcal{G}$. Note that if the optimal strategy $\mathcal{S}_n$ for $\mathcal{G}^n$ were independent over the n coordinates then we would have had $\mathrm{val}(\mathcal{G}^n) = \mathrm{val}(\mathcal{G})^n$, or $\mathrm{val}(\mathcal{G}) = \mathrm{val}(\mathcal{G}^n)^{1/n}$. A more contrived equivalent way of saying this is that if $\mathcal{S}_n$ were independent over the n coordinates, Alice and Bob could have dealt with a challenge (x, y) by embedding (x, y) into a coordinate i of a challenge $((x_1, \ldots, x_n), (y_1, \ldots, y_n))$ (by jointly sampling the remaining pairs (x_{-i}, y_{-i})); having Alice and Bob calculate the strategies $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$ prescribed by $\mathcal{S}_n$, respectively; and having Alice output a_i and Bob output b_i as their response to the challenge (x, y)). Since $\mathcal{S}_n$ is a product strategy, this clearly works.

The challenge is to make this embedding work even when $\mathcal{S}_n$ is a general strategy where each a_i depends on *the entire vector* $(x_1, \ldots, x_n)$. Note that as we know from counterexamples that it can happen that $\mathrm{val}(\mathcal{G}^n) \gg \mathrm{val}(\mathcal{G})^n$, this is not a mere technicality. Still, while the naïve embedding above breaks down, the general mold of the construction is a valid one: (1) embedding (x, y) into the i-th coordinate for some i; (2) sampling some public information R conditioned on (x, y); (3) having Alice and Bob play according to $\mathcal{S}_n$ conditioned on $(R, x_i = x)$ and $(R, y_i = y)$, respectively; (4) arranging R so that we can prove that the success probability of this strategy is sufficiently high.

Some previous parallel repetition proofs use the assumption that the success probability on coordinate i given success on some other coordinates is high as their departure point, and arrive at a contradiction. By proving that many of these conditional probabilities are low, these proofs establish that the probability of winning all coordinates simultaneously is also low by using the fact that

$$\Pr[\text{win on all coords}] = \prod_{i=1}^{n} \Pr[\text{win on coord } i | \text{win on coords} < i].$$

Since our main tool is symmetrization, we instead opt to define the random variable 1_W representing whether the players win on *all* n coordinates, and condition everything on the event W that they do win. Under the assumption that $p_W = \Pr[W]$ is not too small, we hope (and eventually prove) that this conditioning does not distort individual coordinates by too much: after sampling the "big" public variable R conditioned on W, Alice and Bob will sample their respective strategies on x_i and y_i ignoring W altogether. This is achieved by a careful choice of what to include in the variable R. This choice is made to balance the parameters of the problem.

The most naïve strategy would have Alice and Bob sample (x_{-i}, y_{-i}) and then play the strategy prescribed by $\mathcal{S}_n$ as they did in the special "product strategy" case above. Unfortunately, this only leads to a success probability of p_W, which is exponentially worse than what we would hope for. We would like to somehow "zoom" on the strategies of Alice and Bob conditioned on W. In other words, they would like to sample (a_i, b_i) conditioned on (x_i, y_i), and W. The problem is that conditioned on W, a_i is very far from being independent from y_i (or, for that matter from b_i conditioned on x_i). This makes such sampling impossible. To address this issue, we will have Alice and Bob sample a public variable R such that conditioned on R and $x_i = x$, a_i is (almost) independent of 1_W and y_i. Thus to sample a_i conditioned on x_i, R and the event W, Alice can ignore W and the fact she doesn't know $y_i = y$, and just sample her strategy conditioned on x_i, R.

The remaining challenge is carefully selecting the variable R. Ignoring W for the moment, we would like a_i conditioned on x_i, R to be independent from y_i. Note that in general the distribution of the answer a_i in $\mathcal{S}_n$ depends on the distribution of all coordinates x_{-i} and not just on $x_i = x$. We could have R empty, and thus have Alice and Bob sample x_{-i} and y_{-i} on their own, but since x_j and y_j are not independent, this would lead to a wrong distribution of inputs to $\mathcal{S}_n$, and thus to a wrong distribution of outputs. Another extreme solution would be to have $R = \{x_{-i}, y_{-i}\}$ contain *all* coordinates except for the i-th one.

This would solve the dependence problem, but create a new one: conditioned on W, there could be a very high dependence ($\sim \log(1/\Pr[W])$ bits of mutual information) between R and e.g. x_i, thus making it impossible for Alice and Bob (who each only have access to either x_i or y_i but not both) to sample R. As an illustration, consider the following example. Let $M = 1/\Pr[W]$ be an integer, and imagine an n-coordinate game where Alice and Bob win if and only if $\sum_{j=1}^{n}(x_j + y_j) = 0 \mod M$. Then sampling (x_{-i}, y_{-i}) correctly conditioned on W requires the knowledge of $x_i + y_i \mod M$, something neither Alice nor Bob possesses. Our solution is similar to previous solutions, although its exact execution is inspired by the latest developments of information complexity techniques, particularly in the context of direct product theorems for communication complexity. R will contain a set x_G of x's and y_H of y's such that each coordinate $j \neq i$ is contained in $G \cup H$. Thus for each such j either x_j or y_j is publicly sampled. Conditioning on R breaks the dependence between the remaining x's and y's, which can then be sampled privately. Still, R "misses" enough coordinates that the mutual information between R and y_i conditioned on W is small, and thus R can be simultaneously jointly sampled by Alice and Bob (at least with high enough probability).

Such dependence breaking appeared in previous parallel repetition proofs, as well as in information complexity/communication complexity contexts [BYJKS04, BBCR10, BR11]. Here, however, the existence of the arbitrary random variable 1_W on which we are conditioning, creates technical difficulties that do not exist in previous context. We address those by choosing G and H to have a $\Theta(n)$ overlap — a technical innovation that, to the best of our knowledge, was only employed once before [BRWY13b], and the potential applications of which are still not fully understood.

An additional complication that we need to address is that even if the mutual information between y_i and R given x_i is "small" (and thus it should be possible for Alice to sample R without knowing y_i), and similarly for Bob, this mutual information will not be very small. In particular, the best we can hope for is something of the form $O((\log(1/\Pr[W]))/n)$,

which, in the small success probability regime, is still $\omega(1)$. In previous (high success probability) works, this mutual information was $o(1)$, and thus the statistical distance between $R|x_i$, $R|y_i$, and the variable $R|(x_i, y_i)$ Alice and Bob really want to sample is also $o(1)$ by Pinsker's inequality. In this case, an approximate sample from $R|(x_i, y_i)$ is obtained using a joint sampling technique employed by Holenstein and Rao [Hol07, Rao08] in their simplified proofs of the parallel repetition theorem. In the low-success-probability case we end up proving the following statement: if the mutual information $I(Y_i; R|X_i)$ and $I(X_i; R|Y_i)$ are $< \log(1/\delta)$, then R can be correctly jointly sampled with probability $> \text{poly}(\delta)$. Note that this probability is $o(1)$ when $\log(1/\delta) = \omega(1)$, but is still high enough for our purposes. More precisely, we sample a distribution that doesn't over-sample any value of R by more than a factor of 2; it is noteworthy that such a sampling is sufficient for our purposes. The sampling lemma we prove may find other applications in complexity theory.

With R having been sampled, Alice and Bob are able to independently sample a_i and b_i conditioned on $(R, x_i = x)$ and $(R, y_i = y)$, respectively. There is one last concern: to win the game, Alice and Bob need to sample (a_i, b_i) conditioned on R, their respective inputs, and the event W. They are only able to sample these conditioned on R and their inputs. Thus, as discussed above, our final goal is to limit the dependence between 1_W and (a_i, b_i). Here we employ a trick that has been used before, though our presentation perhaps shows it in a slightly different light. To reduce the interaction between (a_i, b_i) and 1_W conditioned on R, we "hide" (a_i, b_i) among $\sim T$ other pairs of answers to challenges in $G \cap H$. This reduces the dependence between 1_W and (a_i, b_i) to $O((\log(1/\Pr[W]))/T)$ bits of information, which becomes small as T increases. However, adding the answers to T creates and additional dependence and adds to $I(Y_i; R|X_i, W)$ and $I(X_i; R|Y_i, W)$. The additional contribution is on the order of $T(\log s)/n$, where s is the size of the answer space (and thus $O(T \log s)$ is the entropy of the publicly sampled answers). Finally, a T is chosen to balance the two constraints.

Discussion of techniques. At a technical level, we further develop the idea of symmetrizing out a dependence through a careful choice of conditioning. Similar to the situation in the study of direct sum and product questions in communication complexity, all we want is to claim that there is a coordinate that is "average" in the effect conditioning on winning has on it. The simplest tool available to us which allows us to make such claims in the information-theoretic domain is the chain rule. Unfortunately, breaking the mutual information of a family of variables using the chain rule produces a family of conditional mutual information expressions, each of which has a different conditioning. The main challenge was thus to select a distribution of conditioning terms consistent with the various chain rules needed in the proof. In particular, as was the case in the proof of the direct product theorem for randomized communication complexity [BRWY13c], we seem to need to condition on a family of overlapping variables. Understanding why this is the case, and systematizing the use of such conditioning remains an interesting challenge.

The second technical innovation is a joint sampling procedure for the high information-discrepancy regime. Informally, it allows Alice and Bob who each have a distribution μ_A, μ_B, respectively, such that $D(\mu||\mu_A), D(\mu||\mu_B) \leq k$ to jointly (approximately) sample from μ with probability $> 2^{-O(k)}$. The proof of the lemma is similar to previous low-success probability constructions in [BW12, KLL$^+$12b], but its current formulation might be of use elsewhere.

We should note that while the notation is somewhat intimidating, the new proof is completely elementary. It only uses basic probability, repeated applications of the chain rule, and some elementary calculus. In particular, it does not use more advanced tools e.g. from linear algebra or spectral graph theory. Still, it is quite possible to draw parallels between our proof and the algebraic proof of Dinur and Steurer for the projection case [DS14]. This raises the tantalizing possibility of finding deeper connections between spectral and information-theoretic tools, and exploiting tools from one to advance the other.

One challenge involving the parallel repetition theorem this chapter does not address is the gap that is present in the case of general games with value close to 1. Assuming $\mathrm{val}(\mathcal{G}) = 1 - \varepsilon$, the best upper bound on $\mathrm{val}(\mathcal{G}^n)$ is $(1 - \varepsilon^3)^{\Omega(n/\log s)}$ [Hol07], while the best counterexample [Raz11] only gives a lower bound of $(1 - \varepsilon^2)^{O(n)}$. If indeed the lower bound is the tight one (in terms of dependence on ε), it would be interesting to see whether our techniques can be used to prove it.

5.1.2 Notation

We will use capital letters, e.g. A, B, X, Y to denote random variables. If X is a random variable, we will use P_X to denote its distribution. We will frequently use expectations of mutual information, so we will have a compact notation for it. Suppose $A_1, \ldots, A_n$, $B_1, \ldots, B_n$ and $C_1, \ldots, C_n$ are random variables. Let S, G, H be random subsets of $[n]$. Then we will use the notation:

$$\mathbb{E}_{P_{S,G,H}} I(A_S; B_G | C_H) := \mathbb{E}_{s,g,h \sim P_{S,G,H}} I(A_s; B_g | C_h)$$

Here A_s denotes $(A_i)_{i \in s}$. Also we will use the notation:

$$\mathbb{E}_{P_{C,D}} D(P_{A|C} || P_{B|D}) := \mathbb{E}_{c,d \sim P_{C,D}} D(P_{A|C=c} || P_{B|D=d})$$

5.1.3 Games

Here we formally define a 2-player 1-round game. Such a game $\mathcal{G}$ consists of a verifier and two provers Alice and Bob. The verifier draws (x, y) from some distribution μ on $\mathcal{X} \times \mathcal{Y}$, and gives x to Alice and y to Bob. Alice and Bob answer $a \in \mathcal{A}$ and $b \in \mathcal{B}$ depending on x and y i.e. there exists functions $f : \mathcal{X} \to \mathcal{A}$ and $g : \mathcal{Y} \to \mathcal{B}$ s.t. $a = f(x)$ and $b = g(y)$. They win the game if some predicate of x, y, a, b is satisfied i.e. there exists a subset $V \subseteq \mathcal{X} \times \mathcal{Y} \times \mathcal{A} \times \mathcal{B}$

such that they win the game if $(x, y, a, b) \in V$. Here V and μ are part of the definition of the game $\mathcal{G}$, so that $\mathcal{G} = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}, V, \mu)$. Value of the game $\text{val}(\mathcal{G})$ is defined as the maximum probability of winning over all strategies of Alice and Bob. Formally

$$\text{val}(\mathcal{G}) = \max_{f,g} \Pr_{(x,y)\sim\mu}[(x, y, f(x), g(y)) \in V]$$

The game $\mathcal{G}^n$ is defined as follows: Alice gets $x_1, \ldots, x_n$ and Bob gets $y_1, \ldots, y_n$, where $(x_1, y_1), \ldots, (x_n, y_n)$ are distributed according to μ^n (n independent copies of μ). Alice outputs $a_1, \ldots, a_n = F(x_1, \ldots, x_n)$, where $F : \mathcal{X}^n \to \mathcal{A}^n$ and Bob outputs $b_1, \ldots, b_n = G(y_1, \ldots, y_n)$, where $G : \mathcal{Y}^n \to \mathcal{B}^n$. They win the game if for all i, $(x_i, y_i, a_i, b_i) \in V$. The value is defined similarly:

$$\text{val}(\mathcal{G}^n) = \max_{F,G} \Pr_{(x_1,y_1),\ldots,(x_n,y_n)\sim\mu^n} \left[\bigwedge_{i=1}^{n} ((x_i, y_i, F(x_1, \ldots, x_n)_i, G(y_1, \ldots, y_n)_i) \in V) \right]$$

It is not hard to see that allowing shared randomness between Alice and Bob doesn't change the value of the game. But we'll allow Alice and Bob to use shared randomness to facilitate the proofs. We'll denote the size of the answer set of the game, $|\mathcal{A}| \cdot |\mathcal{B}|$ by s.

There are two special cases of games which are interesting: unique and projection games. A game is unique if its accepting predicate has the following property: for each x, y, a, there exists a unique b s.t. $(x, y, a, b) \in V$. Also for each x, y, b, there exists a unique a s.t. $(x, y, a, b) \in V$. A game is called a projection game if for each x, y, a, there exists a unique b s.t. $(x, y, a, b) \in V$. Note that in a projection game, there might exist multiple accepting answers of Alice corresponding to an answer of Bob, once we fix the questions.

5.1.4 Previous work

Exponential decay in the value of the game was first proven by Raz [Raz98]. He proved the following theorem:

Theorem 5.1.1 ([Raz98])**.** *Let $\mathcal{G}$ be a game with* $\mathrm{val}(\mathcal{G}) = 1-\epsilon$ *and let* $\log(s)$ *be the answer size of the game. Then* $\mathrm{val}(\mathcal{G}^n) \leq (1-\epsilon^{32}/2)^{\Omega(n/\log(s))}$.

This was improved by Holenstein [Hol07] who proved the following theorem:

Theorem 5.1.2 ([Hol07])**.** *Let $\mathcal{G}$ be a game with* $\mathrm{val}(\mathcal{G}) = 1-\epsilon$ *and let* $\log(s)$ *be the answer size of the game. Then* $\mathrm{val}(\mathcal{G}^n) \leq (1-\epsilon^{3}/2)^{\Omega(n/\log(s))}$.

Holenstein also proved parallel repetition for no-signaling strategies. Later Rao [Rao08] improved the bound for projection games.

Theorem 5.1.3 ([Rao08])**.** *Let $\mathcal{G}$ be a projection game with* $\mathrm{val}(\mathcal{G}) = 1-\epsilon$. *Then* $\mathrm{val}(\mathcal{G}^n) \leq (1-\epsilon^2/2)^{\Omega(n)}$.

Recently Dinur and Steurer proved parallel repetition for projection games in the small value regime.

Theorem 5.1.4 ([DS14])**.** *Let $\mathcal{G}$ be a projection game with* $\mathrm{val}(\mathcal{G}) = \beta$. *Then* $\mathrm{val}(\mathcal{G}^n) \leq (4\beta)^{n/4}$.

There has been a substantial amount of other work on improved parallel repetition for special classes of games, e.g. for free games [BRR$^+$09], expanding games [RR12] and projection games with low threshold rank [TWZ14]. Derandomizing parallel repetition theorems is an important question and there has been some work on it e.g. [Sha13], [DM11]. Recently Moshkovitz [Mos14] has given an operation on projection games, called "fortification", which makes the value of the game to behave nicely under parallel repetition. This enables improvements in the state of the art projection PCP theorem, while bypassing some of the

difficulty with general parallel repetition. There also has been a lot of work around parallel repetition for games with entanglement [CSUU08, KV11, DSV14, JPY14, CS14].

5.2 Results

The main theorem of this chapter is the following:

Theorem 5.2.1. *Let $\mathcal{G}$ be a 2-prover 1-round game. Let s be the size of answer set of the game. If $\mathrm{val}(\mathcal{G}) = \beta$, where $1/s \leq \beta$. Then $\mathrm{val}(\mathcal{G}^n) \leq \beta^{\Omega(n \log(1/\beta)/\log(s))}$, where β is sufficiently small and n sufficiently large.*

The theorem is stated formally in theorem 5.3.12 below.

Remark 5.2.2. *We assume in the theorem that $\beta \geq 1/s$. Note that this is a very natural assumption, since if for all x, y, there exist a, b s.t. the provers win on x, y, a, b, then provers can just output random answers and they win w.p. $\geq 1/s$. Even without the assumption, a simple reduction can be used to handle the case $\beta < 1/s$. In this case, the bound of the theorem is too strong, as the best we can hope for is a bound of the form $\beta^{\Omega(n)}$. Let $\mathcal{G}_w$ be the sub-game of $\mathcal{G}$ over question pairs (x, y) for which there exists some pair of answers that wins the game. Also let p be the probability that we draw such an (x, y) from the distribution for the game, i.e. p is the probability that game is winnable. Then $\mathrm{val}(\mathcal{G}) = p \cdot \mathrm{val}(\mathcal{G}_w)$ and $\mathrm{val}(\mathcal{G}^n) = p^n \cdot \mathrm{val}(\mathcal{G}_w^n)$. Then if $\mathrm{val}(\mathcal{G}) = \beta$ and $\mathrm{val}(\mathcal{G}_w) = \alpha$, where $\beta < 1/s$. There are two cases: (1) If $\log(1/\alpha) < \log(s)/2 \leq \log(1/\beta)/2$, then $\mathrm{val}(\mathcal{G}) \leq p^n = \beta^n/\alpha^n \leq \beta^{n/2}$. (2) If $\log(1/\alpha) \geq \log(s)/2$, then we can apply theorem 5.2.1 to the game $\mathcal{G}_w$:*

$$\mathrm{val}(\mathcal{G}^n) \leq p^n \cdot \alpha^{c \cdot n \log(1/\alpha)/\log(s)} = p^n \cdot \alpha^{\Omega(n)} \leq (p\alpha)^{\Omega(n)} = \beta^{\Omega(n)}$$

Remark 5.2.3. $\mathrm{val}(\mathcal{G}^n) \leq \beta^{\Omega(n/\log(1/\beta) \cdot \log(s))}$ *is what we'll get if we apply the parallel repetition theorem of Raz [Raz98]. It is not clear how to get* $\mathrm{val}(\mathcal{G}^n) \leq \beta^{\Omega(n/\log(s))}$*, however even*

in this bound, there is no "small-value behavior", since $\beta^{\Omega(n/\log(s))} \geq 2^{-\Theta(n)}$*, if* $\beta \geq 1/s$*. However our bound has the "small-value behavior" and it says that we get strong parallel repetition up to constants, if* β *and* $1/s$ *are polynomially related.*

We also show that Feige and Verbitsky's example [FV02] with tweaking of the parameters proves tightness of theorem 5.2.1.

Theorem 5.2.4. *There is a family of games* $\mathcal{G}_k$ *parametrized by* k *with* $\mathrm{val}(\mathcal{G}_k) = \beta_k \to 0$ *s.t.* $\mathrm{val}(\mathcal{G}_k^n) \geq \beta_k^{O(n \log(1/\beta_k)/\log(s_k))}$*, where* $\log(s_k)$ *is the answer size of the game* $\mathcal{G}_k$ *with* $\frac{\log(1/\beta_k)}{\log(s_k)} \to 0$*.*

Remark 5.2.5. *Theorem 5.2.1 is clearly tight when* $\log(1/\beta) = \Theta(\log(s))$*. However we give an example where it is tight even when* $\log(1/\beta) = o(\log(s))$*.*

Remark 5.2.6. *Feige and Verbitsky's example is not tight for games with constant value (it has a slack of* $\log\log(s)$*). Our work shows that for games with sub-constant value, it is exactly tight upto constant factors.*

We also reprove Dinur and Steurer's parallel repetition theorem for projection games in the small value regime. However they get much better constants in their proof. Our proof also extends to the high value regime and it provides an alternate proof for the theorems of Holenstein and Rao.

5.3 Proof for general games

We will denote by $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ inputs to Alice and Bob respectively in the n copy game. If f, g is a strategy for the game, then we'll denote by $A_1, \ldots, A_n = f(X_1, \ldots, X_n)$ and $B_1, \ldots, B_n = g(Y_1, \ldots, Y_n)$ the answers of Alice and Bob respectively. Let W be the event that they win the game on all coordinates and let 1_W be the indicator random variable

for it.

Let S, G, H be random subsets of $[n]$ distributed as follows: Let s_h and s_g be random numbers from $\{3n/4+1, \ldots, n\}$. Let $\sigma : [n] \to [n]$ be a uniformly random permutation. Set $H = \sigma([s_h])$, $G = \sigma(\{n - s_g + 1, \ldots, n\})$. Let I be a uniformly random element of $G \cap H$. Let l be a random number from $[T]$, where $T < n/2$ is a parameter. Let S be a uniformly random subset of $G \cap H \backslash \{I\}$ of size l. Let $R_{S,G,H,I}$ denote the random variable $X_{G\backslash\{I\}} Y_{H\backslash\{I\}} A_S B_S$. We will use s, g, h, i to denote instantiations of the random variables S, G, H, I respectively.

Lemma 5.3.1. $\mathbb{E}_{P_{S,G,H,I}} I(A_I B_I; 1_W | X_I, Y_I, R_{S,G,H,I}) \leq H(1_W)/T$.

Proof. Let $|g \cap h| = m$, and let $l_1, l_2, \ldots, l_m$ be the elements of $g \cap h$. Then the distribution $P_{S,G,H,I}$ can also be described as follows: G, H be distributed as in $P_{S,G,H,I}$. Let κ be a random permutation such that $\kappa(\{l_1, \ldots, l_m\}) = \{l_1, \ldots, l_m\}$, and $t \in_R [T]$. Set $I = \kappa(l_t)$ and $S = \kappa(\{l_{t+1}, \ldots, l_{T+1}\})$. Then

$$
\begin{aligned}
&\mathbb{E}_{P_{S,G,H,I}} I(A_I B_I; 1_W | X_I Y_I R_{S,G,H,I}) \\
&= \mathbb{E}_{P_{G,H}} \mathbb{E}_\kappa \mathbb{E}_{t \in_R [T]} I(A_{\kappa(l_t)} B_{\kappa(l_t)}; 1_W | A_{\kappa(\{l_{t+1}, \ldots, l_{T+1}\})} B_{\kappa(\{l_{t+1}, \ldots, l_{T+1}\})} X_G Y_H) \\
&= \mathbb{E}_{P_{G,H}} \mathbb{E}_\kappa \frac{1}{T} \sum_{t=1}^{T} I(A_{\kappa(l_t)} B_{\kappa(l_t)}; 1_W | A_{\kappa(\{l_{t+1}, \ldots, l_{T+1}\})} B_{\kappa(\{l_{t+1}, \ldots, l_{T+1}\})} X_G Y_H) \\
&= \frac{1}{T} \mathbb{E}_{P_{G,H}} \mathbb{E}_\kappa I(A_{\kappa(\{l_1, \ldots, l_T\})} B_{\kappa(\{l_1, \ldots, l_T\})}; 1_W | A_{\kappa(l_{T+1})} B_{\kappa(l_{T+1})} X_G Y_H) \\
&\leq \frac{H(1_W)}{T}
\end{aligned}
$$

□

Remark 5.3.2. *The variable size of the set S (or the variable sizes of the sets G and H, as we will see in the next lemma) is very important for the symmetrization trick to work (it enables the chain rule via an alternate description of the distribution).*

Lemma 5.3.3. $\mathbb{E}_{P_{S,G,H,I}} I(R_{S,G,H,I}; X_I | Y_I, W) \leq \frac{4}{n} H(1_W)/\Pr[W] + \frac{2(T+1)}{n} \cdot \log(s)$.

Proof. Note that $R_{S,G,H,I}$ consists of two parts : $X_{G\setminus\{I\}} Y_{H\setminus\{I\}}$ and $A_S B_S$. We will prove

$$\mathbb{E}_{P_{S,G,H,I}} I(X_{G\setminus\{I\}} Y_{H\setminus\{I\}}; X_I | Y_I, W) \leq \frac{4}{n} H(1_W)/\Pr[W] \tag{5.1}$$

and

$$\mathbb{E}_{P_{S,G,H,I}} I(A_S B_S; X_I | X_{G\setminus\{I\}} Y_{H\setminus\{I\}} Y_I, W) \leq \frac{2(T+1)}{n} \cdot \log(s) \tag{5.2}$$

which together will prove the lemma. To prove the first statement, we first prove the following statement:

$$\mathbb{E}_{P_{G,H,I}} I(X_I; 1_W | Y_I, X_{G\setminus\{I\}} Y_{H\setminus\{I\}}) \leq 4H(1_W)/n$$

The distribution $P_{G,H,I}$ can be seen in the following way: let H be distributed as in $P_{G,H,I}$. Let κ_H be a random permutation that maps $[|H|]$ to H. Choose a random number $l \in \{1, \ldots, n/4\}$. Set $I = \kappa_H(l)$ and $G = \kappa_H(\{l, \ldots, n\})$. Then

$$\begin{aligned}
\mathbb{E}_{P_{G,H,I}} I(X_I; 1_W | Y_I, X_{G\setminus\{I\}} Y_{H\setminus\{I\}}) &= \mathbb{E}_H \mathbb{E}_{\kappa_H} \mathbb{E}_{l \in_R [n/4]} I(X_{\kappa_H(l)}; 1_W | X_{\kappa_H(\{l+1,\ldots,n\})} Y_H) \\
&= \mathbb{E}_H \mathbb{E}_{\kappa_H} \frac{4}{n} \sum_{l=1}^{n/4} I(X_{\kappa_H(l)}; 1_W | X_{\kappa_H(\{l+1,\ldots,n\})} Y_H) \\
&= \frac{4}{n} \mathbb{E}_H \mathbb{E}_{\kappa_H} I(X_{\kappa_H(\{1,\ldots,n/4\})}; 1_W | X_{\kappa_H(\{n/4+1,\ldots,n\})} Y_H) \\
&\leq 4H(1_W)/n
\end{aligned}$$

Now we relate $I(X_i; 1_W | Y_i, X_{g\setminus\{i\}} Y_{h\setminus\{i\}})$ to $I(X_{g\setminus\{i\}} Y_{h\setminus\{i\}}; X_i | Y_i, W)$. Consider the mutual

information term $I(X_i; X_{g\setminus\{i\}} Y_{h\setminus\{i\}} 1_W | Y_i)$.

$$\begin{aligned} & I(X_i; X_{g\setminus\{i\}} Y_{h\setminus\{i\}} 1_W | Y_i) \\ & = I(X_i; X_{g\setminus\{i\}} Y_{h\setminus\{i\}} | Y_i) + I(X_i; 1_W | Y_i X_{g\setminus\{i\}} Y_{h\setminus\{i\}}) \\ & = I(X_i; 1_W | Y_i X_{g\setminus\{i\}} Y_{h\setminus\{i\}}) \end{aligned} \tag{5.3}$$

Also writing it in another way, we get

$$\begin{aligned} & I(X_i; X_{g\setminus\{i\}} Y_{h\setminus\{i\}} 1_W | Y_i) \\ & = I(X_i; 1_W | Y_i) + I(X_i; X_{g\setminus\{i\}} Y_{h\setminus\{i\}} | Y_i 1_W) \\ & \geq \Pr[W] \cdot I(X_i; X_{g\setminus\{i\}} Y_{h\setminus\{i\}} | Y_i, W) \end{aligned} \tag{5.4}$$

Combining (5.3) and (5.4), we get $\mathbb{E}_{P_{S,G,H,I}} I(X_I; X_{G\setminus\{I\}} Y_{H\setminus\{I\}} | Y_I, W) \leq \frac{4}{n} H(1_W) / \Pr[W]$. To prove (5.2), notice that the distribution $P_{S,G,H,I}$ can also be described as follows: Let S, H be distributed as in $P_{S,G,H,I}$. Let $\kappa_{S,H}$ be a random permutation conditioned on $\kappa_{S,H}([|S|]) = S$ and $\kappa_{S,H}([|H|]) = H$. Choose a random number l from $\{|S|+1, \ldots, |S|+n/4\}$. Set

$I = \kappa_{S,H}(l)$ and $G = S \cup \kappa_{S,H}(\{l, \ldots, n\})$. Then

$$
\begin{aligned}
&\mathbb{E}_{P_{S,G,H,I}} I(A_S B_S; X_I | X_{G\setminus\{I\}} Y_{H\setminus\{I\}} Y_I, W) \\
&= \mathbb{E}_{S,H} \mathbb{E}_{\kappa_{S,H}} \mathbb{E}_{l \in_R \{|S|+1,\ldots,|S|+n/4\}} I(A_S B_S; X_{\kappa_{S,H}(l)} | X_{\kappa_{S,H}(\{l+1,\ldots,n\})} X_S Y_H, W) \\
&= \mathbb{E}_{S,H} \mathbb{E}_{\kappa_{S,H}} \frac{4}{n} \sum_{l=|S|+1}^{|S|+n/4} I(A_S B_S; X_{\kappa_{S,H}(l)} | X_{\kappa_{S,H}(\{l+1,\ldots,n\})} X_S Y_H, W) \\
&= \frac{4}{n} \mathbb{E}_{S,H} \mathbb{E}_{\kappa_{S,H}} I(A_S B_S; X_{\kappa_{S,H}(\{|S|+1,\ldots,|S|+n/4\})} | X_{\kappa_{S,H}(\{|S|+n/4+1,\ldots,n\})} X_S Y_H, W) \\
&\leq \frac{4}{n} \mathbb{E}_S H(A_S B_S | W) \\
&\leq \frac{4}{n} \mathbb{E}_S |S| \cdot \log(s) \\
&= \frac{2(T+1)}{n} \cdot \log(s)
\end{aligned}
$$

□

Remark 5.3.4. *Note that the similar statement*

$$\mathbb{E}_{P_{S,G,H,I}} I(R_{S,G,H,I}; Y_I | X_I, W) \leq \frac{4}{n} H(1_W) / \Pr[W] + \frac{2(T+1)}{n} \cdot \log(s)$$

is also true since the distribution of G and H is symmetric.

The next lemma considers the following situation: Alice and Bob have rough estimates of a distribution P and they want to jointly sample from it. This is very similar to the settings in [BW12] and [KLL$^+$12b].

Lemma 5.3.5. *Suppose Alice knows a distribution P_1 and Bob knows a distribution P_2, and they want to jointly sample from a distribution P (all three are distributions over $\mathcal{U}$). Also $D(P||P_1) \leq \log(1/\eta)$ and $D(P||P_2) \leq \log(1/\eta)$, where $\eta \leq \frac{1}{2}$. Then there is a sampling procedure (using shared randomness) such that*

1. *Suppose Alice outputs p_1 and Bob outputs p_2. There is an event E (which depends just on the shared randomness of the sampling procedure) with $\Pr[E] \geq \eta^{10}$, such that $\Pr[p_1 = p_2 | E] = 1$.*

2. *The distribution of $p_1|E$ is multiplicatively bounded by P i.e. $\forall u$, $\Pr[p_1 = u|E] \leq 2 \cdot P(u)$.*

Proof. Consider the sampling procedure described in Protocol 4. Let

$$\mathcal{A} = \{i \text{ s.t. } q_i < P_1(u_i)/\eta^8\}$$
$$\mathcal{B} = \{i \text{ s.t. } q_i < P_2(u_i)/\eta^8\}$$
$$\mathcal{C} = \{i \text{ s.t. } q_i < P(u_i)\}$$

Let E be the event: first index in $\mathcal{A} \cup \mathcal{B}$ lies in $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$. Let us first prove that $\Pr[E] \geq \eta^{10}$. Let (u, q) be a uniformly random element of $\mathcal{U} \times [0, 1]$. Then

$$\begin{aligned}
\Pr[E] &= \frac{\Pr[q \leq \min(P_1(u)/\eta^8, P_2(u)/\eta^8, P(u))]}{\Pr[q \leq \max(P_1(u)/\eta^8, P_2(u)/\eta^8)]} \\
&\geq \frac{\Pr[q \leq \min(P_1(u)/\eta^8, P_2(u)/\eta^8, P(u))]}{\Pr[q \leq P_1(u)/\eta^8] + \Pr[q \leq P_2(u)/\eta^8]} \\
&\geq \frac{1}{2} \cdot \eta^8 \cdot |\mathcal{U}| \cdot \Pr[q \leq \min(P_1(u)/\eta^8, P_2(u)/\eta^8, P(u))]
\end{aligned}$$

Let $\mathcal{U}' = \{u \in \mathcal{U} | P(u) \leq \min(P_1(u)/\eta^8, P_2(u)/\eta^8)\}$. Then

$$\Pr[q \leq \min(P_1(u)/\eta^8, P_2(u)/\eta^8, P(u))] \geq \frac{1}{|\mathcal{U}|} \cdot P(\mathcal{U}')$$

and hence

$$\Pr[E] \geq \frac{1}{2} \cdot \eta^8 \cdot P(\mathcal{U}')$$

Since $\Pr_{x\sim P}[P(x)/P_1(x) \geq 1/\eta^8] \leq (\log(1/\eta)+1)/(8\log(1/\eta)) \leq 1/4$ (by Lemma 1.2.22), and $\Pr_{x\sim P}[P(x)/P_2(x) \geq 1/\eta^8] \leq 1/4$, we have that $P(\mathcal{U}') \geq 1/2$. Thus $\Pr[E] \geq \eta^{10}$.

1. Using shared randomness, get many uniformly random samples from $\mathcal{U}\times[0,1]$. Denote these samples by $(u_i, q_i)_{i=1}^{\infty}$.
2. Alice outputs the first u_i s.t. $q_i < P_1(u_i)/\eta^8$ and Bob outputs the first u_j s.t. $q_j < P_2(u_j)/\eta^8$.

Protocol 4: Sampling strategy

Since a subevent of E is the event that first index in $\mathcal{A}\cup\mathcal{B}$ lies in $\mathcal{A}\cap\mathcal{B}$, $\Pr[p_1 = p_2|E] = 1$. It remains to prove $\forall u$, $\Pr[p_1 = u|E] \leq 2\cdot P(u)$.

$$\begin{aligned}\Pr[p_1 = u|E] &= \frac{\min(P_1(u)/\eta^8, P_2(u)/\eta^8, P(u))}{\sum_{u\in\mathcal{U}}\min(P_1(u)/\eta^8, P_2(u)/\eta^8, P(u))}\\ &\leq \frac{P(u)}{\sum_{u\in\mathcal{U}'}\min(P_1(u)/\eta^8, P_2(u)/\eta^8, P(u))}\\ &= \frac{P(u)}{P(\mathcal{U}')} \leq 2\cdot P(u)\end{aligned}$$

This completes the proof of the lemma. □

Lemma 5.3.6. $\mathbb{E}_{P_I} D\left(P_{X_I,Y_I|W}||P_{X_IY_I}\right) \leq \frac{\log(1/\Pr[W])}{n}$.

Proof.

$$\begin{aligned}\mathbb{E}_{P_I} D\left(P_{X_I,Y_I|W}||P_{X_IY_I}\right) &= \frac{1}{n}\sum_{i=1}^{n} D\left(P_{X_i,Y_i|W}||P_{X_iY_i}\right)\\ &\leq \frac{1}{n} D\left(P_{X_1,Y_1,\ldots X_n,Y_n|W}||P_{X_1,Y_1,\ldots X_n,Y_n}\right)\\ &\leq \frac{\log(1/\Pr[W])}{n}\end{aligned}$$

The first equality is true because P_I is uniform over $[n]$. First inequality follows from Lemma 1.2.19. The second inequality follows from the Fact 1.2.20. □

Lemma 5.3.7. *Suppose* $2^{-20} \geq \Pr[W] \geq \delta^{n\log(1/\delta)/\log(s)}$*, where* $\delta \geq 1/s^{1/4}$*, and* $n \geq \frac{4\log(s)}{\log(1/\delta)}$*. Fix the parameter* T *in the definition of* $P_{S,G,H,I}$ *to be* $\frac{n\log(1/\delta)}{2\log(s)} - 1$ *(we needed* $T < n/2$ *which is true). Then there exists a fixing of* s, g, h, i *such that:*

1. $\mathbb{E}_{x,y\sim\mu_i} D\left(P_{R_{s,g,h,i}|X_i=x,Y_i=y,W} || P_{R_{s,g,h,i}|X_i=x,W}\right) \leq 10\log(1/\delta)$.
2. $\mathbb{E}_{x,y\sim\mu_i} D\left(P_{R_{s,g,h,i}|X_i=x,Y_i=y,W} || P_{R_{s,g,h,i}|Y_i=y,W}\right) \leq 10\log(1/\delta)$.
3. $D\left(P_{X_iY_i|W} || P_{X_iY_i}\right) \leq 10\log(1/\delta)$.
4. $\mathbb{E}_{P_{R_{s,g,h,i},X_i,Y_i|W}} D\left(P_{A_iB_i|X_i,Y_i,R_{s,g,h,i},W} || P_{A_iB_i|X_i,Y_i,R_{s,g,h,i}}\right) \leq 10\log(1/\delta)$.

Here μ_i *denotes the distribution* $P_{X_i,Y_i|W}$*.*

Proof. Lemma 5.3.3 proves that

$$\mathbb{E}_{P_{S,G,H,I}} I(R_{S,G,H,I}; X_I | Y_I, W) \leq \frac{4}{n} H(1_W)/\Pr[W] + \frac{2(T+1)}{n} \cdot \log(s)$$

Similarly one can prove that

$$\mathbb{E}_{P_{S,G,H,I}} I(R_{S,G,H,I}; Y_I | X_I, W) \leq \frac{4}{n} H(1_W)/\Pr[W] + \frac{2(T+1)}{n} \cdot \log(s)$$

Since $\Pr[W] \leq 2^{-20}$, we have

$$\begin{aligned} H(1_W) &= \Pr[W]\log(1/\Pr[W]) + (1-\Pr[W])\log(1/(1-\Pr[W])) \\ &\leq \Pr[W]\log(1/\Pr[W]) + \log(1 + 2\cdot\Pr[W]) \\ &\leq \Pr[W]\log(1/\Pr[W]) + 4\cdot\Pr[W] \\ &\leq 1.2\cdot\Pr[W]\log(1/\Pr[W]) \end{aligned}$$

The first inequality follows from $\frac{1}{1-x} \leq 1 + 2x$, for all $0 \leq x \leq 1/2$. The second inequality is true since $\log(1+2x) \leq 4x$, for all $x \geq 0$. The third inequality follows from $\Pr[W] \leq 2^{-20}$.

Now we have $T = \frac{n \log(1/\delta)}{2 \log(s)} - 1$ and $\frac{2(T+1)}{n} \cdot \log(s) = \log(1/\delta)$. Also

$$\begin{aligned}\frac{4}{n} H(1_W) / \Pr[W] &\leq \frac{4 \cdot 1.2 \cdot \log(1/\Pr[W])}{n} \\ &\leq \frac{4 \cdot 1.2 \cdot \log(1/\delta)^2}{\log(s)} \\ &\leq 1.2 \log(1/\delta)\end{aligned}$$

Thus

$$\begin{aligned}&\mathbb{E}_{s,g,h,i \sim P_{S,G,H,I}} \mathbb{E}_{x,y \sim \mu_i} D\left(P_{R_{s,g,h,i}|X_i=x,Y_i=y,W} || P_{R_{s,g,h,i}|X_i=x,W}\right) = \\ &\mathbb{E}_{P_{S,G,H,I}} I(R_{S,G,H,I}; Y_I | X_I, W) \leq 2.2 \log(1/\delta)\end{aligned} \tag{5.5}$$

Similarly

$$\mathbb{E}_{s,g,h,i \sim P_{S,G,H,I}} \mathbb{E}_{x,y \sim \mu_i} D\left(P_{R_{s,g,h,i}|X_i=x,Y_i=y,W} || P_{R_{s,g,h,i}|Y_i=y,W}\right) \leq 2.2 \log(1/\delta) \tag{5.6}$$

By Lemma 5.3.6, we get that

$$\mathbb{E}_{i \sim P_I} D\left(P_{X_i Y_i|W} || P_{X_i Y_i}\right) \leq \log(1/\delta)^2 / \log(s) \leq \log(1/\delta)/4 \tag{5.7}$$

Also, by Lemma 5.3.1

$$\mathbb{E}_{P_{S,G,H,I}} I(A_I B_I; 1_W | X_I, Y_I, R_{S,G,H,I}) \leq H(1_W)/T \tag{5.8}$$

Note that

$$I(A_iB_i; 1_W | X_i, Y_i, R_{s,g,h,i}) \geq$$
$$\mathbb{E}_{P_{X_i,Y_i,R_{s,g,h,i}}} \Pr[W|X_i, Y_i, R_{s,g,h,i}] \cdot D\left(P_{A_iB_i|X_i,Y_i,R_{s,g,h,i},W} || P_{A_iB_i|X_i,Y_i,R_{s,g,h,i}}\right) \tag{5.9}$$

Combining (5.8) and (5.9), we get

$$\frac{H(1_W)}{T \cdot \Pr[W]} \geq$$
$$\mathbb{E}_{P_{S,G,H,I}} \mathbb{E}_{P_{X_I,Y_I,R_{S,G,H,I}}} \left(\frac{\Pr[W|X_I, Y_I, R_{S,G,H,I}]}{\Pr[W]}\right) \cdot D\left(P_{A_IB_I|X_I,Y_I,R_{S,G,H,I},W} || P_{A_IB_I|X_I,Y_I,R_{S,G,H,I}}\right)$$
$$= \mathbb{E}_{P_{S,G,H,I}} \mathbb{E}_{P_{X_I,Y_I,R_{S,G,H,I}|W}} D\left(P_{A_IB_I|X_I,Y_I,R_{S,G,H,I},W} || P_{A_IB_I|X_I,Y_I,R_{S,G,H,I}}\right)$$

Now

$$\begin{aligned}
\frac{H(1_W)}{T \cdot \Pr[W]} &\leq 1.2 \cdot \frac{\log(1/\Pr[W])}{T} \\
&\leq 1.2 \cdot \frac{n \log(1/\delta)^2}{\log(s)} \cdot \frac{1}{\frac{n\log(1/\delta)}{2\log(s)} - 1} \\
&= 2.4 \cdot \frac{\log(1/\delta)}{1 - \frac{2\log(s)}{n\log(1/\delta)}} \\
&\leq 4.8 \log(1/\delta) \qquad (\text{since } n \geq \frac{4\log(s)}{\log(1/\delta)})
\end{aligned}$$

This gives:

$$\mathbb{E}_{s,g,h,i \sim P_{S,G,H,I}} \mathbb{E}_{P_{R_{s,g,h,i},X_i,Y_i|W}} D\left(P_{A_iB_i|X_i,Y_i,R_{s,g,h,i},W} || P_{A_iB_i|X_i,Y_i,R_{s,g,h,i}}\right) \leq 4.8 \log(1/\delta) \tag{5.10}$$

Applying a Markov argument to (5.5), (5.6), (5.7) and (5.10) completes the proof.

□

Lemma 5.3.8. *Let i satisfy the condition in Lemma 5.3.7 i.e. $D(\mu_i||\mu) \leq 10\log(1/\delta)$, where*

μ_i is the distribution $P_{X_i,Y_i|W}$ and μ is the distribution P_{X_i,Y_i}. Also suppose $\delta^{120} \leq 1/2$. Then there exists a distribution ν_i s.t. $\nu_i \leq 2 \cdot \mu_i$ and $\mu \geq \delta^{380} \cdot \nu_i$.

Proof. Let $\mathcal{B} = \{(x,y)|\mu_i(x,y) \geq \mu(x,y)/\delta^{260}\}$. Then

$$\mu_i(\mathcal{B}) \leq (10\log(1/\delta)+1)/(260\log(1/\delta)) \leq 1/2$$

(by Lemma 1.2.22 and $\delta^{120} \leq 1/2$). Now define the distribution ν_i as follows:

$$\nu_i(x,y) = \begin{cases} 0 & : (x,y) \in \mathcal{B} \\ \frac{\mu_i(x,y)}{1-\mu_i(\mathcal{B})} & : (x,y) \notin \mathcal{B} \end{cases}$$

It is clear from the definition of ν_i that $\nu_i \leq 2 \cdot \mu_i$. Now, if $(x,y) \in \mathcal{B}$, then clearly $\mu(x,y) \geq \nu_i(x,y) = 0$. If $(x,y) \notin \mathcal{B}$, then $\nu_i(x,y) \leq 2 \cdot \mu_i(x,y) \leq 2 \cdot \frac{1}{\delta^{260}} \cdot \mu(x,y) \leq \mu(x,y)/\delta^{380}$. This completes the proof. □

The next lemma is about breaking dependencies between Alice and Bob, which will be very crucial in the proof of main theorem.

Lemma 5.3.9. *Let $\mathcal{G}$ be a 2-prover 1-round game. Suppose $(X_1,Y_1),..,(X_n,Y_n)$ are inputs for $\mathcal{G}^n$ and let f, g be a strategy for $\mathcal{G}^n$. Let $A_1,\ldots,A_n = f(X_1,\ldots,X_n)$ and $B_1,\ldots,B_n = g(Y_1,\ldots,Y_n)$. Suppose $G,H,S_a,S_b \subset [n]$ and $i \in [n]$ be such that $G \cup H = [n]\backslash\{i\}$. Then*

$$P_{A_i,B_i|X_G=\bar{x},Y_H=\bar{y},A_{S_a}=\bar{a},B_{S_b}=\bar{b},X_i=x,Y_i=y} =$$

$$P_{A_i|X_G=\bar{x},Y_H=\bar{y},A_{S_a}=\bar{a},X_i=x} \otimes P_{B_i|X_G=\bar{x},Y_H=\bar{y},B_{S_b}=\bar{b},Y_i=y}$$

if $\Pr[X_G=\bar{x},Y_H=\bar{y},A_{S_a}=\bar{a},B_{S_b}=\bar{b},X_i=x,Y_i=y] > 0$.

Proof. Note that

$$P_{A_i,B_i|X_G=\bar{x},Y_H=\bar{y},A_{S_a}=\bar{a},B_{S_b}=\bar{b},X_i=x,Y_i=y}(a,b) =$$
$$P_{A_i|X_G=\bar{x},Y_H=\bar{y},A_{S_a}=\bar{a},B_{S_b}=\bar{b},X_i=x,Y_i=y}(a) \cdot P_{B_i|X_G=\bar{x},Y_H=\bar{y},A_{S_a}=\bar{a},B_{S_b}=\bar{b},X_i=x,Y_i=y,A_i=a}(b)$$

Lets first prove

$$P_{A_i|X_G=\bar{x},Y_H=\bar{y},A_{S_a}=\bar{a},B_{S_b}=\bar{b},X_i=x,Y_i=y}(a) =$$
$$P_{A_i|X_G=\bar{x},Y_H=\bar{y},A_{S_a}=\bar{a},X_i=x}(a)$$

The other part,

$$P_{B_i|X_G=\bar{x},Y_H=\bar{y},A_{S_a}=\bar{a},B_{S_b}=\bar{b},X_i=x,Y_i=y,A_i=a}(b) =$$
$$P_{B_i|X_G=\bar{x},Y_H=\bar{y},B_{S_b}=\bar{b},Y_i=y}(b)$$

would follow similarly with the set S_a changed to $S_a \cup \{i\}$.

Let $\mathcal{X}^a$ be the set of $x'_1, \ldots, x'_n$ s.t. $f(x'_1, \ldots, x'_n) = a$, $(x'_j)_{j\in G} = \bar{x}$ and $x'_i = x_i$ i.e. set of all completions of $\bar{x}, x_i$ which evaluate to a under the strategy f. Also let Q be the distribution of $X_1, \ldots, X_n$ conditioned on $X_G = \bar{x}, Y_H = \bar{y}, A_{S_a} = \bar{a}, B_{S_b} = \bar{b}, X_i = x, Y_i = y$. This is the same as distribution of $X_1, \ldots, X_n$ conditioned on $X_G = \bar{x}, Y_H = \bar{y}, A_{S_a} = \bar{a}, X_i = x$, since $[n]\backslash(G \cup \{i\}) \subseteq H$. Denote this distribution by Q'. Then

$$\begin{aligned} P_{A_i|X_G=\bar{x},Y_H=\bar{y},A_{S_a}=\bar{a},B_{S_b}=\bar{b},X_i=x,Y_i=y}(a) &= Q(\mathcal{X}^a) \\ &= Q'(\mathcal{X}^a) \\ &= P_{A_i|X_G=\bar{x},Y_H=\bar{y},A_{S_a}=\bar{a},X_i=x}(a) \end{aligned}$$

□

Remark 5.3.10. *A weaker statement is:*

$$P_{A_i,B_i|X_G=\bar{x},Y_H=\bar{y},A_{S_a}=\bar{a},B_{S_b}=\bar{b},X_i=x,Y_i=y} =$$

$$P_{A_i|X_G=\bar{x},Y_H=\bar{y},A_{S_a}=\bar{a},B_{S_b}=\bar{b},X_i=x} \otimes P_{B_i|X_G=\bar{x},Y_H=\bar{y},A_{S_a}=\bar{a},B_{S_b}=\bar{b},Y_i=y}$$

which is all we will need for the proof of Lemma 5.3.11.

Lemma 5.3.11. *If* $2^{-20} \geq \Pr[W] \geq \delta^{n\log(1/\delta)/\log(s)}$, *where* $\delta \geq 1/s^{1/4}$, $\delta^{120} \leq 1/2$ *and* $n \geq \frac{4\log(s)}{\log(1/\delta)}$, *then there exists a strategy for winning a single game w.p.* $> \delta^{2000}$.

Proof. Consider the strategy described in Protocol 5 for a single copy of the game. We prove that if $\Pr[W] \geq \delta^{n\log(1/\delta)/\log(s)}$, then the strategy wins w.p. $\geq \delta^{1940}$. Let $Q(x,y)$ denote the probability of winning when Alice and Bob get x and y, respectively. Note that the probability of winning is $\mathbb{E}_{x,y\sim\mu}Q(x,y)$. By Lemma 5.3.8, there exists a distribution ν_i s.t. $\nu_i \leq 2\cdot\mu_i$ and $\mu \geq \delta^{380}\cdot\nu_i$. We will prove that $\mathbb{E}_{x,y\sim\nu_i}Q(x,y) \geq \delta^{1560}$, which will imply that $\mathbb{E}_{x,y\sim\mu}Q(x,y) \geq \delta^{1940}$.

Inputs : Alice gets x, Bob get y, $(x,y)\sim\mu$.

1. Let s,g,h,i be as in Lemma 5.3.7.
2. Alice knows the distribution $P_{R_{s,g,h,i}|X_i=x,W}$ and Bob knows the distribution $P_{R_{s,g,h,i}|Y_i=y,W}$. They use the sampling procedure in Lemma 5.3.5 to sample from $P_{R_{s,g,h,i}|X_i=x,Y_i=y,W}$. Suppose Alice samples r_1 and Bob samples r_2.
3. Alice outputs according to the distribution $P_{A_i|X_i=x,R_{s,g,h,i}=r_1}$ and Bob outputs according the distribution $P_{B_i|Y_i=y,R_{s,g,h,i}=r_2}$.

Protocol 5: Strategy for a single game

Lemma 5.3.7 together with $\nu_i \leq 2 \cdot \mu_i$ implies that (the lemma applies since $\delta \geq 1/s^{1/4}$):

$$\mathbb{E}_{x,y\sim\nu_i} D\left(P_{R_{s,g,h,i}|X_i=x,Y_i=y,W} || P_{R_{s,g,h,i}|X_i=x,W}\right) \leq 20\log(1/\delta)$$
$$\mathbb{E}_{x,y\sim\nu_i} D\left(P_{R_{s,g,h,i}|X_i=x,Y_i=y,W} || P_{R_{s,g,h,i}|Y_i=y,W}\right) \leq 20\log(1/\delta))$$
$$\mathbb{E}_{x,y\sim\nu_i} \mathbb{E}_{P_{R_{s,g,h,i}|X_i=x,Y_i=y|W}} D\left(P_{A_iB_i|X_i=x,Y_i=y,R_{s,g,h,i},W} || P_{A_iB_i|X_i=x,Y_i=y,R_{s,g,h,i}}\right) \leq 20\log(1/\delta))$$

Let $\mathcal{S} \subset \mathcal{X} \times \mathcal{Y}$ be the set of x, y s.t.

$$D\left(P_{R_{s,g,h,i}|X_i=x,Y_i=y,W} || P_{R_{s,g,h,i}|X_i=x,W}\right) \leq 120\log(1/\delta)$$
$$D\left(P_{R_{s,g,h,i}|X_i=x,Y_i=y,W} || P_{R_{s,g,h,i}|Y_i=y,W}\right) \leq 120\log(1/\delta)$$
$$\mathbb{E}_{P_{R_{s,g,h,i}|X_i=x,Y_i=y|W}} D\left(P_{A_iB_i|X_i=x,Y_i=y,R_{s,g,h,i},W} || P_{A_iB_i|X_i=x,Y_i=y,R_{s,g,h,i}}\right) \leq 120\log(1/\delta)$$

Note that $\nu_i(\mathcal{S}) \geq 1/2$. Fix a pair $x, y \in \mathcal{S}$. We will prove that $Q(x,y) \geq \delta^{1440}$, which will imply that $\mathbb{E}_{x,y\sim\nu_i} Q(x,y) \geq \delta^{1560}$ (since $1/2 \geq \delta^{120}$). Applying Lemma 5.3.5 with $\eta = \delta^{120}$ (note that $\eta \leq 1/2$), we get that there exists an event E with $\Pr[E] \geq \delta^{1200}$, $\Pr[r_1 = r_2|E] = 1$, and the distribution of $r_1|E$ is bounded by $2 \cdot P_{R_{s,g,h,i}|X_i=x,Y_i=y,W}$. This implies that:

$$\mathbb{E}_{r\sim r_1|E} D\left(P_{A_iB_i|X_i=x,Y_i=y,R_{s,g,h,i}=r,W} || P_{A_iB_i|X_i=x,Y_i=y,R_{s,g,h,i}=r}\right) \leq 240\log(1/\delta) \tag{5.11}$$

Let $\mathcal{G}_{x,y} = \{(a,b)|V(x,y,a,b) = 1\}$, that is the set of accepting answers when the questions are x, y. Note that $P_{A_iB_i|X_i=x,Y_i=y,R_{s,g,h,i}=r,W}(\mathcal{G}_{x,y}) = 1$. This implies (by Fact 1.2.23):

$$P_{A_iB_i|X_i=x,Y_i=y,R_{s,g,h,i}=r}(\mathcal{G}_{x,y}) \geq 2^{-D\left(P_{A_iB_i|X_i=x,Y_i=y,R_{s,g,h,i}=r,W} || P_{A_iB_i|X_i=x,Y_i=y,R_{s,g,h,i}=r}\right)}$$

which along with (5.11) and convexity of the function 2^{-x} implies that:

$$\mathbb{E}_{r\sim r_1|E} P_{A_iB_i|X_i=x,Y_i=y,R_{s,g,h,i}=r}(\mathcal{G}_{x,y}) \geq \delta^{240}$$

Let $Q_E(x,y)$ be the probability of winning conditioned on event E. A very important observation is that:

$$Q_E(x,y) = \mathbb{E}_{r\sim r_1|E} P_{A_iB_i|X_i=x,Y_i=y,R_{s,g,h,i}=r}(\mathcal{G}_{x,y})$$

This is true because $P_{A_iB_i|X_i=x,Y_i=y,R_{s,g,h,i}=r} = P_{A_i|X_i=x,R_{s,g,h,i}=r} \otimes P_{B_i|Y_i=Y,R_{s,g,h,i}=r}$ (Lemma 5.3.9, it applies since $\Pr[X_i = x, Y_i = y, R_{s,g,h,i} = r] > 0$), and $\Pr[r_1 = r_2|E] = 1$. Thus $Q_E(x,y) \geq \delta^{240}$, and hence $Q(x,y) \geq \Pr[E] \cdot Q_E(x,y) \geq \delta^{1440}$. Note that

$$P_{A_iB_i|X_i=x,Y_i=y,R_{s,g,h,i}=r} = P_{A_i|X_i=x,R_{s,g,h,i}=r} \otimes P_{B_i|Y_i=Y,R_{s,g,h,i}=r}$$

is very crucial for us, otherwise the whole proof breaks down. It is crucial to break the dependencies between Alice and Bob and all the weird conditionings were needed so that this property is true.

□

Theorem 5.3.12. *Let the probability of winning of single game be β, where $\beta \leq 1/2^{20}$ and $\beta \geq 1/s$. Then probability of winning n copies of the game $\leq \beta^{n\log(1/\beta)/(2000)^2\log(s)}$. Here $n \geq \frac{4\log(s)}{\log(1/\delta)}$.*

Proof. Suppose that $\Pr[W] \geq \beta^{n\log(1/\beta)/(2000)^2\log(s)}$. Then apply Lemma 5.3.11 with $\delta = \beta^{1/2000}$. Since $\beta \leq 1/2^{20}$, we get $\delta^{120} \leq 1/2$ and $\Pr[W] \leq \beta \leq 2^{-20}$. Also since $\beta \geq 1/s$, we have $\delta \geq 1/s^{1/4}$. Note that $\beta^{n\log(1/\beta)/(2000)^2\log(s)} = \delta^{n\log(1/\delta)/\log(s)}$. Hence there exists a strategy for winning a single game w.p. $> \delta^{2000} = \beta$, a contradiction. □

5.4 Projection games

Theorem 5.4.1. *Suppose $\mathcal{G}$ is a projection game and* $\mathrm{val}(\mathcal{G}) \leq \beta$*, for β sufficiently small. Then* $\mathrm{val}(\mathcal{G}^n) \leq \beta^{\Omega(n)}$.

We recall the definition of a projection game. A game is called a projection game if for each x, y, a, there exists a unique b s.t. $(x, y, a, b) \in V$ i.e. the provers win on the tuple (x, y, a, b). We will denote by $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ inputs to Alice and Bob respectively in the n copy game. If f, g is a strategy for the game, then we'll denote by $A_1, \ldots, A_n = f(X_1, \ldots, X_n)$ and $B_1, \ldots, B_n = g(Y_1, \ldots, Y_n)$ the answers of Alice and Bob respectively. Let W be the event that they win the game on all coordinates and let 1_W be the indicator random variable for it.

We will use a slightly different proof strategy. As before, let S, G, H be random subsets of $[n]$ distributed as follows: Let s_h and s_g be random numbers from $\{3n/4 + 1, \ldots, n\}$. Let $\sigma : [n] \to [n]$ be a uniformly random permutation. Set $H = \sigma([s_h])$, $G = \sigma(\{n - s_g + 1, \ldots, n\})$. Let I be a uniformly random element of $G \cap H$. Let l be a random number from $[T]$, where $T = n/4$. Let S be a uniformly random subset of $G \cap H \backslash \{I\}$ of size l. Let $L_{S,G,H,I}$ denote the random variable $X_{G\backslash\{I\}} Y_{H\backslash\{I\}} B_S$. **The upshot is that we can afford a larger $T(= n/4)$ here, whereas in the general games proof, we could only afford $T = \Theta(n \log(1/\beta)/\log(s))$.**

Lemma 5.4.2. $\mathbb{E}_{P_{S,G,H,I}} I(A_I; Y_I | X_I, L_{S,G,H,I}, W) \leq 4 \cdot \frac{\log(1/\Pr[W])}{n}$

Proof. As in the proof of Lemma 5.3.3, the distribution $P_{S,G,H,I}$ can also be described as follows: Let S, G be distributed as in $P_{S,G,H,I}$. Let $\kappa_{S,G}$ be a random permutation conditioned on $\kappa_{S,G}([|S|]) = S$ and $\kappa_{S,G}([|G|]) = G$. Choose a random number l from $\{|S| + 1, \ldots, |S| +$

$n/4\}$. Set $I = \kappa_{S,G}(l)$ and $H = S \cup \kappa_{S,G}(\{l, \ldots, n\})$.

$$
\begin{aligned}
&\mathbb{E}_{P_{S,G,H,I}} I(A_I; Y_I | X_I, L_{S,G,H,I}, W) \\
&= \mathbb{E}_{P_{S,G,H,I}} I(A_I; Y_I | X_G, Y_{H\setminus\{I\}}, B_S, W) \\
&\leq \mathbb{E}_{P_{S,G,H,I}} I(X_{[n]\setminus G}; Y_I | X_G, Y_{H\setminus\{I\}}, B_S, W) \\
&= \mathbb{E}_{S,G} \mathbb{E}_{\kappa_{S,G}} \mathbb{E}_{l \in_R \{|S|+1,\ldots,|S|+n/4\}} I(X_{[n]\setminus G}; Y_{\kappa_{S,G}(l)} | Y_{\kappa_{S,G}(\{l+1,\ldots,n\})}, Y_S, X_G, B_S, W) \\
&= \frac{4}{n} \cdot \mathbb{E}_{S,G} \mathbb{E}_{\kappa_{S,G}} \sum_{l=|S|+1}^{|S|+n/4} I(X_{[n]\setminus G}; Y_{\kappa_{S,G}(l)} | Y_{\kappa_{S,G}(\{l+1,\ldots,n\})}, Y_S, X_G, B_S, W) \\
&= \frac{4}{n} \cdot \mathbb{E}_{S,G} \mathbb{E}_{\kappa_{S,G}} I(X_{[n]\setminus G}; Y_{\kappa_{S,G}(\{|S|+1,\ldots,|S|+n/4\})} | Y_{\kappa_{S,G}(\{|S|+n/4+1,\ldots,n\})}, Y_S, X_G, B_S, W) \\
&\leq 4 \cdot \frac{\log(1/\Pr[W])}{n}
\end{aligned}
$$

The first inequality is true since $X_{[n]}$ determines A_i. The last inequality follows from Fact 1.2.21 and that $I(X_{[n]\setminus g}; Y_{\kappa_{s,g}(\{|s|+1,\ldots,|s|+n/4\})} | Y_{\kappa_{s,g}(\{|s|+n/4+1,\ldots,n\})}, Y_s, X_g, B_s) = 0$. This is because $|g| > 3n/4 > T + n/4 \geq |s| + n/4$, therefore $\kappa_{s,g}(\{|s|+1, \ldots, |s|+n/4\}) \subseteq g$ and hence $([n]\setminus g) \subseteq \kappa_{s,g}(\{|s| + n/4 + 1, \ldots, n\})$. Note that conditioning on B_s creates dependencies between $Y_1, \ldots, Y_n$, however conditioned on $Y_{[n]\setminus g}$, there is no dependency between $X_{[n]\setminus g}$ and other Y_j's. □

Lemma 5.4.3. $\mathbb{E}_{P_{S,G,H,I}} I(L_{S,G,H,I}; Y_I | X_I, W) \leq 8 \cdot \frac{\log(1/\Pr[W])}{n}$

Proof. $L_{S,G,H,I}$ consists of two parts: $X_{G\setminus\{I\}} Y_{H\setminus\{I\}}$ and B_S. We know from the proof of Lemma 5.3.3 that

$$
\mathbb{E}_{P_{S,G,H,I}} I(Y_I; X_{G\setminus\{I\}} Y_{H\setminus\{I\}} | X_I, W) \leq \frac{4}{n} \cdot \log(1/\Pr[W]) \tag{5.12}
$$

So we care about:

$$\begin{aligned}\mathbb{E}_{P_{S,G,H,I}} I(B_S; Y_I | X_G, Y_{H\setminus\{I\}}, W) &\leq \mathbb{E}_{P_{S,G,H,I}} I(A_S; Y_I | X_G, Y_{H\setminus I}, W) \\ &\leq \mathbb{E}_{P_{S,G,H,I}} I(X_{[n]\setminus G}; Y_I | X_G, Y_{H\setminus I}, W) \end{aligned} \tag{5.13}$$

The first inequality is extremely important and this is where **we use the projection property**. The inequality holds because conditioned on W, X_s, Y_s and A_s determine B_s. Note that we use the fact that $s \subseteq (g \cap h)\setminus\{i\}$. The second inequality is true since $X_{[n]}$ determines A_s. Now by an averaging argument similar to the proof of Lemma 5.4.2, we have that:

$$\mathbb{E}_{P_{S,G,H,I}} I(X_{[n]\setminus G}; Y_I | X_G, Y_{H\setminus I}, W) \leq 4 \cdot \frac{\log(1/\Pr[W])}{n} \tag{5.14}$$

The only difference from the proof of Lemma 5.4.2 is that we will use

$$I(X_{[n]\setminus g}; Y_{\kappa_{s,g}(\{|s|+1,\ldots,|s|+n/4\})} | Y_{\kappa_{s,g}(\{|s|+n/4+1,\ldots,n\})}, Y_s, X_g) = 0$$

instead of

$$I(X_{[n]\setminus g}; Y_{\kappa_{s,g}(\{|s|+1,\ldots,|s|+n/4\})} | Y_{\kappa_{s,g}(\{|s|+n/4+1,\ldots,n\})}, Y_s, X_g, B_s) = 0.$$

Combining equations (5.12), (5.13) and (5.14) proves the lemma. □

Lemma 5.4.4. $\mathbb{E}_{P_{S,G,H,I}} I(L_{S,G,H,I}; X_I | Y_I, W) \leq 8 \cdot \frac{\log(1/\Pr[W])}{n}$

Proof. The proof of Lemma 5.3.3 gives:

$$\mathbb{E}_{P_{S,G,H,I}} I(X_{G\setminus\{I\}} Y_{H\setminus\{I\}}; X_I | Y_I, W) \leq 4 \cdot \frac{\log(1/\Pr[W])}{n} \tag{5.15}$$

Also

$$\begin{aligned}\mathbb{E}_{P_{S,G,H,I}} I(B_S; X_I | X_{G\setminus I}, Y_H, W) &\leq \mathbb{E}_{P_{S,G,H,I}} I(Y_{[n]\setminus H}; X_I | X_{G\setminus I}, Y_H, W) \\ &\leq 4 \cdot \frac{\log(1/\Pr[W])}{n} \end{aligned} \tag{5.16}$$

The first inequality holds because $Y_{[n]}$ determines B_s. The second inequality is similar to the proof of Lemma 5.4.3. Combining equations (5.15) and (5.16) proves the lemma. □

Lemma 5.4.5. $\mathbb{E}_{P_{S,G,H,I}} I(B_I; 1_W | X_I, Y_I, L_{S,G,H,I}, A_I) \leq H(1_W)/T = \frac{4H(1_W)}{n}$

Proof. Since $I(B_i; X_{[n]\setminus g} | X_i, Y_i, L_{s,g,h,i}, A_i) = 0$, we have by Fact 1.2.14 that:

$$\begin{aligned} I(B_i; 1_W | X_i, Y_i, L_{s,g,h,i}, A_i) &\leq I(B_i; 1_W | X_i, Y_i, L_{s,g,h,i}, A_i, X_{[n]\setminus g}) \\ &\leq I(B_i; 1_W | X_i, Y_i, L_{s,g,h,i}, X_{[n]\setminus g}) \end{aligned}$$

The second inequality follows from the fact that A_i is a deterministic function of $X_{[n]}$. Also

$$X_i, Y_i, L_{s,g,h,i}, X_{[n]\setminus g} = X_{[n]}, Y_h, B_s$$

Hence

$$\mathbb{E}_{P_{S,G,H,I}} I(B_I; 1_W | X_I, Y_I, L_{S,G,H,I}, A_I) \leq \mathbb{E}_{P_{S,G,H,I}} I(B_I; 1_W | X_{[n]}, Y_H, B_S) \tag{5.17}$$

As in the proof of Lemma 5.3.1, the distribution $P_{S,G,H,I}$ can also be described as follows: G, H be distributed as in $P_{S,G,H,I}$. Let κ be a random permutation such that $\kappa(\{l_1, \ldots, l_m\}) = \{l_1, \ldots, l_m\}$, and $t \in_R [T]$. Set $I = \kappa(l_t)$ and $S = \kappa(\{l_{t+1}, \ldots, l_{T+1}\})$. Here $G \cap H =$

$\{l_1, \ldots, l_m\}$. Now

$$\begin{aligned}
\mathbb{E}_{P_{S,G,H,I}} I(B_I; 1_W | X_{[n]}, Y_H, B_S) &= \mathbb{E}_{P_{G,H}} \mathbb{E}_\kappa \mathbb{E}_{t \in_R [T]} I(B_{\kappa(l_t)}; 1_W | X_{[n]}, Y_H, B_{\kappa(\{l_{t+1}, \ldots, l_{T+1}\})}) \\
&= \frac{1}{T} \cdot \mathbb{E}_{P_{G,H}} \mathbb{E}_\kappa \sum_{t=1}^{T} I(B_{\kappa(l_t)}; 1_W | X_{[n]}, Y_H, B_{\kappa(\{l_{t+1}, \ldots, l_{T+1}\})}) \\
&= \frac{1}{T} \cdot \mathbb{E}_{P_{G,H}} \mathbb{E}_\kappa I(B_{\kappa(\{l_1, \ldots, l_T\})}; 1_W | X_{[n]}, Y_H, B_{\kappa(l_{T+1})}) \\
&\leq \frac{H(1_W)}{T} \qquad (5.18)
\end{aligned}$$

Combining (5.17) and (5.18) completes the proof of the lemma. □

Lemma 5.4.6. *Let $\mathcal{G}$ be a projection game. Suppose f, g is a strategy for $\mathcal{G}^n$ and let W be the event of winning in all coordinates. If $2^{-20} \geq \Pr[W] \geq \delta^n$, then there exists a fixing of s, g, h, i such that:*

1. $\mathbb{E}_{x,y \sim P_{X_i,Y_i|W}} D\left(P_{L_{s,g,h,i}|X_i=x,Y_i=y,W} || P_{L_{s,g,h,i}|Y_i=y,W}\right) \leq O(\log(1/\delta))$
2. $\mathbb{E}_{x,y \sim P_{X_i,Y_i|W}} D\left(P_{L_{s,g,h,i}|X_i=x,Y_i=y,W} || P_{L_{s,g,h,i}|X_i=x,W}\right) \leq O(\log(1/\delta))$
3. $D(P_{X_i,Y_i|W} || P_{X_i,Y_i}) \leq O(\log(1/\delta))$
4. $\mathbb{E}_{x,y \sim P_{X_i,Y_i|W}} \mathbb{E}_{r \sim L_{s,g,h,i}|X_i=x,Y_i=y,W} D\left(P_{A_i|X_i=x,Y_i=y,L_{s,g,h,i}=r,W} || P_{A_i|X_i=x,L_{s,g,h,i}=r,W}\right)$
   $\leq O(\log(1/\delta))$
5. $\mathbb{E}_{x,y \sim P_{X_i,Y_i|W}} \mathbb{E}_{r \sim L_{s,g,h,i}|X_i=x,Y_i=y,W}$
   $D\left(P_{A_i,B_i|X_i=x,Y_i=y,L_{s,g,h,i}=r,W} || P_{A_i|X_i=x,L_{s,g,h,i}=r,W} \otimes P_{B_i|Y_i=y,L_{s,g,h,i}=r}\right) \leq O(\log(1/\delta))$

Proof. The proof is similar to the proof of Lemma 5.3.7. The proof is a Markov bound applied to the expected versions (expectation over $P_{S,G,H,I}$) of the statements. The expected versions of *1* and *2* follow from Lemma 5.4.4 and 5.4.3 respectively, as in Lemma 5.3.7. The expected version of *3* is Lemma 5.3.6. The expected version of *4* follows from Lemma 5.4.2.

For the expected version of *5*, note that:

$$
\begin{aligned}
&\mathbb{E}_{s,g,h,i\sim P_{S,G,H,I}}\mathbb{E}_{x,y\sim P_{X_i,Y_i|W}}\mathbb{E}_{r\sim L_{s,g,h,i}|X_i=x,Y_i=y,W}\\
&D\left(P_{A_i,B_i|X_i=x,Y_i=y,L_{s,g,h,i}=r,W}||P_{A_i|X_i=x,L_{s,g,h,i}=r,W}\otimes P_{B_i|Y_i=y,L_{s,g,h,i}=r}\right)\\
&=\mathbb{E}_{s,g,h,i\sim P_{S,G,H,I}}\mathbb{E}_{x,y\sim P_{X_i,Y_i|W}}\mathbb{E}_{r\sim L_{s,g,h,i}|X_i=x,Y_i=y,W}D\left(P_{A_i|X_i=x,Y_i=y,L_{s,g,h,i}=r,W}||P_{A_i|X_i=x,L_{s,g,h,i}=r,W}\right)\\
&+\mathbb{E}_{s,g,h,i\sim P_{S,G,H,I}}\mathbb{E}_{x,y\sim P_{X_i,Y_i|W}}\mathbb{E}_{r\sim L_{s,g,h,i}|X_i=x,Y_i=y,W}\mathbb{E}_{a\sim P_{A_i|X_i=x,Y_i=y,L_{s,g,h,i}=r,W}}\\
&D\left(P_{B_i|A_i=a,X_i=x,Y_i=y,L_{s,g,h,i}=r,W}||P_{B_i|Y_i=y,L_{s,g,h,i}=r}\right)\\
&\leq O(\log(1/\delta))+O(\log(1/\delta))\\
&=O(\log(1/\delta))
\end{aligned}
$$

The first inequality is expected version of *4*. The second inequality we prove below, which will complete the proof of the lemma. We want to prove that:

$$\mathbb{E}_{s,g,h,i\sim P_{S,G,H,I}}\mathbb{E}_{P_{X_i,Y_i,L_{s,g,h,i},A_i|W}}D\left(P_{B_i|X_i,Y_i,L_{s,g,h,i},A_i,W}||P_{B_i|Y_i,L_{s,g,h,i}}\right)\leq O(\log(1/\delta))$$

which is the same as

$$\mathbb{E}_{s,g,h,i\sim P_{S,G,H,I}}\mathbb{E}_{P_{X_i,Y_i,L_{s,g,h,i},A_i|W}}D\left(P_{B_i|X_i,Y_i,L_{s,g,h,i},A_i,W}||P_{B_i|X_i,Y_i,L_{s,g,h,i},A_i}\right)\leq O(\log(1/\delta))$$

since by Lemma 5.3.9, $P_{B_i|X_i,Y_i,L_{s,g,h,i},A_i}$ is the same as $P_{B_i|Y_i,L_{s,g,h,i}}$. Now note that:

$$
\begin{aligned}
&\frac{4H(1_W)}{n\Pr[W]}\geq\mathbb{E}_{s,g,h,i\sim P_{S,G,H,I}}\frac{I(B_i;1_W|X_i,Y_i,L_{s,g,h,i},A_i)}{\Pr[W]}\geq\\
&\mathbb{E}_{s,g,h,i\sim P_{S,G,H,I}}\mathbb{E}_{P_{X_i,Y_i,L_{s,g,h,i},A_i}}\frac{\Pr[W|X_i,Y_i,L_{s,g,h,i},A_i]}{\Pr[W]}\cdot D\left(P_{B_i|X_i,Y_i,L_{s,g,h,i},A_i,W}||P_{B_i|X_i,Y_i,L_{s,g,h,i},A_i}\right)\\
&=\mathbb{E}_{s,g,h,i\sim P_{S,G,H,I}}\mathbb{E}_{P_{X_i,Y_i,L_{s,g,h,i},A_i|W}}D\left(P_{B_i|X_i,Y_i,L_{s,g,h,i},A_i,W}||P_{B_i|X_i,Y_i,L_{s,g,h,i},A_i}\right)
\end{aligned}
\tag{5.19}
$$

The first inequality is Lemma 5.4.5. The second inequality follows by writing mutual infor-

mation as an expected divergence. Now since $Pr[W] \leq 2^{-20}$, $\frac{4H(1_W)}{n \Pr[W]} \leq O\left(\frac{\log(1/\Pr[W])}{n}\right) \leq O(\log(1/\delta))$, which completes the proof. □

Lemma 5.4.7. *Let $\mathcal{G}$ be a projection game. Suppose* $\mathrm{val}(\mathcal{G}^n) \geq \delta^n$*, for δ sufficiently small, then* $\mathrm{val}(\mathcal{G}) \geq \delta^{O(1)}$.

Proof. The proof is very similar to proof of Lemma 5.3.11. We use the strategy for $\mathcal{G}^n$ to obtain a strategy for $\mathcal{G}$. Suppose $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be inputs to Alice and Bob in $\mathcal{G}^n$ and $A_1, \ldots, A_n$ and $B_1, \ldots, B_n$ be their answers. W be the event of winning on all copies. Consider the strategy defined in Protocol 6. Let $Q(x, y)$ denote the probability of winning when Alice and Bob get x and y, respectively. The probability of winning is $\mathbb{E}_{x,y\sim\mu} Q(x, y)$. Let μ_i denote the distribution $P_{X_i,Y_i|W}$. Since by Lemma 5.4.6, $D(\mu_i||\mu) \leq O(\log(1/\delta))$, we get by Lemma 5.3.8, there exists a distribution ν_i s.t. $\nu_i \leq 2 \cdot \mu_i$ and $\mu \geq \delta^{O(1)} \cdot \nu_i$. We'll prove that $\mathbb{E}_{x,y\sim\nu_i} Q(x, y) \geq \delta^{O(1)}$, which will imply that $\mathbb{E}_{x,y\sim\mu} Q(x, y) \geq \delta^{O(1)}$.

Inputs : Alice gets x, Bob get y, $(x, y) \sim \mu$.

1. Let s, g, h, i be as in Lemma 5.4.6.
2. Alice knows the distribution $P_{L_{s,g,h,i}|X_i=x,W}$ and Bob knows the distribution $P_{L_{s,g,h,i}|Y_i=y,W}$. They use the sampling procedure in Lemma 5.3.5 to sample from $P_{L_{s,g,h,i}|X_i=x,Y_i=y,W}$. Suppose Alice samples r_1 and Bob samples r_2.
3. Alice outputs according to the distribution $P_{A_i|X_i=x,L_{s,g,h,i}=r_1,W}$ and Bob outputs according the distribution $P_{B_i|Y_i=y,L_{s,g,h,i}=r_2}$.

Protocol 6: Strategy for a single game: Projection case

Lemma 5.4.6 together with $\nu_i \leq 2 \cdot \mu_i$ implies that:

$$\mathbb{E}_{x,y\sim\nu_i} D\left(P_{L_{s,g,h,i}|X_i=x,Y_i=y,W} || P_{L_{s,g,h,i}|Y_i=y,W}\right) \leq O(\log(1/\delta))$$

$$\mathbb{E}_{x,y\sim\nu_i} D\left(P_{L_{s,g,h,i}|X_i=x,Y_i=y,W} || P_{L_{s,g,h,i}|X_i=x,W}\right) \leq O(\log(1/\delta))$$

$$\mathbb{E}_{x,y\sim\nu_i} \mathbb{E}_{r\sim L_{s,g,h,i}|X_i=x,Y_i=y,W} D\left(P_{A_i,B_i|X_i=x,Y_i=y,L_{s,g,h,i}=r,W} || P_{A_i|X_i=x,L_{s,g,h,i}=r,W} \otimes P_{B_i|Y_i=y,L_{s,g,h,i}=r}\right)$$
$$\leq O(\log(1/\delta))$$

Let $\mathcal{S} \subset \mathcal{X} \times \mathcal{Y}$ be the set of x, y s.t.

$$D\left(P_{L_{s,g,h,i}|X_i=x,Y_i=y,W} || P_{L_{s,g,h,i}|Y_i=y,W}\right) \leq 6 \cdot O(\log(1/\delta)) = O(\log(1/\delta))$$
$$D\left(P_{L_{s,g,h,i}|X_i=x,Y_i=y,W} || P_{L_{s,g,h,i}|X_i=x,W}\right) \leq 6 \cdot O(\log(1/\delta)) = O(\log(1/\delta))$$
$$\mathbb{E}_{r\sim L_{s,g,h,i}|X_i=x,Y_i=y,W} D\left(P_{A_i,B_i|X_i=x,Y_i=y,L_{s,g,h,i}=r,W} || P_{A_i|X_i=x,L_{s,g,h,i}=r,W} \otimes P_{B_i|Y_i=y,L_{s,g,h,i}=r}\right)$$
$$\leq 6 \cdot O(\log(1/\delta)) = O(\log(1/\delta))$$

Then $\nu_i(\mathcal{S}) \geq 1/2$. Fix a pair $x, y \in \mathcal{S}$. We will prove that $Q(x,y) \geq \delta^{O(1)}$, which will imply that $\mathbb{E}_{x,y\sim\nu_i} Q(x,y) \geq \delta^{O(1)}$, for δ sufficiently small. Applying Lemma 5.3.5 with $\eta = \delta^{O(1)}$ (note that $\eta \leq 1/2$ for δ sufficiently small), we get that there exists an event E with $\Pr[E] \geq \delta^{O(1)}$, $\Pr[r_1 = r_2|E] = 1$, and the distribution of $r_1|E$ is bounded by $2 \cdot P_{L_{s,g,h,i}|X_i=x,Y_i=y,W}$. This implies that:

$$\mathbb{E}_{r\sim r_1|E} D\left(P_{A_i,B_i|X_i=x,Y_i=y,L_{s,g,h,i}=r,W} || P_{A_i|X_i=x,L_{s,g,h,i}=r,W} \otimes P_{B_i|Y_i=y,L_{s,g,h,i}=r}\right) \leq O(\log(1/\delta)) \tag{5.20}$$

Let $\mathcal{G}_{x,y} = \{(a,b)|V(x,y,a,b) = 1\}$, that is the set of accepting answers when the questions are x, y. Note that $P_{A_iB_i|X_i=x,Y_i=y,L_{s,g,h,i}=r,W}(\mathcal{G}_{x,y}) = 1$. This implies (by Fact 1.2.23):

$$P_{A_i|X_i=x,L_{s,g,h,i}=r,W} \otimes P_{B_i|Y_i=y,L_{s,g,h,i}=r}(\mathcal{G}_{x,y})$$
$$\geq 2^{-D\left(P_{A_i,B_i|X_i=x,Y_i=y,L_{s,g,h,i}=r,W} || P_{A_i|X_i=x,L_{s,g,h,i}=r,W} \otimes P_{B_i|Y_i=y,L_{s,g,h,i}=r}\right)}$$

which along with (5.20) and convexity of the function 2^{-x} implies that:

$$\mathbb{E}_{r\sim r_1|E} P_{A_i|X_i=x,L_{s,g,h,i}=r,W} \otimes P_{B_i|Y_i=y,L_{s,g,h,i}=r}(\mathcal{G}_{x,y}) \geq \delta^{O(1)} \tag{5.21}$$

Let $Q_E(x,y)$ be the probability of winning conditioned on event E. Then by (5.21), $Q_E(x,y) \geq$

$\delta^{O(1)}$, which implies $Q(x,y) \geq \Pr[E] \cdot Q_E(x,y) \geq \delta^{O(1)}$.

□

Proof. (Of theorem 5.4.1) Follows from Lemma 5.4.7. □

5.5 Unique games

For unique games, we can obtain a simpler proof for the following theorem:

Theorem 5.5.1. *Let $\mathcal{G}$ be a unique game. Then if* $\mathrm{val}(\mathcal{G}) = \beta$*, then for β sufficiently small,* $\mathrm{val}(\mathcal{G}^n) \leq \beta^{\Omega(n)}$.

The idea is the same as in the proof of general games, but we can afford to sample $\Omega(n)$ answers. Let S, G, H be random subsets of $[n]$ distributed as follows: Let s_h and s_g be random numbers from $\{3n/4+1, \ldots, n\}$. Let $\sigma : [n] \to [n]$ be a uniformly random permutation. Set $H = \sigma([s_h])$, $G = \sigma(\{n - s_g + 1, \ldots, n\})$. Let I be a uniformly random element of $G \cap H$. Let l be a random number from $[T]$, where $T < n/2$ is a parameter. Let S be a uniformly random subset of $G \cap H \backslash \{I\}$ of size l. Let $R_{S,G,H,I}$ denote the random variable $X_{G\backslash\{I\}} Y_{H\backslash\{I\}} A_S B_S$. Here we will choose $T = n/4$. Lemma 5.3.1 gives us:

$$\mathbb{E}_{P_{S,G,H,I}} I(A_I B_I; 1_W | X_I Y_I R_{S,G,H,I}) \leq H(1_W)/T = O(\Pr[W] \cdot \log(1/\beta))$$

The other term we want to analyze is from Lemma 5.3.3: $\mathbb{E}_{P_{S,G,H,I}} I(R_{S,G,H,I}; X_I | Y_I, W)$. Here the analysis slightly deviates from the proof of general games. **We use the following property of unique games**: conditioned on W, X_i, Y_i, A_i fixes B_i and similarly X_i, Y_i, B_i fixes A_i. This is the only place where we will use the unique game property. It affects the analysis of the following term in the proof of Lemma 5.3.3 (rest of proof remains the same).

$$
\begin{aligned}
&\mathbb{E}_{P_{S,G,H,I}} I(A_S B_S; X_I | X_{G\setminus\{I\}} Y_{H\setminus\{I\}} Y_I, W) \\
&= \mathbb{E}_{P_{S,G,H,I}} I(B_S; X_I | X_{G\setminus\{I\}} Y_{H\setminus\{I\}} Y_I, W) + \mathbb{E}_{P_{S,G,H,I}} I(A_S; X_I | X_{G\setminus\{I\}} Y_{H\setminus\{I\}} Y_I, B_S, W) \\
&= \mathbb{E}_{P_{S,G,H,I}} I(B_S; X_I | X_{G\setminus\{I\}} Y_{H\setminus\{I\}} Y_I, W) \\
&\leq \mathbb{E}_{P_{S,G,H,I}} I(Y_{[n]\setminus H}; X_I | X_{G\setminus\{I\}} Y_{H\setminus\{I\}} Y_I, W) \\
&\leq O\left(\frac{\log(1/\Pr[W])}{n}\right) \\
&\leq O(\log(1/\beta))
\end{aligned}
$$

The second inequality follows because $S \subset G\setminus\{I\}$ and $S \subset H\setminus\{I\}$, hence

$$H(A_S | X_{G\setminus\{I\}} Y_{H\setminus\{I\}} Y_I, B_S, W) = 0$$

by the observation about unique games. The rest of the steps are similar to the proofs in the projection games section. This gives us:

$$\mathbb{E}_{P_{S,G,H,I}} I(R_{S,G,H,I}; X_I | Y_I, W) \leq O(\log(1/\beta))$$

and

$$\mathbb{E}_{P_{S,G,H,I}} I(R_{S,G,H,I}; Y_I | X_I, W) \leq O(\log(1/\beta))$$

from where we can finish the proof similar to the one for general games.

5.6 Tight lower bound

Theorem 5.6.1. *There is a family of games $\mathcal{G}_k$ parametrized by k with* $\text{val}(\mathcal{G}_k) = \beta_k \to 0$ *s.t.* $\text{val}(\mathcal{G}_k^n) \geq \beta_k^{O(n \log(1/\beta_k)/\log(s_k))}$, *where* $\log(s_k)$ *is the answer size of the game* $\mathcal{G}_k$ *with* $\frac{\log(1/\beta_k)}{\log(s_k)} \to 0$.

We show that different parameters in Feige and Verbitsky's counterexample [FV02] give a tight lower matching theorem 5.2.1. We describe our example below (based on [FV02], we just tweak the parameters):

- There is a parameter k and another parameter $r = k^{1/3}$.
- There is a bipartite graph G where each side has k^r vertices, the properties needed of this bipartite graph will be described later.
- Alice and Bob get uniformly distributed $(x, y) \in_R [k] \times [k]$.
- Alice needs to output $(s_a, l_a) \in [k]^r \times [r]$ and Bob needs to output $(s_b, l_b) \in [k]^r \times [r]$. They win the game if $l_a = l_b$, $s_a(l_a) = x$, $s_b(l_b) = y$ and there is an edge between s_a and s_b in G. The answer length of the game $\log(s) = 2(r \log(k) + \log(r)) = \Theta(k^{1/3} \log(k))$. Lets call this game $\mathcal{G}_k$.
- The properties we need from the graph G are the following: (1) It has at least $k^{2r}/2k^{1/5}$ edges. (2) Every k by k vertex induced subgraph of G has at most $k^2/k^{1/10}$ edges.

We'll prove the existence of such a graph G later. First lets use it to obtain a tight lower bound.

Lemma 5.6.2. $\text{val}(\mathcal{G}_k^n) \geq \left(\frac{1}{2k^{1/5}}\right)^{n/r}$

Proof. Divide the n copies into chunk of size r each. We'll give a strategy which is independent over different chunks and wins w.p. $\geq \frac{1}{2k^{1/5}}$ in each chunk and this will prove the

lemma. Suppose in a chunk Alice gets $\bar{x} = x_1, \ldots, x_r$ and Bob gets $\bar{y} = y_1, \ldots, y_r$. Then Alice outputs $(\bar{x}, 1), \ldots, (\bar{x}, r)$ and Bob outputs $(\bar{y}, 1), \ldots, (\bar{y}, r)$. The players win the all the copies in the chunk if there is an edge between $\bar{x}$ and $\bar{y}$ in G which happens w.p. $\geq \frac{1}{2k^{1/5}}$, since this is the fraction of edges in the graph G. □

Lemma 5.6.3. $\text{val}(\mathcal{G}_k) \leq 1/k^{1/20}$

Proof. Fix a strategy f, g for $\mathcal{G}_k$. We define a k by k bipartite graph G'. There is an edge between x and y if the players win under the strategy f, g on inputs x and y. Note that $\text{val}(\mathcal{G}_k) = \frac{\#\text{ of edges in } G'}{k^2}$. Suppose $f(x) = (s_a, l_a)$ and $g(y) = (s_b, l_b)$. There is an edge between x and y iff $l_a = l_b$, $s_a(l_a) = x$, $s_b(l_b) = y$ and there is an edge between s_a and s_b in G. Now look at a connected component of G' and the answer (s, l) corresponding to a vertex v in the connected component. l should be the same for all vertices in the component, and also it should hold that $s(l) = v$ for all vertices v. Because of this the answer strings corresponding to vertices on Alice's side in the component are all distinct and similarly for Bob's side. Also # of edges in the component $\leq k^2/k^{1/10}$ because of the property of G. Thus G' has the property that every connected component has at most $k^2/k^{1/10}$ edges. Now using the following claim, we get that $\text{val}(\mathcal{G}_k) = \frac{\#\text{ of edges in } G'}{k^2} \leq 1/k^{1/20}$

Claim 5.6.4. *Let G' be a k by k bipartite graph with the property that every connected component has at most $\delta \cdot k^2$ edges. Then G' has at most $\sqrt{\delta} \cdot k^2$ edges.*

Proof. (Of claim) Let $c_1, \ldots, c_t$ be the number of vertices in the components. Then $\sum_{i=1}^{t} c_i = 2k$. In each component, the number of edges $\leq \min\{c_i^2/4, \delta \cdot k^2\}$, since in a bipartite graph with c vertices, number of edges $\leq c^2/4$. Then number of edges in the graph:

$$\leq \sum_{i=1}^{t} \min\{c_i^2/4, \delta \cdot k^2\} \leq \sum_{i=1}^{t} \sqrt{(c_i^2/4) \cdot \delta \cdot k^2}$$

$$= \sqrt{\delta} \cdot \left(\sum_{i=1}^{t} c_i \right) \cdot k/2 = \sqrt{\delta} \cdot k^2$$

□

□

If $\mathrm{val}(\mathcal{G}_k) = \beta$, lemma 5.6.2 and 5.6.3 give

$$\mathrm{val}(\mathcal{G}_k^n) \geq \beta^{\Theta(n/r)} = \beta^{\Theta(n\log(k)/\log(s))} = \beta^{\Theta(n\log(1/\beta)/\log(s))}$$

Now let us prove that a graph G with required properties exists. We want it to have at least $k^{2r}/2k^{1/5}$ edges and every k by k induced subgraph to have at most $k^2/k^{1/10}$ edges. Pick a random graph with each edge included w.p. $1/k^{1/5}$. Then it has at least $k^{2r}/2k^{1/5}$ edges w.p. $1 - o(1)$. The probability that some k by k induced subgraph has at least $k^2/k^{1/10}$ edges is:

$$\begin{aligned} \leq \binom{k^r}{k}^2 \cdot \binom{k^2}{k^2/k^{1/10}} \cdot \left(\frac{1}{k^{1/5}}\right)^{k^{19/10}} &\leq \frac{k^{2rk} \cdot 2^{H(1/k^{1/10})\cdot k^2}}{2^{k^{19/10}\log(k)/5}} \\ &\leq \frac{k^{2k^{4/3}} \cdot 2^{k^{19/10}\log(k)/8}}{2^{k^{19/10}\log(k)/5}} = o(1) \end{aligned}$$

The third inequality follows from the fact that for large enough k, $H(1/k^{1/10}) \leq \frac{\log(k)}{8k^{1/10}}$. Since both the bad events occur w.p. $o(1)$, the required graph exists.

5.7 Games with value close to 1

We provide an alternate proof for the parallel repetition theorem of Holenstein [Hol07].

Theorem 5.7.1 ([Hol07])**.** *Let $\mathcal{G}$ be a game with $\mathrm{val}(\mathcal{G}) = 1 - \epsilon$ and let $\log(s)$ be the answer size of the game. Then $\mathrm{val}(\mathcal{G}^n) \leq (1-\epsilon^3)^{\Omega(n/\log(s))}$, if $n \geq \log(s)/\epsilon^3$ and $\epsilon <= 1/2$.*

The proof techniques for the small value regime readily extend to the case when $\mathrm{val}(\mathcal{G}) = 1 - \epsilon$. The only difference is that we have to replace our sampling Lemma 5.3.5 with the

correlated sampling lemma of Holenstein [Hol07]. The following variant of the lemma is proven in [Rao08].

Lemma 5.7.2. *Suppose Alice knows a distribution P_1 and Bob knows a distribution P_2 such that $||P - P_1||_1 \leq \epsilon$ and $||P - P_2||_1 \leq \epsilon$. Then there is a sampling procedure s.t.*

1. *Suppose Alice outputs p_1 and Bob outputs p_2. There exists an event E with $\Pr[E] \geq 1 - O(\epsilon)$ s.t. $\Pr[p_1 = p_2|E] = 1$.*

2. *The distribution of $p_1|E$ is P.*

Let us provide a rough sketch of our proof strategy for the high value case. Suppose W be the event of winning in all coordinates. We want to show that $\Pr[W] \leq 2^{-\Omega\left(\epsilon^3 n/\log(s)\right)}$. Assume on the contrary. As in the proof of the small value case, let S, G, H be random subsets of $[n]$ distributed as follows: Let s_h and s_g be random numbers from $\{3n/4+1, \ldots, n\}$. Let $\sigma : [n] \to [n]$ be a uniformly random permutation. Set $H = \sigma([s_h])$, $G = \sigma(\{n - s_g + 1, \ldots, n\})$. Let I be a uniformly random element of $G \cap H$. Let l be a random number from $[T]$, where $T < n/2$ is a parameter. Let S be a uniformly random subset of $G \cap H \backslash \{I\}$ of size l. Let $R_{S,G,H,I}$ denote the random variable $X_{G\backslash\{I\}} Y_{H\backslash\{I\}} A_S B_S$. We'll choose $T = \epsilon^2 n/\log(s)$ here.

Recall that the proof of Lemma 5.3.7 gives us:

$$\begin{aligned}
\mathbb{E}_{P_{S,G,H,I}} \mathbb{E}_{P_{X_I,Y_I,R_{S,G,H,I}|W}} D\left(P_{A_I B_I|X_I,Y_I,R_{S,G,H,I},W} || P_{A_I B_I|X_I,Y_I,R_{S,G,H,I}}\right) &\leq \frac{H(1_W)}{T \cdot \Pr[W]} \\
&\leq O(\epsilon) + \frac{1 - \Pr[W]}{T \cdot \Pr[W]} \cdot \log\left(\frac{1}{1 - \Pr[W]}\right) \\
&\leq O(\epsilon) + O(1/T) \\
&\leq O(\epsilon)
\end{aligned}$$

The last inequality is true for $n \geq \log(s)/\epsilon^3$. Similarly following other steps of Lemma 5.3.7,

we will get the following analogue to it: there exists a fixing of s, g, h, i s.t.

$$\mathbb{E}_{x,y\sim\mu_i} D\left(P_{R_{s,g,h,i}|X_i=x,Y_i=y,W}||P_{R_{s,g,h,i}|X_i=x,W}\right) \leq O(\epsilon^2) \tag{5.22}$$

$$\mathbb{E}_{x,y\sim\mu_i} D\left(P_{R_{s,g,h,i}|X_i=x,Y_i=y,W}||P_{R_{s,g,h,i}|Y_i=y,W}\right) \leq O(\epsilon^2) \tag{5.23}$$

$$D\left(P_{X_iY_i|W}||P_{X_iY_i}\right) \leq O(\epsilon^2) \tag{5.24}$$

$$\mathbb{E}_{P_{R_{s,g,h,i}},X_i,Y_i|W} D\left(P_{A_iB_i|X_i,Y_i,R_{s,g,h,i},W}||P_{A_iB_i|X_i,Y_i,R_{s,g,h,i}}\right) \leq O(\epsilon) \tag{5.25}$$

Here μ_i denotes the distribution $P_{X_i,Y_i|W}$. Then consider the strategy described in Protocol 7 for a single copy. We will prove that it wins w.p. $1 - O(\epsilon)$ w.r.t. the distribution μ, which will lead to a contradiction (after scaling ϵ appropriately).

Inputs : Alice gets x, Bob get y, $(x, y) \sim \mu$.

1. Let s, g, h, i be as described above.
2. Alice knows the distribution $P_{R_{s,g,h,i}|X_i=x,W}$ and Bob knows the distribution $P_{R_{s,g,h,i}|Y_i=y,W}$. They use the sampling procedure in Lemma 5.7.2 to sample from $P_{R_{s,g,h,i}|X_i=x,Y_i=y,W}$. Suppose Alice samples r_1 and Bob samples r_2.
3. Alice outputs according to the distribution $P_{A_i|X_i=x,R_{s,g,h,i}=r_1}$ and Bob outputs according the distribution $P_{B_i|Y_i=y,R_{s,g,h,i}=r_2}$.

Protocol 7: Strategy for a single game: high value case

By equation (5.24) and Pinsker's inequality, we have that: $||\mu_i - \mu||_1 \leq O(\epsilon)$. Thus it is enough to say that the strategy in Protocol 7 wins w.p. $1 - O(\epsilon)$ w.r.t. μ_i. Suppose

$$p_{x,y} := ||P_{R_{s,g,h,i}|X_i=x,Y_i=y,W} - P_{R_{s,g,h,i}|X_i=x,W}||_1$$

$$l_{x,y} := ||P_{R_{s,g,h,i}|X_i=x,Y_i=y,W} - P_{R_{s,g,h,i}|Y_i=y,W}||_1$$

$$D_{x,y} := \mathbb{E}_{P_{R_{s,g,h,i}|X_i=x,Y_i=y,W}} D\left(P_{A_iB_i|X_i=x,Y_i=y,R_{s,g,h,i},W}||P_{A_iB_i|X_i=x,Y_i=y,R_{s,g,h,i}}\right)$$

By equation (5.22), Pinsker's inequality and convexity of the function $f(z) = z^2$, we get $E_{x,y\sim\mu_i} p_{x,y} \leq O(\epsilon)$. Similarly, $E_{x,y\sim\mu_i} l_{x,y} \leq O(\epsilon)$. Also equation (5.25) gives us that $E_{x,y\sim\mu_i} D_{x,y} \leq O(\epsilon)$. Now fix a particular x, y and look at the probability of winning $Q(x, y)$. The claim is that

$$Q(x, y) \geq 1 - O(l_{x,y} + p_{x,y} + D_{x,y}) \tag{5.26}$$

This is enough to prove that $\mathbb{E}_{x,y\sim\mu_i} Q(x, y) \geq 1 - O(\epsilon)$, which is what we need. So let us prove (5.26). By Lemma 5.7.2, there exists an event $E_{x,y}$ s.t. $\Pr[r_1 = r_2 | E_{x,y}] = 1$, $r_1 | E \sim P_{R_{s,g,h,i}|X_i=x,Y_i=y,W}$ and $\Pr[E_{x,y}] \geq 1 - O(l_{x,y} + p_{x,y})$. Let $Q_E(x, y)$ be the probability of winning conditioned on $E_{x,y}$. By Fact 1.2.23 and convexity of the function $f(z) = 2^{-z}$, we have that:

$$Q_E(x, y) \geq 2^{-D_{x,y}} \geq 1 - O(D_{x,y})$$

Then

$$Q(x, y) \geq \Pr[E_{x,y}] \cdot Q_E(x, y) \geq (1 - O(l_{x,y} + p_{x,y})) \cdot (1 - O(D_{x,y})) \geq 1 - O(l_{x,y} + p_{x,y} + D_{x,y})$$

This completes the proof sketch.

Remark 5.7.3. *The proofs for unique and projection games for the small value case extend similarly to the high value case.*

Remark 5.7.4. *A remarkable feature of our proof for the high value case (a property that seems essential in the small value regime) is that we don't need the players to sample $P_{R_{s,g,h,i}|X_i=x,Y_i=y,W}$ conditioned on an event E of probability $1 - O(\epsilon)$. It would have sufficed for our purposes to samples from a distribution which is multiplicatively bounded by $P_{R_{s,g,h,i}|X_i=x,Y_i=y,W}$ (say by a factor of 2) conditioned on E. However we don't know yet how*

*to exploit it and it would be interesting if this can lead to improvements for parallel repetition for general and free games in the value-close-to-*1 *regime. Note that an improved proof has to work around the tightness of the bound for unique and projection games implied by Raz's counterexample [Raz08].*

Chapter 6

Bounded-round Quantum Communication Complexity Lower Bounds for Disjointness

The results in this chapter are based on joint work with Mark Braverman, Young Kun Ko, Jieming Mao and Dave Touchette [BGK+15].

6.1 Introduction

We prove near-optimal bounds on the bounded-round quantum communication complexity of disjointness. Quantum communication complexity, introduced by Yao [Yao93], studies the amount of quantum communication that two parties, Alice and Bob, need to exchange in order to compute a function (usually boolean) of their private inputs. It is the natural quantum extension of classical communication complexity [Yao79]. While the inputs are classical and the end result is classical, the players are allowed to use quantum resources while communicating. The motivation for the introduction of quantum communication was

to study questions in quantum computation. For example, in [Yao93], Yao used it to prove that the majority function does not have any linear size quantum formulas.

For disjointness with input size n, Grover's search [Gro96, BBHT98] can be used to obtain a quantum communication protocol (with probability of error $1/3$) with communication cost $O(\sqrt{n}\log n)$ [BCW98]. The bound was later improved to $O(\sqrt{n})$ in [AA03]. The protocols attaining this upper bound are very interactive and require $\Theta(\sqrt{n})$ rounds of interaction. The $O(\sqrt{n})$ upper bound on the quantum communication complexity of disjointness has been shown to be tight in [Raz02].

If we restrict the players to allow only r rounds of interaction, then it is not hard to use the $O(\sqrt{n})$ protocol discussed above as a black-box to obtain an $O(n/r)$ communication protocol for $n \geq r^2$. The best known lower bound was $\Omega(n/r^2)$ [JRS03]. We prove a lower bound of $\tilde{\Omega}(n/r)$, which is optimal up to logarithmic factors:

Theorem A. (Theorem 6.6.3, rephrased) *The r-round quantum communication complexity of $DISJ_n$ is $\Omega\left(\frac{n}{r\log^8(r)}\right)$.*

The analogous result for query complexity of quantum search, an $\Omega(n/r)$ lower bound for the number of queries when r sets of nonadaptive queries are allowed, was known before [Zal99]. Our lower bound does not give a new proof of the $\Omega(\sqrt{n})$ bound on the quantum communication complexity of disjointness [Raz02] since our proof uses that lower bound (in fact we use something much stronger, a strengthening of the strong direct product theorem for disjointness [KSDW04] due to [She12]).

There is a rich history of papers studying lower bounds on bounded-round communication complexity, for example for the pointer jumping problem [NW93b, PRV01, Kla98, KNTSZ01], for sparse set disjointness [ST13], for equality [BCK14] and several other examples. Most of these lower bounds are proven via a round elimination strategy: show that an r-round protocol can be converted into an $(r-1)$-round protocol without too much increase in communication cost and error; arrive at contradition by obtaining a too-good-to-be-true

1-round or 0-round protocol. Even the result of [JRS03] can be viewed as round elimination on quantum information complexity of the 2-bit AND. Despite substantial effort, obtaining the optimal $\Omega(1/r)$ lower bound on the r-round quantum information complexity of AND via round elimination has remained elusive. We prove:

Theorem B. (Corollary 6.6.2, rephrased) *The r-round quantum information complexity of AND with prior $1/3, 1/3, 1/3, 0$ is $\Omega\left(\frac{1}{r\log^8(r)}\right)$.*

As discussed below, we obtain this result by using existing lower bounds for the communication complexity of quantum disjointness. A direct proof of a quantum information complexity lower bound for the 2-bit AND remains an intriguing open problem. In light of the fact that disjointness has a sub-linear quantum communication complexity, it is not surprising that the quantum information complexity of AND vanishes with the number of rounds. This phenomenon is closely related to the Elitzur-Vaidman bomb tester [EV93, KWHZ95], which gives a sequence of quantum measurements that allows one to test whether a bomb is loaded without detonating it. The loss of the protocol (i.e. the probability that the bomb will explode — which loosely corresponds to the amount of information revealed about the bomb) behaves like $1/r$, where r is the number of measurements performed.

Our proof relies on the notion of quantum information complexity, defined recently in [Tou15], where it is used to prove a direct sum theorem for constant round quantum communication. It is harder to manipulate quantum information than in the classical case, and tools that are standard in the classical setting are yet to be developed for the quantum case. However, it could still be useful in proving partial direct sum and direct product theorems, which we know in the classical world [BBCR10], [BRWY13b]. Moreover, a model similar to that of quantum communication complexity is connected to proving SDP extension complexity lower bounds [JSWZ13]. Although the recent breakthrough for SDP lower bounds [LRS15] does not follow this direction, it is likely that a quantum information complexity viewpoint will provide further insights as information complexity has provided in the classical

case (LP extension complexity) [BM13, BP13]. Further development of tools for quantum communication and information complexity is likely to further the SDP extension complexity program.

We also prove that for all boolean functions, prior-free quantum information complexity is lower bounded by the generalized discrepancy method:

Theorem C. (Theorem 6.4.7, rephrased) *For any boolean function f and a sufficiently small constant error $\eta > 0$, the prior-free quantum information complexity of f with error η is lower bounded by the generalized discrepancy bound for f.*

Previously no lower bounds were known on the quantum information complexity of general boolean functions. Our proof relies on the strong direct product theorem for quantum communication complexity in terms of the generalized discrepancy method [She12]. Note that in the classical setting such a result can be proven directly using zero-communication protocols [KLL$^+$12b]. It remains to be seen whether such a direct proof can be obtained in the quantum setting.

As a corollary we also get that the quantum communication complexity of any boolean function is at most exponential in the prior-free quantum information complexity.

Theorem D. (Corollary 6.4.8, rephrased) *For any boolean function f, quantum communication complexity of f with error $1/3$ is at most $2^{O(QIC(f,1/3)+1)}$, where $QIC(f, 1/3)$ is the prior-free quantum information complexity of f with error $1/3$.*

Note that the classical analogue of this is proven via a compression argument [Bra12], but we prove this via an indirect argument. It would be interesting to prove this directly via a quantum compression argument.

Acknowledgments

We would like to thank Andris Ambainis, Rahul Jain, Ashwin Nayak, Jaikumar Radhakrishnan, Iordanis Kerenidis and Mathieu Lauriere for helpful discussions.

6.2 Proof overview and discussion

High-level strategy. At a high-level, the proof builds on the connection between quantum information complexity and quantum communication complexity of the disjointness function $DISJ_m$ with various values of m. There are two parts to the proof:

1. Suppose there is a r-round quantum protocol for disjointness of input size $n \geq r^2$ with communication cost $\frac{n}{r \cdot \text{polylog}(r)}$. Then there exists a protocol for disjointness of input size r^2 with quantum information cost $\leq o(r)$.

2. Lower bound on quantum information complexity of disjointness: we prove that the (prior-free) quantum information complexity of any boolean function is lower bounded by the generalized discrepancy method, which by results in [She07] implies that quantum information complexity of disjointness with input size r^2 is $\Omega(r)$.

Note that these two steps imply a lower bound on the bounded round quantum communication complexity of disjointness. Also the above statements are about computation with some constant error (say 1/3).

Both directions are proven via a connection between the information complexity of a problem and its communication complexity. In one direction, a protocol for a large sized disjointness can be converted into a low-information protocol for a smaller size disjointness. Using the converse direction of the connection, a low-information protocol for $DISJ_{r^2}$ leads to a protocol for many copies of the problem that violate known direct product results. The former connection has been at the heart of many classical lower bounds involving information

complexity [BYJKS04, BGPW13a]. The latter connection (deriving information complexity lower bound from known communication lower bound on an "amortized" version of the problem) has been previously explored in the classical setting by [BGPW13c].

Let us start by giving a high level overview of the first step. If there is a r-round quantum protocol for disjointness of input size n with communication cost $\frac{n}{r \cdot \text{polylog}(r)}$ and $1/3$ probability of error, then by a direct sum argument in [Tou15], there exists a r-round quantum protocol π for AND with $1/3$ probability of error (for a worst case input) and quantum information cost $\leq \frac{1}{r \cdot \text{polylog}(r)}$ w.r.t **any distribution $\boldsymbol{\mu}$ s.t. $\boldsymbol{\mu(1,1) = 0}$**. Now we want to use π to obtain a low information protocol for disjointness of size r^2. One can imagine if we run π on each coordinate of the disjointness instance, we get an r-round protocol τ of information cost $\leq \frac{r}{\text{polylog}(r)}$ and also it solves disjointness with small error (assuming we first amplify the error of π to $1/r^3$ losing a log factor in information cost). However, the issue is that information cost of τ is low only w.r.t. **distributions $\boldsymbol{\nu}$ supported on disjoint pairs of sets**. The information cost of τ may increase dramatically when it is run on a pair of sets with many intersections. To deal with this we use a trick used in [BGPW13a].

Note that if there are too many intersections in a disjointness instance, then the players can just subsample some of the coordinates and check for an intersection in those coordinates. Hence we can assume wlog that the intersection size in a typical input distributed according to ν is small. This means that if we look at a typical coordinate i, the marginal distribution ν_i has small mass on $(1,1)$. And in this case, we can run π on each coordinate. The only thing left to understand is: how does the information cost of π change if we place a small mass, say w, on $(1,1)$? The answer to this turns out to be $r \cdot H(w)$, where π has r-rounds. Note that this is in contrast to the classical case, where the answer would be just $H(w)$. Later we will give an example of a quantum protocol for AND whose information cost does go up by $r \cdot H(w)$. Also **this is the only place where we use the fact that the protocol we started with had only $\boldsymbol{r}$ rounds**. Such a dependence is necessary here, since an $\Omega(n/r)$

lower bound for general (non-r-round) protocols would violate the $O(\sqrt{n})$ upper bound.

For the second step, we use compression along with a strong direct product theorem for quantum communication complexity of f in terms of the generalized discrepancy lower bound $GDM_{1/5}(f)$ due to Sherstov [She12]. It says that to compute k copies of a boolean function f with success probability $2^{-\Omega(k)}$, it requires at least $k \cdot GDM_{1/5}(f)$ qubits of communication (with arbitrary amount of entanglement). Note that a strong direct product theorem for quantum communication complexity of disjointness was already known [KSDW04], but we need a stronger version for our proof which shows that even computing a large fraction of the copies is hard and Sherstov's result also holds in this case[1].

Suppose there is a protocol π for a function f with quantum information cost $\leq I$ w.r.t a distribution μ and probability of error $\leq \epsilon$, then by quantum information equals amortized communication [Tou15], we get a protocol π_k for f^k which computes at least $(1 - 2\epsilon)k$ coordinates correctly with probability ≥ 0.99 (w.r.t. μ^k) and $QCC(\pi_k) \leq k \cdot I + o(k)$. To apply Sherstov's theorem, we need such a protocol which works for worst case inputs. We show how to obtain such a worst case to average case reduction, whence applying Sherstov's result gives us the lower bound on information complexity.

Discussion and open problems

In its entirety our proof shows how from a r-round protocol for disjointness, one can obtain a protocol for k copies of disjointness of size r^2. But to achieve this reduction, we have to move to information complexity, since the number of rounds r only comes up in an information theoretic context in our proof.

Thus the reduction structure of the proof is communication→information→communication, with the latter communication problem having a known lower bound. Lower bounds for

[1]We could probably base our result off the lower bound of [KSDW04], but the reduction would be considerably more complicated.

disjointness in the classical setting [BYJKS04, BGPW13a] only do a reduction of the form communication $\rightarrow$ information, with an information complexity lower bound on the resulting problem proven directly.

Open Problem 6.2.1. *Give a direct proof of a lower bound for the information complexity of* $DISJ_{r^2}$.

One possible attack route would be along the lines of the proof for the classical case using zero-communication protocols [KLL$^+$12b]. In the past, techniques developed for two-party quantum communication, e.g. the pattern matrix method [She07], turned out to be useful for multiparty number-on-forehead communication [CA08, She14]. It could be that techniques developed for quantum information also result in similar progress.

Another natural question is whether the lower bound on the information complexity of AND can be proved using a direct argument:

Open Problem 6.2.2. *Give a direct proof of Theorem B.*

Even though efforts since [JRS03] to-date have been unsuccessful, it still could be possible to directly obtain Theorem B via round elimination or other techniques and that would be really interesting, since it would also yield a new proof of the lower bound for quantum communication complexity of disjointness [Raz02, She07]. The recent breakthrough results in lower bounding conditional quantum mutual information [FR14, BHOS14, BT15] should be relevant.

Remark 6.2.3. *Our proofs can be adapted to show that the (unbounded round) zero-error quantum information complexity of AND w.r.t the prior* $(1-\epsilon)/3, (1-\epsilon)/3, (1-\epsilon)/3, \epsilon$ *is* $\tilde{\Omega}(\sqrt{\epsilon})$. *It is another intriguing question whether it is possible to have a direct proof for this. Note that this requires a global view of quantum information complexity, even though it is defined round by round. By a continuity argument this would also resolve open problem 6.2.2.*

More generally, our understanding of the relationship between quantum information and communication complexity is in its early stages of development. Questions of interactive protocol compression occupy a central position in understanding the connection between classical information and communication complexity [BBCR10, Bra12, GKR14b]. In particular, [BBCR10] shows that a protocol π with information cost I and communication cost C can be compressed into a protocol with communication cost $\tilde{O}(\sqrt{I \cdot C})$. It remains open whether this (or an analogous) fact is true in the quantum setting:

Open Problem 6.2.4. *Given a quantum protocol π over a distribution μ of inputs whose communication cost is C and whose quantum information cost is I, can π be simulated (with a small error) using a quantum protocol π' whose communication cost is $\tilde{O}(\sqrt{I \cdot C})$?*

We refer the reader to Sections 1.4, 1.5 and 1.6 for preliminaries on quantum information theory, quantum communication and quantum information complexity.

6.3 Properties of Quantum Information Complexity

In this section, we prove general results about quantum information complexity that we use to obtain the main results. These may be of independent interest.

6.3.1 Prior-free Quantum Information Complexity

We want to define a sensible notion of quantum information complexity for classical tasks. Like in the classical setting [Bra12], there are two sensible orderings for the optimization over inputs and protocols. We provide the two corresponding definitions and then investigate the link between them. We denote by $\mathcal{D}_{XY}$ the set of all distributions μ on input space $X \times Y$.

Definition 6.3.1. The *max-distributional quantum information complexity* of a relation T

with error $\epsilon \in [0,1]$ is

$$QIC_D(T,\epsilon) = \max_{\mu \in \mathcal{D}_{XY}} QIC(T,\mu,\epsilon).$$

When restricting to r-round protocols, it is

$$QIC_D^r(T,\epsilon) = \max_{\mu \in \mathcal{D}_{XY}} QIC^r(T,\mu,\epsilon).$$

Definition 6.3.2. The *quantum information complexity* of a relation T with error $\epsilon \in [0,1]$ is

$$QIC(T,\epsilon) = \inf_{\Pi \in \mathcal{T}(T,\epsilon)} \max_{\mu \in \mathcal{D}_{XY}} QIC(\Pi,\mu).$$

When restricting to r-round protocols, it is

$$QIC^r(T,\epsilon) = \inf_{\Pi \in \mathcal{T}^r(T,\epsilon)} \max_{\mu \in \mathcal{D}_{XY}} QIC(\Pi,\mu).$$

Lemma 6.3.3 (Information lower bounds communication)**.** *For any relation T, error parameter $\epsilon \in [0,1]$, and number of rounds $r \in \mathbb{N}$, the following holds:*

$$QIC^r(T,\epsilon) \leq QCC^r(T,\epsilon),$$
$$QIC(T,\epsilon) \leq QCC(T,\epsilon).$$

Proof. Let Π be a protocol computing T correctly except with probability ϵ on all input and satisfying $QCC(\Pi) = QCC(T,\epsilon)$. We get the result by noting that $QIC(T,\epsilon) \leq \max_\mu QIC(\Pi,\mu) \leq QCC(\Pi)$. □

Clearly, $QIC_D(T,\epsilon) \leq QIC(T,\epsilon)$, and $QIC_D^r(T,\epsilon) \leq QIC^r(T,\epsilon)$. We prove that we can

almost reverse the quantifiers. The proof idea follows the lines of the proof of Theorem 3.5 in Ref. [Bra12], but special care must be taken for quantum protocols. The idea we use is to take an ϵ-net over $\mathcal{D}_{XY}$, and then take a δ-optimal protocol for each distribution in the net. To extend this result to the unbounded round quantum setting, we adapt a compactness argument from Ref. [BGPW13a], itself adapted from Ref. [Ter72]. The following results will be used.

Lemma 6.3.4 (Continuity in average error)**.** *Quantum information complexity is continuous in the error. This holds uniformly in the input. That is, for all T, r and $\epsilon, \delta > 0$, there exists $\epsilon' \in (0, \epsilon)$ such that for all $\epsilon'' \in (\epsilon', \epsilon)$ and for all μ,*

$$|QIC(T, \mu, \epsilon - \epsilon'') - QIC(T, \mu, \epsilon)| \leq \delta,$$
$$|QIC^r(T, \mu, \epsilon - \epsilon'') - QIC^r(T, \mu, \epsilon)| \leq \delta.$$

Proof. Note that we can drop the absolute values and also work at ϵ' since quantum information complexity is non-increasing in the error, i.e. $QIC(T, \mu, \epsilon) \leq QIC(T, \mu, \epsilon - \epsilon'') \leq QIC(T, \mu, \epsilon - \epsilon')$. Let $0 < p < \frac{1}{2}$ and use Corollary 1.6.9 with $\epsilon_1 = 0, \epsilon_2 = \epsilon, \epsilon' = p\epsilon$ for the current ϵ. We get

$$\begin{aligned} QIC(T, \mu, \epsilon - \epsilon') &\leq pQIC(T, \mu, 0) + (1 - p)QIC(T, \mu, \epsilon) \\ &\leq pQCC(T, 0) + QIC(T, \mu, \epsilon). \end{aligned}$$

Rearranging terms, we get

$$|QIC(T, \mu, \epsilon - \epsilon') - QIC(T, \mu, \epsilon)| \leq \frac{\epsilon'}{\epsilon} QCC(T, 0).$$

This bound is independent of μ, and goes to zero as p and ϵ' do, so the result follows. The bounded round result is proved in the same way, obtaining $QCC^r(T,0)$ in the final bound instead. □

Lemma 6.3.5 (Convexity in error)**.** *For any $p \in [0,1], T$ and $\epsilon, \epsilon_1, \epsilon_2 \in [0,1]$ satisfying $\epsilon = p\epsilon_1 + (1-p)\epsilon_2$ and for any bound $r = \max(r_1, r_2), r_1, r_2 \in \mathbb{N}$ on the number of rounds, the following holds:*

$$QIC(T,\epsilon) \leq pQIC(T,\epsilon_1) + (1-p)QIC(T,\epsilon_2),$$
$$QIC^r(T,\epsilon) \leq pQIC^{r_1}(T,\epsilon_1) + (1-p)QIC^{r_2}(T,\epsilon_2).$$

Proof. The proof is similar to the one for the analogous result with fixed input. Given $\delta > 0$, let Π^1 and Π^2 be protocols satisfying, for all μ, for $i \in \{1,2\}, \Pi^i \in \mathcal{T}(T,\epsilon_i), QIC(\Pi^i,\mu) \leq QIC(T,\epsilon_i) + \delta$, and take the corresponding protocol Π of Lemma 1.6.8. First, it holds that protocol Π successfully accomplish its task, i.e. it implements task T on all inputs with error bounded by $\epsilon = p\epsilon_1 + (1-p)\epsilon_2$. We must now verify that the quantum information cost satisfies the convexity property:

$$\begin{aligned} QIC(T,\epsilon) &\leq \max_{\mu} QIC(\Pi,\mu) \\ &= \max_{\mu} \left(pQIC(\Pi^1,\mu) + (1-p)QIC(\Pi^2,\mu) \right) \\ &\leq p \max_{\mu} QIC(\Pi^1,\mu) + (1-p) \max_{\mu} QIC(\Pi^2,\mu) \\ &\leq pQIC(T,\epsilon_1) + (1-p)QIC(T,\epsilon_2) + 2\delta. \end{aligned}$$

Keeping track of rounds, we get the bounded round result. □

Corollary 6.3.6 (Continuity in error)**.** *Quantum information complexity is continuous in*

the error. That is, for all T, r and $\epsilon, \delta > 0$, there exists $\epsilon' \in (0, \epsilon)$ such that for all $\epsilon'' \in (\epsilon', \epsilon)$

$$|QIC(T, \epsilon - \epsilon'') - QIC(T, \epsilon)| \leq \delta,$$
$$|QIC^r(T, \epsilon - \epsilon'') - QIC^r(T, \epsilon)| \leq \delta.$$

Lemma 6.3.7 (Quasi-convexity in input)**.** *For any $p \in [0, 1]$, define $\rho = p\rho_1 + (1-p)\rho_2$ for any two input states ρ_1, ρ_2. Then the following holds for any r-round protocol Π:*

$$QIC(\Pi, \rho) \geq pQIC(\Pi, \rho_1) + (1-p)QIC(\Pi, \rho_2)$$
$$QIC(\Pi, \rho) \leq pQIC(\Pi, \rho_1) + (1-p)QIC(\Pi, \rho_2) + rH(p).$$

Proof. The first inequality is Lemma 1.6.10, and the second is obtained by keeping track of the remainder terms discarded in its proof. Let R be a register holding a purification of ρ_1 and ρ_2, then we can purify ρ with two copies S_1, S_2 of a selector reference register, such that $|\rho\rangle^{A_{in}B_{in}RS_1S_2} = \sqrt{p}\,|\rho_1\rangle^{A_{in}B_{in}R}\,|1\rangle^{S_1}\,|1\rangle^{S_2} + \sqrt{1-p}\,|\rho_2\rangle^{A_{in}B_{in}R}\,|2\rangle^{S_1}\,|2\rangle^{S_2}$. We can then expand each term as

$$I(C_i; RS_1S_2|B_i)_\rho = I(C_i; S_1|B_i)_\rho + I(C_i; R|B_iS_1)_\rho + I(C_i; S_2|B_iRS_1)_\rho,$$

and similarly for terms conditioning on Alice's systems A_i. The result follows by summing over all rounds since

$$I(C_i; R|B_iS_1)_\rho = pI(C_i; R|B_i)_{\rho_1} + (1-p) \cdot I(C_i; R|B_i)_{\rho_2},$$

and then $H(S) = H(p)$ upper bounds the two remainder terms in each of the r rounds. □

Lemma 6.3.8 (Continuity in input)**.** *Quantum information cost for r-round protocols is*

uniformly continuous in the input distribution. This holds uniformly over all r-round protocols over input $X \times Y$. That is, for all $r, |X|, |Y|$, and $\epsilon > 0$, there exists $\delta > 0$ such that for all μ_1 and μ_2 that are δ-close and all r-round protocols Π,

$$|QIC(\Pi, \mu_1) - QIC(\Pi, \mu_2)| \leq \epsilon.$$

Proof. Let $\delta > 0$ and fix μ_1 and μ_2 that are δ-close. We can then write, for some common part μ_0 and remainder parts μ_1', μ_2',

$$\begin{aligned}
\mu_1 &= (1-\delta)\mu_0 + \delta\mu_1', \\
\mu_2 &= (1-\delta)\mu_0 + \delta\mu_2', \\
\mu_0(x,y) &= \frac{\min(\mu_1(x,y), \mu_2(x,y))}{\sum_{x',y'} \min(\mu_1(x',y'), \mu_2(x',y'))}.
\end{aligned}$$

Using the bounds in the lemma above once on each of μ_1 and μ_2, we get

$$\begin{aligned}
QIC(\Pi, \mu_1) &\leq (1-\delta)QIC(\Pi, \mu_0) + \delta QIC(\Pi, \mu_1') + rH(\delta) \\
&\leq (1-\delta)QIC(\Pi, \mu_0) + \delta QIC(\Pi, \mu_2') + \delta QIC(\Pi, \mu_1') + rH(\delta) \\
&\leq QIC(\Pi, \mu_2) + \delta \cdot r(\log|X| + \log|Y|) + rH(\delta).
\end{aligned}$$

Similarly, we get a bound on $QIC(\Pi, \mu_2)$ in terms of $QIC(\Pi, \mu_1)$, so the following holds:

$$|QIC(\Pi, \mu_1) - QIC(\Pi, \mu_2)| \leq \delta \cdot r(\log|X| + \log|Y|) + rH(\delta).$$

This bound is independent of μ_1, μ_2, depends on Π only through r and $|X|, |Y|$, and goes to zero as δ does, so the result follows. □

Corollary 6.3.9. *Suppose we have a r-round protocol Π for AND. Then,*

$$QIC(\Pi, \mu) \leq QIC(\Pi, \mu_0) + O(rH(w)) \tag{6.1}$$

where $w = \mu(1,1) \leq 1/2$, $\mu_0(1,1) = 0$, and $\mu_0(x_i, y_i) = \frac{1}{1-w}\mu(x_i, y_i)$ otherwise.

Proof. This just follows from the proof of lemma 6.3.8, since the input size is constant. □

Theorem 6.3.10. *For a relation $T \subset X \times Y \times Z_A \times Z_B$, an error parameter $\epsilon \in (0,1)$, a number of rounds r and each value $\alpha \in (0,1)$,*

$$QIC^r(T, \frac{\epsilon}{\alpha}) \leq \frac{QIC^r_D(T, \epsilon)}{1-\alpha}.$$

Proof. Fix T, r, ϵ, α and denote $I = QIC^r_D(T, \epsilon)$. For any $\delta_1 \in (0,1)$, we want to prove the existence of a protocol $\Pi \in \mathcal{T}^r(T, \frac{\epsilon}{\alpha} \cdot (1 + 2\delta_1))$ satisfying $QIC(\Pi, \mu) \leq \frac{I \cdot (1+2\delta_1)}{1-\alpha}$ for all $\mu \in \mathcal{D}_{XY}$. This shows that $QIC^r(T, \frac{\epsilon}{\alpha} \cdot (1+2\delta_1)) \leq \frac{I}{1-\alpha} \cdot (1+2\delta_1)$, and then by continuity of quantum information complexity in the error, we get the result by taking δ_1 to 0. The proof follows along the lines of the one for the analogous result for classical information complexity [Bra12], using a minimax argument. We take extra care to account for the continuum of quantum protocols, the round-by-round definition of quantum information cost, and the fact that we do not have a bound on the size of the entanglement. Let $\delta_2 \in (0, \epsilon\delta_1)$ satisfy the following two properties for all μ_1, μ_2 that are δ_2-close, and for all r-round protocols Π:

$$|QIC(\Pi, \mu_1) - QIC(\Pi, \mu_2)| \leq I \cdot \frac{\delta_1}{10}, \tag{6.2}$$

$$|QIC^r(T, \mu_1, \epsilon - \delta_2) - QIC^r(T, \mu_1, \epsilon)| \leq I \cdot \frac{\delta_1}{10}. \tag{6.3}$$

The first inequality is possible by Lemma 6.3.8, i.e. by the uniform continuity of quantum information cost in the input, uniformly over all r-rounds protocols, and the second is possible

by Lemma 6.3.4, i.e. the continuity of quantum information complexity in the error, uniformly over all inputs. Fix a finite δ_2-net for $\mathcal{D}_{XY}$, that we denote N_{XY}. For each $\mu \in N_{XY}$, fix a protocol $\Pi_\mu \in \mathcal{T}^r(T, \mu, \epsilon - \delta_2)$ such that $QIC(\Pi_\mu, \mu) \leq QIC^r(T, \mu, \epsilon - \delta_2) \cdot (1 + \frac{\delta_1}{10})$ and denote the set of all such protocols P_N. We then have $|P_N| = |N_{XY}| < \infty$, and we get using (6.3) that

$$\begin{aligned} QIC(\Pi_\mu, \mu) &\leq QIC^r(T, \mu, \epsilon - \delta_2) \cdot (1 + \frac{\delta_1}{10}) \\ &\leq \big(QIC^r(T, \mu, \epsilon) + I \cdot \frac{\delta_1}{10}\big)(1 + \frac{\delta_1}{10}) \\ &\leq I(1 + \frac{\delta_1}{10})^2 \\ &\leq I(1 + \frac{\delta_1}{2}). \end{aligned} \tag{6.4}$$

We define the following two-player zero-sum game over these two sets. Player A comes up with a quantum protocol $\Pi \in P_N$. Player B comes up with a distribution $\mu \in N_{XY}$. Player B's payoff is given by

$$P_B(\Pi, \mu) = (1 - \alpha) \cdot \frac{QIC(\Pi, \mu)}{I} + \alpha \cdot \frac{Pr_\mu[\Pi \notin T]}{\epsilon},$$

and then player A's is given by $P_A(\Pi, \mu) = -P_B(\Pi, \mu)$. We first show the following.

Claim 6.3.11. *The value of the game for player B is bounded by $1 + \delta_1$.*

Proof. Let ν_B be a probability distribution over N_{XY} representing a mixed strategy for player B. To prove the claim, it suffices to show that there is a protocol $\Pi \in P_N$ such that $\mathbb{E}_{\nu_B}[P_B(\Pi, \mu)] < 1 + \delta_1$. Let $\bar{\mu}$ be the distribution corresponding to averaging over ν_B, that is

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

Let $\mu' \in N_{XY}$ be a distribution that is δ_2-close to $\bar{\mu}$, and $\Pi' \in P_N$ the corresponding protocol. We will show that Π' is also good for $\bar{\mu}$. We first have

$$\begin{aligned} Pr_{\bar{\mu}}[\Pi' \notin T] &\leq Pr_{\mu'}[\Pi' \notin T] + \delta_2 \\ &\leq \epsilon - \delta_2 + \delta_2 \\ &= \epsilon, \end{aligned}$$

in which the first inequality follows from the fact that $\bar{\mu}$ and μ' are δ_2-close and the second inequality from the fact that $\Pi' \in P_N$ is the protocol corresponding to $\mu' \in N_{XY}$, i.e. $\Pi' \in \mathcal{T}^r(T, \mu', \epsilon - \delta_2)$. We also have

$$\begin{aligned} QIC(\Pi', \bar{\mu}) &\leq QIC(\Pi', \mu') + I \cdot \frac{\delta_1}{2} \\ &\leq I \cdot (1 + \delta_1), \end{aligned}$$

in which the first inequality follows from (6.2) and the second from the fact that $\Pi' \in P_N$ is the protocol corresponding to $\mu' \in N_{XY}$ along with (6.4). We obtain

$$\begin{aligned} \mathbb{E}_{\nu_B}[P_B(\Pi', \mu)] &= \mathbb{E}_{\nu_B}\big[(1-\alpha) \cdot \frac{QIC(\Pi', \mu)}{I} + \alpha \cdot \frac{Pr_{\mu}[\Pi' \notin T]}{\epsilon}\big] \\ &= (1-\alpha) \cdot \mathbb{E}_{\nu_B}\big[\frac{QIC(\Pi', \mu)}{I}\big] + \alpha \cdot \frac{Pr_{\bar{\mu}}[\Pi' \notin T]}{\epsilon} \\ &\leq (1-\alpha) \cdot \big[\frac{QIC(\Pi', \bar{\mu})}{I}\big] + \alpha \cdot \frac{Pr_{\bar{\mu}}[\Pi' \notin T]}{\epsilon} \\ &< (1-\alpha) \cdot (1 + \delta_1) + \alpha \\ &< 1 + \delta_1, \end{aligned}$$

in which the first equality is by definition, the second by linearity of expectation, the first inequality is by Lemma 1.6.10, i.e. concavity of quantum information cost in the input state, and the second inequality is by the above results about Π'. This concludes the proof of the

claim. □

By the minimax theorem for zero-sum games, the above claim implies that there exists a probability distribution ν_A over P_N representing a mixed strategy for player A and such that the value of the game for player B is at most $1 + \delta_1$. That is, for all $\mu \in N_{XY}$,

$$\mathbb{E}_{\nu_A}(P_B(\Pi, \mu)) < 1 + \delta_1.$$

Let $\bar{\Pi} = \mathbb{E}_{\nu_A}(\Pi)$ be the r-round protocol obtained by publicly averaging over ν_A, as per Lemma 1.6.8. This is the protocol we are looking for. The following claim holds.

Claim 6.3.12. *For all $\mu \in \mathcal{D}_{XY}$, $(1-\alpha) \cdot \frac{QIC(\bar{\Pi},\mu)}{I} + \alpha \cdot \frac{Pr_\mu[\bar{\Pi} \notin T]}{\epsilon} < 1 + 2\delta_1$.*

Proof. Fix any $\mu \in \mathcal{D}_{XY}$, and let $\mu' \in N_{XY}$ be a distribution that is δ_2-close to μ. Then we obtain

$$\begin{aligned}
&(1-\alpha) \cdot \frac{QIC(\bar{\Pi}, \mu)}{I} + \alpha \cdot \frac{Pr_\mu[\bar{\Pi} \notin T]}{\epsilon} \\
&\leq (1-\alpha) \cdot \frac{QIC(\bar{\Pi}, \mu') + I\delta_1}{I} + \alpha \cdot \frac{Pr_{\mu'}[\bar{\Pi} \notin T] + \delta_2}{\epsilon} \\
&= (1-\alpha) \cdot \frac{QIC(\bar{\Pi}, \mu')}{I} + \alpha \cdot \mathbb{E}_{\nu_A} \frac{Pr_{\mu'}[\Pi \notin T]}{\epsilon} + (1-\alpha) \cdot \delta_1 + \alpha \cdot \frac{\delta_2}{\epsilon} \\
&\leq (1-\alpha) \cdot \mathbb{E}_{\nu_A}\big[\frac{QIC(\Pi, \mu')}{I}\big] + \alpha \cdot \mathbb{E}_{\nu_A}\big[\frac{Pr_{\mu'}[\Pi \notin T]}{\epsilon}\big] + \delta_1 \\
&= \mathbb{E}_{\nu_A}[P_B(\Pi, \mu')] + \delta_1 \\
&< 1 + 2\delta_1,
\end{aligned}$$

in which the first inequality follows from (6.2) and the fact that μ, μ' are δ_2-close, the first equality is because we take expectation over a probability, the second inequality is because $\delta_2 \leq \epsilon \cdot \delta_1$ and by Lemma 1.6.8, i.e. by the convexity of quantum information cost in the protocol, the second equality is by linearity of expectation and the definition of $P_B(\Pi, \mu')$,

and the last inequality is because ν_A represents the mixed strategy obtained by the minimax theorem. Since this holds for all $\mu \in \mathcal{D}_{XY}$, this conclude the proof of the claim. □

To conclude the proof of the theorem, we first note that the above claim implies that for all $\mu \in \mathcal{D}_{XY}$,

$$QIC(\bar{\Pi}, \mu) \leq \frac{I}{1-\alpha}(1 + 2\delta_1),$$

so $\bar{\Pi}$ satisfies the quantum information cost property we are looking for. Is left to verify that it also has low error on all inputs. The above claim also implies that for all μ,

$$Pr_{\mu}[\bar{\Pi} \notin T] \leq \frac{\epsilon}{\alpha} \cdot (1 + 2\delta_1).$$

Letting μ run over all atomic distributions, we get the desired error property, and so

$$QIC^r(T, \frac{\epsilon}{\alpha} \cdot (1 + 2\delta_1)) \leq \frac{I}{1-\alpha}(1 + 2\delta_1),$$

as desired. □

Theorem 6.3.13. *For a relation $T \subset X \times Y \times Z_A \times Z_B$, an error parameter $\epsilon \in (0, 1)$ and each value $\alpha \in (0, 1)$,*

$$QIC(T, \frac{\epsilon}{\alpha}) \leq \frac{QIC_D(T, \epsilon)}{1-\alpha}.$$

Proof. Let $I = QIC_D(T, \epsilon)$, and denote by $P_e^{\mu}(\Pi)$ the average error of Π for computing T on μ, and by P_T the set of all protocols over the same input and output spaces as T. Then for any Π, $P_e^{\mu}(\Pi)$ is continuous in μ by properties of the statistical distance. Given $\delta > 0$,

define

$$A(\Pi) = \{\mu \in \mathcal{D}_{XY} : QIC(\Pi, \mu) \geq I + 2 \cdot \delta \text{ or } P_e^{\mu}(\Pi) \geq \epsilon + \delta\}.$$

By continuity of $QIC(\Pi, \mu)$ and $P_e^{\mu}(\Pi)$ in μ, these sets are closed for all $\Pi \in P_T$. Then, by definition of I, for all μ there exists $\Pi_\mu \in \mathcal{T}(T, \mu, \epsilon)$ such that $QIC(\Pi_\mu, \mu) \leq I + \delta$, and so $\cap_{\Pi \in P_T} A(\Pi) = \emptyset$. Since $\mathcal{D}_{XY}$ is compact and the sets $A(\Pi)$ are closed, we get that there exists a finite set $Q \subset P_T$ such that $\cap_{\Pi \in Q} A(\Pi) = \emptyset$. We get that for all μ, there exists $\Pi_\mu \in Q$ such that $QIC(\Pi_\mu, \mu) < I + 2\delta$ and $P_e^{\mu}(\Pi_\mu) < \epsilon + \delta$. Let $r_M = \max\{r :$ there is $\Pi \in Q$ with r rounds $\}$, then

$$\begin{aligned} I + 2\delta &\geq \max_{\mu} \min_{\Pi \in Q \cap \mathcal{T}(T, \mu, \epsilon+\delta)} QIC(\Pi, \mu) \\ &\geq QIC_D^{r_M}(T, \epsilon + \delta) \\ &\geq (1 - \alpha) \cdot QIC^{r_M}(T, \frac{\epsilon}{\alpha} + \frac{\delta}{\alpha}) \\ &\geq (1 - \alpha) \cdot QIC(T, \frac{\epsilon}{\alpha} + \frac{\delta}{\alpha}). \end{aligned}$$

The result follows by continuity of QIC and by taking δ to zero. □

6.3.2 Subadditivity

Lemma 6.3.14. *For any two protocols Π^1, Π^2 with r_1, r_2 rounds, respectively, there exists a r-round protocol Π_2, satisfying $\Pi_2 = \Pi^1 \otimes \Pi^2, r = \max(r_1, r_2)$, such that the following holds for any joint input state $\rho_{12} \in \mathcal{D}(A_{in}^1 \otimes B_{in}^1 \otimes A_{in}^2 \otimes B_{in}^2)$:*

$$QIC(\Pi_2, \rho_{12}) \leq QIC(\Pi^1, \rho_1) + QIC(\Pi^2, \rho_2),$$

with $\rho_1 = \mathrm{Tr}_{A_{in}^2 B_{in}^2}(\rho_{12})$ and $\rho_2 = \mathrm{Tr}_{A_{in}^1 B_{in}^1}(\rho_{12})$.

Proof. Given protocols Π^1 and Π^2, we assume without loss of generality that $r_1 \geq r_2$, and we define the protocol Π_2 in the following way.

1. Run protocols Π^1, Π^2 in parallel for r_2 rounds, on corresponding input registers $A^1_{in}, B^1_{in}, A^2_{in}, B^2_{in}$ until Π^2 has finished.

2. Finish running protocol Π^1

3. Take as output the output registers $A^1_{out}, B^1_{out}, A^2_{out}, B^2_{out}$ of both Π^1 and Π^2.

It is clear that the channel that Π_2 implements is $\Pi_2 = \Pi^1 \otimes \Pi^2$, and the number of rounds satisfies $r = \max(r_1, r_2)$, so is left to analyze its quantum information cost on input ρ_{12}. Let R_{12} be a purifying register such that $\rho_{12}^{A^1_{in}B^1_{in}A^2_{in}B^2_{in}R_{12}}$ is a pure state. Also, denote the purified joint state in round i as $(\rho^i_{12})^{A^1_iB^1_iC^1_iA^2_iB^2_iC^2_iR_{12}}$, and the local state for protocol Π^1 as

$$(\rho^i_1)^{A^1_iB^1_iC^1_i} = \mathrm{Tr}_{A^2_iB^2_iC^2_iR_{12}}((\rho^i_{12})^{A^1_iB^1_iC^1_iA^2_iB^2_iC^2_iR_{12}}), \tag{6.5}$$

and similarly for that of protocol Π^2. Notice that for all i, $(\rho^i_1)^{A^1_iB^1_iC^1_i}$ is purified by $(\rho^i_1)^{A^1_iB^1_iC^1_iA^2_{in}B^2_{in}R_{12}} \otimes \phi_2^{T^2_AT^2_B}$, with $A^2_{in}B^2_{in}R_{12}$ the registers of state ρ_{12} before application of the unitaries corresponding to Π^1, and ϕ_2 is the pure entangled state used in Π_2. If we denote, for $i \geq r_2 + 1, A^2_i = A^2_{out} \otimes (A')^2, B^2_i = B^2_{out} \otimes (B')^2$, then by the definition of QIC

and application of chain rule,

$$\begin{aligned}
2 \cdot QIC(\Pi_2, \rho_{12}) &= \sum_{i=1,i\ odd}^{r_2} I(C_i^1 C_i^2; R_{12}|B_i^1 B_i^2)_{\rho_{12}} + \sum_{i=1,i\ even}^{r_2} I(C_i^1 C_i^2; R_{12}|A_i^1 A_i^2)_{\rho_{12}} \\
&+ \sum_{i=r_2+1,i\ odd}^{r_1} I(C_i^1; R_{12}|B_i^1 B_i^2)_{\rho_{12}} + \sum_{i=r_2+1,i\ even}^{r_1} I(C_i^1; R_{12}|A_i^1 A_i^2)_{\rho_{12}} \\
&= \sum_{i=1,i\ odd}^{r_2} I(C_i^2; R_{12}|B_i^1 B_i^2 C_i^1)_{\rho_{12}} + \sum_{i=1,i\ even}^{r_2} I(C_i^2; R_{12}|A_i^1 A_i^2 C_i^1)_{\rho_{12}} \\
&+ \sum_{i=1,i\ odd}^{r_1} I(C_i^1; R_{12}|B_i^1 B_i^2)_{\rho_{12}} + \sum_{i=1,i\ even}^{r_1} I(C_i^1; R_{12}|A_i^1 A_i^2)_{\rho_{12}}.
\end{aligned}$$

Now for protocol Π^1, as noted above, the registers $A_{in}^2 B_{in}^2 R_{12} T_A^2 T_B^2$ purify $(\rho_1^i)^{A_i^1 B_i^1 C_i^1}$ for all i, so

$$\begin{aligned}
2 \cdot QIC(\Pi^1, \rho_1) &= \sum_{i=1,i\ odd}^{r_1} I(C_i^1; A_{in}^2 B_{in}^2 R_{12} T_A^2 T_B^2|B_i^1)_{\rho_1} + \sum_{i=1,i\ even}^{r_1} I(C_i^1; A_{in}^2 B_{in}^2 R_{12} T_A^2 T_B^2|A_i^1)_{\rho_1} \\
&= \sum_{i=1,i\ odd}^{r_1} I(C_i^1; A_i^2 B_i^2 C_i^2 R_{12}|B_i^1)_{\rho_{12}} + \sum_{i=1,i\ even}^{r_1} I(C_i^1; A_i^2 B_i^2 C_i^2 R_{12}|A_i^1)_{\rho_{12}} \\
&= \sum_{i=1,i\ odd}^{r_1} I(C_i^1; B_i^2|B_i^1)_{\rho_{12}} + \sum_{i=1,i\ even}^{r_1} I(C_i^1; A_i^2|A_i^1)_{\rho_{12}} \\
&+ \sum_{i=1,i\ odd}^{r_1} I(C_i^1; R_{12}|B_i^1 B_i^2)_{\rho_{12}} + \sum_{i=1,i\ even}^{r_1} I(C_i^1; R_{12}|A_i^1 A_i^2)_{\rho_{12}} \\
&+ \sum_{i=1,i\ odd}^{r_1} I(C_i^1; A_i^2 C_i^2|B_i^1 B_i^2 R_{12})_{\rho_{12}} + \sum_{i=1,i\ even}^{r_1} I(C_i^1; B_i^2 C_i^2|A_i^1 A_i^2 R_{12})_{\rho_{12}} \\
&\geq \sum_{i=1,i\ odd}^{r_1} I(C_i^1; R_{12}|B_i^1 B_i^2)_{\rho_{12}} + \sum_{i=1,i\ even}^{r_1} I(C_i^1; R_{12}|A_i^1 A_i^2)_{\rho_{12}},
\end{aligned}$$

in which the first equality is by definition, the second is by isometric invariance of the conditional quantum mutual information (CQMI), the third by the chain rule for CQMI, and the inequality is by non-negativity of CQMI. Similarly for protocol Π^2, with a slightly

different application of the chain rule, we get

$$
\begin{aligned}
2 \cdot QIC(\Pi^2, \rho_2) &= \sum_{i=1, i\ odd}^{r_2} I(C_i^2; A_{in}^1 B_{in}^1 R_{12} T_A^1 T_B^1 | B_i^2)_{\rho_2} + \sum_{i=1, i\ even}^{r_2} I(C_i^2; A_{in}^1 B_{in}^1 R_{12} T_A^1 T_B^1 | A_i^2)_{\rho_2} \\
&= \sum_{i=1, i\ odd}^{r_2} I(C_i^2; A_i^1 B_i^1 C_i^1 R_{12} | B_i^2)_{\rho_{12}} + \sum_{i=1, i\ even}^{r_2} I(C_i^2; A_i^1 B_i^1 C_i^1 R_{12} | A_i^2)_{\rho_{12}} \\
&= \sum_{i=1, i\ odd}^{r_2} I(C_i^2; B_i^1 C_i^1 | B_i^2)_{\rho_{12}} + \sum_{i=1, i\ even}^{r_2} I(C_i^2; A_i^1 C_i^1 | A_i^2)_{\rho_{12}} \\
&+ \sum_{i=1, i\ odd}^{r_2} I(C_i^2; R_{12} | B_i^1 B_i^2 C_i^1)_{\rho_{12}} + \sum_{i=1, i\ even}^{r_2} I(C_i^2; R_{12} | A_i^1 A_i^2 C_i^1)_{\rho_{12}} \\
&+ \sum_{i=1, i\ odd}^{r_2} I(C_i^2; A_i^1 | B_i^1 B_i^2 C_i^1 R_{12})_{\rho_{12}} + \sum_{i=1, i\ even}^{r_2} I(C_i^2; B_i^2 | A_i^1 A_i^2 C_i^1 R_{12})_{\rho_{12}} \\
&\geq \sum_{i=1, i\ odd}^{r_2} I(C_i^2; R_{12} | B_i^1 B_i^2 C_i^1)_{\rho_{12}} + \sum_{i=1, i\ even}^{r_2} I(C_i^2; R_{12} | A_i^1 A_i^2 C_i^1)_{\rho_{12}}.
\end{aligned}
$$

The result then follows by comparing terms. □

6.3.3 Reducing the Error for Functions

Similarly to communication, it is possible to reduce the error when computing functions without increasing too much the information.

Lemma 6.3.15. *For any function f and error parameter $\epsilon > 0$, the following holds:*

$$QIC(f, \epsilon) \leq O\big(\log 1/\epsilon \cdot QIC(f, 1/3)\big).$$

Proof. Given $\delta > 0$, let Π be a protocol computing f correctly except with probability $1/3$ on every input and satisfying $QIC(\Pi, \mu) \leq QIC(f, 1/3) + \delta$ for all μ. Let $n \in O(\log 1/\epsilon)$ be given by the Chernoff bound such that protocol Π_n running Π n times in parallel as per Lemma 6.3.14, with each input being a copy of the instance to f, and taking a majority vote (with arbitrary tie-breaking) computes f correctly except with probability ϵ on every input.

This n can be chosen independently of δ. We now argue on the quantum information cost of Π_n. Consider an arbitrary distribution μ for f, and let μ_n be the distribution once the n copies have been made. If we denote the marginal for the i-th copy by μ^i, then $\mu^i = \mu$. By Lemma 6.3.14 and an easy induction, we then get that

$$\begin{aligned} QIC(f, \epsilon) &\leq QIC(\Pi_n, \mu_n) \\ &\leq nQIC(\Pi, \mu) \\ &\leq n(QIC(f, 1/3) + \delta). \end{aligned}$$

The result follows by taking δ to 0. □

6.3.4 Reduction from DISJ to AND

With the following definition, the above proof also establishes the following corollary.

Definition 6.3.16. For all $r \in \mathbb{N}, \epsilon \in [0, 1]$,

$$QIC_0^r(AND, \epsilon) = \inf_{\Pi \in \mathcal{T}^r(AND, \epsilon)} \max_{\mu_0} QIC(\Pi, \mu_0),$$

in which the maximum ranges over all μ_0 satisfying $\mu_0(1, 1) = 0$.

Corollary 6.3.17. *For any $\epsilon > 0$ and $r \in \mathbb{N}$,*

$$QIC_0^r(AND, \epsilon) \leq O\big(\log 1/\epsilon \cdot QIC_0^r(AND, 1/3)\big).$$

We provide a slight variant of the argument of [Tou15] to obtain a low information protocol for AND from a protocol for disjointness.

Lemma 6.3.18. *For any n, r, ϵ and μ_0 such that $\mu_0(1,1) = 0$,*

$$\inf_{\Pi_A \in \mathcal{T}^r(AND,\epsilon)} QIC(\Pi_A, \mu_0) \leq \inf_{\Pi_D \in \mathcal{T}^r(DISJ_n,\epsilon)} \frac{1}{n} QIC(\Pi_D, \mu_0^{\otimes n}).$$

Proof. Let $I_n = \inf_{\Pi_D \in \mathcal{T}^r(DISJ_n,\epsilon)} QIC(\Pi_D, \mu_0^{\otimes n})$. We prove the result by induction on n. The base case is trivial since $DISJ_1 = \neg AND$, and so a protocol to compute $DISJ_1$ with error ϵ can be used to compute AND with error ϵ and vice-versa. In particular, we get $I_1 = \inf_{\Pi_A \in \mathcal{T}^r(AND,\epsilon)} QIC(\Pi_A, \mu_0)$. For the induction, suppose the result holds for $DISJ_{n-1}$, we will use Lemma 1.6.7 to go from $DISJ_n$ to $DISJ_1$ and $DISJ_{n-1}$. Indeed, given $\delta > 0$ and Π_D computing $DISJ_n$ with error ϵ and satisfying $QIC(\Pi_D, \mu_0^{\otimes n}) \leq I_n + \delta$, we can use Lemma 1.6.7 with $\rho_1 = \mu_0, \rho_2 = \mu_0^{\otimes n-1}$ and then it is clear that Π^1 computes $DISJ_1$ with error ϵ and Π^2 computes $DISJ_{n-1}$ with error ϵ. We get

$$\begin{aligned} I_n + \delta &\geq QIC(\Pi_D, \mu_0^{\otimes n}) \\ &= QIC(\Pi^1, \mu_0) + QIC(\Pi^2, \mu_0^{\otimes n-1}) \\ &\geq I_1 + I_{n-1} \\ &\geq nI_1. \end{aligned}$$

□

The following lemma is very similar to Theorem 6.3.10. The only difference is that the distributions we consider are restricted and on the right hand side the error of the protocol is measured in the worst case. Since the error is worst case, there is no loss in the error, and the payoff function would be simply $P_B(\Pi, \mu) = QIC(\Pi, \mu)/I$.

Lemma 6.3.19.

$$QIC_0^r(AND, \epsilon) = \max_{\mu_0, \mu_0(1,1)=0} \inf_{\Pi \in \mathcal{T}^r(AND,\epsilon)} QIC(\Pi, \mu_0)$$

Lemma 6.3.20. *For all $r, n \in \mathbb{N}$,*

$$QCC^r(DISJ_n, 1/3) \geq n \cdot QIC_0^r(AND, 1/3)$$

Proof. The result follows from the following chain of inequality:

$$\begin{aligned} QCC^r(DISJ_n, 1/3) &\geq QIC^r(DISJ_n, 1/3) \\ &\geq \max_{\mu_0} \inf_{\Pi_D \in \mathcal{T}^r(DISJ_n, 1/3)} QIC(\Pi_D, \mu_0^{\otimes n}) \\ &\geq \max_{\mu_0} \inf_{\Pi_A \in \mathcal{T}^r(AND, 1/3)} n \cdot QIC(\Pi_A, \mu_0) \\ &\geq n \cdot QIC_0^r(AND, 1/3). \end{aligned}$$

The first inequality is by Lemma 6.3.3, the second since, on the r.h.s., the maximization is over a smaller set of product distributions with $\mu_0(1,1) = 0$ and the minimization over a larger set of protocols, the third is by Lemma 6.3.18, and the last is by Lemma 6.3.19. □

6.4 Lower bound on QIC by generalized discrepancy method

6.4.1 Compression

Definition 6.4.1. We say that $QCC(f^k, \mu^k, \eta_1 k, \eta_2) \leq C$ if there exists a protocol π for f^k s.t. $QCC(\pi) \leq C$ and

$$Pr[\pi \text{ computes} \geq \eta_1 k \text{ coordinates correctly}] \geq 1 - \eta_2$$

Here the probability is both over the distribution μ^k and the randomness of protocol (which includes the randomness due to quantum measurements). We don't require the protocol to declare which coordinates were computed correctly.

Lemma 6.4.2. *If there exists a protocol* Π *for* f *with error* $\leq \epsilon$ *w.r.t* μ *s.t.* $QIC(\Pi, \mu) = I$*, then for all* $\epsilon', \delta > 0$*, there exists* $k_0(\Pi, \mu, \epsilon', \delta)$ *such that for all* $k \geq k_0$*,* $QCC(f^k, \mu^k, (1 - 2\epsilon)k, e^{-2\epsilon^2 k} + \epsilon') \leq k(I + \delta)$.

Proof. Suppose $(E_1, \ldots, E_k)$ is the vector of indicator random variables of the errors in various coordinates of $\Pi^{\otimes k}$ i.e. $E_i = 1$ if error occurred on the i^{th} coordinate. Also look at Π_k obtained from lemma 1.6.11 for large enough k with parameters $2\epsilon'$, δ and where ρ is μ. Suppose $(E_1^{'}, \ldots, E_k^{'})$ is the vector of errors for Π_k. According to lemma 1.6.11, Π_k satisfies the following:

$$\mathbb{E}_{((x_1,\ldots,x_k),(y_1,\ldots,y_k))\sim\mu^k} ||\Pi_k((x_1, \ldots, x_k), (y_1, \ldots, y_k)) - \Pi^{\otimes k}((x_1, \ldots, x_k), (y_1, \ldots, y_k))||_1 \leq 2\epsilon'$$

Hence it follows that

$$||(E_1, \ldots, E_k) - (E_1^{'}, \ldots, E_k^{'})||_{\text{TV}} \leq \epsilon'$$

Here $||P - Q||_{\text{TV}}$ is the total variation distance between the distributions P and Q (we are not distinguishing between random variables and their distributions). Since $\Pr[\sum_i E_i \geq 2\epsilon k] \leq e^{-2\epsilon^2 k}$ by Chernoff bounds, it follows that

$$\Pr\left[\sum_i E_i^{'} \geq 2\epsilon k\right] \leq e^{-2\epsilon^2 k} + \epsilon'$$

which implies the lemma along with the fact that $QCC(\Pi_k) \leq (I + \delta)k$. □

6.4.2 Average case to worst case

In this section, we prove the following lemma which turns a protocol for average case input to a protocol for worst case input.

Lemma 6.4.3. *Suppose* $f_n : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ *is an arbitrary boolean function. Let* $k \geq 2^{5n}$ *and* $\epsilon > 10k^{-0.005}$. *Assume for any product input distribution* μ^k, *there exists a protocol* π_{μ^k} *with* $QCC\left(\pi_{\mu^k}\right) \leq l$ *that computes at least* $(1-\alpha)\,k$ *coordinates of* f_n^k *correctly with probability at least* γ. *Then there exists a protocol* τ *s.t. for any input* $((x_1, \cdots, x_k), (y_1, \cdots, y_k))$, *for any integer* $c \geq 3$ *and constant* $\epsilon > 0$, τ *computes at least* $\left(1 - 2^{-c/2} - c\alpha\right) k$ *coordinates of* f_n^k *correctly with probability at least* $\frac{1}{2}\left(\left(\frac{\gamma}{(1+\epsilon)^k}\right)^c - 2^{-2^{2-2c}k}\right)$. *Also* $QCC(\tau) \leq c \cdot l + o(k)$.

Proof. In this lemma, we want to construct a protocol τ which works for an arbitrary input based on protocols which work on product input distributions (product across coordinates). The main idea of the proof is that corresponding to any input $((x_1, ..., x_k), (y_1, ..., y_k))$ (x_i and y_i are inputs of a f_n instance and have n bits), we can associate a μ, which is the empirical distribution:

$$\mu(x, y) = \frac{\#\text{ of } i, (x_i, y_i) = (x, y)}{k}.$$

So it makes sense to construct τ from π_{μ^k}. The players can simulate μ^k by sampling independent coordinates from their input (with replacement). However the issue is that the players don't know μ, so they have no idea what π_{μ^k} is. So in the actual protocol Alice and Bob will first sample some coordinates to get an estimate $\tilde{\mu}$ of μ and then run protocol $\pi_{\tilde{\mu}^k}$. The protocol τ is described in Protocol 8.

Now let's analyze this protocol. We first need the following two lemmas to show how to get an estimate $\tilde{\mu}$ of μ.

Lemma 6.4.4. *After communicating* $O(k^{0.52} \log k)$ *bits, for some specific input* (x, y), *with*

Inputs: $(x_1, \ldots, x_k)$ and $(y_1, \ldots, y_k)$

1. Get an estimate $\tilde{\mu}$ of μ.
2. Alice and Bob use shared randomness to obtain random independent samples from $[k]$, $j_1, \ldots, j_{ck}$. Run the protocol $\pi_{\tilde{\mu}^k}$ c times. In the t^{th} iteration, the protocol is run on inputs $(x_{j_{(t-1)k+1}}, \ldots, x_{j_{tk}}), (y_{j_{(t-1)k+1}}, \ldots, y_{j_{tk}})$. In the process we obtain answers for various coordinates (some of the coordinates will be sampled multiple times and we will obtain multiple answers for them).
3. If a coordinate was sampled in the previous step, output the answer $\pi_{\tilde{\mu}^k}$ gave for it. If they got multiple results on one coordinate, they will output the first one. If a coordinate was not sampled, output 0 on that coordinate.

Protocol 8: Protocol τ

success probability at least $1 - 1/k$*, Alice and Bob know* $\mu(x, y)$ *exactly if* $\mu(x, y) \cdot k < k^{0.02}$*, otherwise Alice and Bob know that* $\mu(x, y) \cdot k \geq k^{0.02}$*.*

Proof. In [BCW98], they showed that to compute the disjointness between two inputs of length k, the quantum communication complexity is $O(\sqrt{k} \log k)$. The corresponding protocol has constant error rate and will find one intersection place. We will use this protocol to solve our problem by the following reduction. For each input (x_i, y_i), we set $a_i = 1_{x_i = x}$ and $b_i = 1_{y_i = y}$. Then finding (x, y) in the input is just like finding intersection between $a = (a_1, ..., a_k)$ and $b = (b_1, ..., b_k)$. Protocol 9 shows how to finish the task described in the lemma.

1. Set a and b as we just described. Set $cnt = 0$.
2. Do the following step $c_1 \cdot k^{0.02}$ times, c_1 is some constant to be figured out in the proof:
3. Use protocol for DISJ in [BCW98] to find the intersection between a and b, let it be at place j, Alice and Bob communicate 2 bits to check if $a_j = b_j = 1$. If it is true, then set $cnt = cnt + 1$, $a_j = 0$, $b_j = 0$.

Protocol 9: Protocol count

Let's analyze this protocol. First its quantum communication cost is clear to be $O(k^{0.52} \log k)$ as the DISJ protocol has quantum communication cost $O(\sqrt{k} \log k)$. Then for each repeat

of step 3, if the DISJ protocol gives wrong answer, we will not do anything. And if the DISJ protocol gives the correct intersection, the counter will be increased by one and the intersection place will be removed and we can find other intersections. Thus we only have to show with probability at least $1 - 1/k$, DISJ protocol gives a correct answer for at least $k^{0.02}$ times. Assume the DISJ protocol succeeds with some constant probability p. Let Cr denote the random variable for the number of correct answers DISJ protocol gives. We know $\mathbb{E}[Cr] = p \cdot c_1 \cdot k^{0.02}$. By the additive Chernoff bound, the probability that DISJ protocol give a correct answer for at least $k^{0.02}$ times is

$$\Pr[Cr \geq k^{0.02}] = 1 - \Pr[Cr < k^{0.02}] \geq 1 - e^{-2(p \cdot c_1 \cdot k^{0.02} - k^{0.02})^2/(c_1 \cdot k^{0.02})}.$$

By picking c_1 properly, for example $c_1 = 2/p$, we get $\Pr[Cr \geq k^{0.02}] \geq 1 - 1/k$. □

Lemma 6.4.5. *Let $\epsilon > 10k^{-0.005}$ be some constant. After communicating $O\left(k^{0.99} \cdot n + 2^{2n} \cdot k^{0.52} \log k\right)$ bits, with probability at least $1/2$, Alice and Bob agree on some $\tilde{\mu}$, such that for any (x, y), $\frac{\tilde{\mu}(x,y)}{\mu(x,y)} < 1 + \epsilon$.*

Proof. We use the following protocol to estimate μ:

Inputs: $(x_1, \ldots, x_k)$ and $(y_1, \ldots, y_k)$

1. Sample the coordinates randomly $k^{0.99}$ times using public randomness (with replacement). Alice and Bob exchange their input for these coordinates. For each $(x, y) \in \{0,1\}^n \times \{0,1\}^n$, count the number of times it appears in these coordinates and denote the count by $c_1(x, y)$.
2. For all (x, y), use Lemma 6.4.4 to count the number of times (x, y) appears in the input and denote the count obtained by $c_2(x, y)$.
3. We combine c_1 and c_2 as c_3. For each (x, y), if $c_2(x, y) \geq k^{0.02}$, let $c_3(x, y) = c_1(x, y) \cdot k^{0.01}$ otherwise $c_3(x, y) = c_2(x, y)$.
4. $\tilde{\mu}(x, y) = \frac{c_3(x,y)}{\sum_{x',y'} c_3(x',y')}$.

Protocol 10: Estimate μ

Let's first analyze the communication cost of this part. It's clear that the first step needs at most $O\left(k^{0.99}n\right)$ communication. For second step, by Lemma 6.4.4, it needs at most $O\left(2^{2n}\cdot k^{0.52}\log k\right)$ communication. Sum them up, this protocol needs $O\left(k^{0.99}\cdot n+2^{2n}\cdot k^{0.52}\log k\right)$ bits of communication.

Then let's consider the following events:

1. For all (x,y) such that $\mu\left(x,y\right)\cdot k\geq k^{0.02}$, $\left|c_1\left(x,y\right)\cdot k^{0.01}-\mu\left(x,y\right)\cdot k\right|<\frac{\epsilon}{3}\mu\left(x,y\right)\cdot k$.

2. For any (x,y), the protocol described in Lemma 6.4.4 does not fail.

If these two events happen, then we know that $\left|c_3\left(x,y\right)-\mu\left(x,y\right)\cdot k\right|<\frac{\epsilon}{3}\mu\left(x,y\right)\cdot k$, therefore as desired,

$$\tilde{\mu}\left(x,y\right)=\frac{c_3\left(x,y\right)}{\sum_{x',y'}c_3\left(x',y'\right)}\leq\frac{\left(1+\frac{\epsilon}{3}\right)\mu\left(x,y\right)\cdot k}{\left(1-\frac{\epsilon}{3}\right)\cdot k}<\left(1+\epsilon\right)\mu\left(x,y\right).$$

Finally, we only have to make sure that these two events happen with probability at least $1/2$. For the first event, by the multiplicative Chernoff bound and union bound, it does not happen with probability

$$2^{2n}\cdot\Pr[\left|c_3\left(x,y\right)/k^{0.99}-\mu\left(x,y\right)\right|>\frac{\epsilon}{3}\mu\left(x,y\right)]<2ke^{-\frac{\left(\epsilon/3\right)^2\mu(x,y)k^{0.99}}{3}}\leq2ke^{-\epsilon^2k^{0.01}/27}<1/4.$$

For the second event, by Lemma 6.4.4 and the union bound, it does not happen with probability at most $2^{2n}\cdot\frac{1}{k}<1/4$. Thus these two events happen with probability at least $1/2$. $\square$

Let's consider the communication cost of τ. For the first step, the cost is

$$O\left(k^{0.99}\cdot n+2^{2n}\cdot k^{0.52}\log k\right)=o\left(k\right)$$

For the second step, the quantum communication complexity is at most $c\cdot l$. For the third step, the cost is 0. Therefore $QCC\left(\tau\right)\leq c\cdot l+o\left(k\right)$.

Let's say that the protocol τ succeeds when the following things happen:

1. For all (x, y), $\frac{\tilde{\mu}(x,y)}{\mu(x,y)} < 1 + \epsilon$.

2. The c runs of protocol $\pi_{\tilde{\mu}^k}$ in step 2 of protocol τ all compute at least $(1 - \alpha) k$ coordinates correctly.

3. Number of $i \in [k]$ such that the coordinate i is not sampled in step 2 of protocol τ is at most $2^{-c/2}k$.

If τ succeeds, then it computes at least $\left(1 - 2^{-c/2} - c\alpha\right) k$ coordinates correctly. This is because errors come from two possible ways:

1. Some coordinates are not sampled. When τ succeeds, the number of coordinates that are not sampled is at most $2^{-c/2}k$.

2. Some coordinates' results are wrong in step 2. When τ succeeds, the number of errors from step 2 is at most αck.

Finally, let's analyze the success probability of protocol τ. Let's analyze step by step:

1. For step one, by Lemma 6.4.5, it is clear that we succeed with probability 1/2.

2. For step two, first we know that when running $\pi_{\tilde{\mu}^k}$ on distribution $\tilde{\mu}^k$, we succeed with probability at least γ. And since we have for any (x, y), $\frac{\tilde{\mu}(x,y)}{\mu(x,y)} < 1 + \epsilon$, if we run $\pi_{\tilde{\mu}^k}$ on distribution μ^k, the success probability will be at least $\frac{\gamma}{(1+\epsilon)^k}$. When running this protocol c times independently, the success probability will be at least $\left(\frac{\gamma}{(1+\epsilon)^k}\right)^c$. Note that when we sample coordinates independently at random, the distribution we induce is μ^k.

3. It is only left to analyze the probability that number of coordinates not sampled in step 2 of protocol τ is at least $2^{-c/2}k$. For each coordinate i, define s_i to be the random

variable that indicates whether coordinate i is sampled or not (1 means not sampled and 0 means sampled). Then we have $\mathbb{E}[s_i] = \left(1 - \frac{1}{k}\right)^{ck} < 2^{-c}$. In order to show the failure probability small by Chernoff bound, we will show that all the s_i's are negatively correlated. To show they are negatively correlated, we only have to show

$$\forall I \subseteq [k], \Pr\left[\prod_{i \in I} s_i = 1\right] \leq \prod_{i \in I} \Pr[s_i = 1].$$

Notice that $\Pr\left[\prod_{i \in I} s_i = 1\right] = \left(1 - \frac{|I|}{k}\right)^{kc}$ and $\Pr[s_i = 1] = \left(1 - \frac{1}{k}\right)^{kc}$. So we have,

$$\forall I \subseteq [k], \Pr\left[\prod_{i \in I} s_i = 1\right] = \left(1 - \frac{|I|}{k}\right)^{kc} \leq \left(\left(1 - \frac{1}{k}\right)^{|I|}\right)^{kc} = \prod_{i \in I} \Pr[s_i = 1].$$

Since all the s_i's are negatively correlated, by Chernoff bound for negatively correlated random variables, for example see [DP], we have that the failure probability

$$\Pr\left[\sum_{i=1}^{k} s_i \geq 2^{-c/2}k\right] < e^{-2k\left(2^{-c/2} - 2^{-c}\right)^2} < e^{-2^{2-2c}k} < 2^{-2^{2-2c}k}.$$

The second inequality holds for all $c \geq 3$. Notice that the event that we err in the first step is independent from the event that we err in the second step. So the success probability of τ is at least $\frac{1}{2}\left(\left(\frac{\gamma}{(1+\epsilon)^k}\right)^c - 2^{-2^{2-2c}k}\right)$.

□

6.4.3 Lower bound on QIC

Definition 6.4.6. We say that $QCC(f^k, \eta_1 k, \eta_2) \leq C$ if there exists a protocol π for f^k s.t. $QCC(\pi) \leq C$ and

$$Pr[\pi \text{ computes} \geq \eta_1 k \text{ coordinates correctly}] \geq 1 - \eta_2$$

Here the probability is over randomness of protocol (which includes the randomness due to quantum measurements). We don't require the protocol to declare which coordinates were computed correctly.

Theorem 6.4.7. *There exists an absolute constant $\eta > 0$ s.t. for any boolean function f, $QIC_D(f, \eta) \geq \Omega(GDM_{1/5}(f) - O(1))$.*

Proof. Let $\eta > 0$ be a sufficiently small constant to be fixed later. Suppose $\max_\mu QIC(f, \mu, \eta) = I$. We will show that for sufficiently large k, it holds that

$$QCC(f^k, (1 - \epsilon_{\text{sh}})k, 1 - 2^{-\epsilon_{\text{sh}} k}) \leq O(k \cdot (I + 2)) + o(k)$$

from which the theorem follows from Theorem 1.5.6.

By definition, for all μ, there exists a protocol Π_μ for f s.t. $QIC(\Pi_\mu, \mu) \leq I + 1$ and error $\leq \eta$ w.r.t μ. By lemma 6.4.2, for sufficiently large k, there exists a protocol $\Pi_{k,\mu,\epsilon'}$ s.t. $QCC(\Pi_{k,\mu,\epsilon'}) \leq k(I + 2)$ and

$$\Pr[\Pi_{k,\mu,\epsilon'} \text{ computes} \geq (1 - 2\eta)k \text{ coordinates of } f^k \text{ correctly}] \geq 1 - e^{-2\eta^2 k} - \epsilon'$$

Here the probability is over the distribution μ^k and the randomness of the protocol. Choose k large enough and ϵ' small enough so that $1 - e^{-2\eta^2 k} - \epsilon' \geq 0.9$. Then by lemma 6.4.3, for any integer $c > 0$, any constant $\epsilon > 0$, there exists a protocol τ s.t.

$$\begin{aligned} &\Pr[\tau \text{ computes} \geq (1 - 2^{-c/2} - 2c\eta)k \text{ coordinates correctly}] \\ &\geq \frac{1}{2}\left(\left(\frac{0.9}{(1+\epsilon)^k}\right)^c - 2^{-2^{2-2c}k}\right) \end{aligned}$$

This holds for any input $(x_1, \ldots, x_k, y_1, \ldots, y_k)$ and the probability is only over the random-

ness of the protocol. Also $QCC(\tau) \leq c \cdot k \cdot (I+2) + o(k)$. Choose $c = \lceil 2 \log\left(\frac{2}{\epsilon_{\text{sh}}}\right) \rceil$. Also choose $\eta = \frac{\epsilon_{\text{sh}}}{4c}$. Then

$$1 - 2^{-c/2} - 2c\eta \geq 1 - \epsilon_{\text{sh}}$$

Since $2^{2x} \geq 1 + x$ for all $x > 0$, it follows that

$$\left(\frac{0.9}{(1+\epsilon)^k}\right)^c \geq 0.9^c \cdot 2^{-2 \cdot \epsilon \cdot c \cdot k} \geq 2^{-(2\epsilon k+1) \cdot c} \geq 2^{-4 \cdot \epsilon \cdot c \cdot k}$$

The last inequality is true for sufficiently large k. Now choose $\epsilon = \epsilon_{\text{sh}}^4/100c$. Then since

$$2^{-2^{2-2c}k} \leq 2^{-\epsilon_{\text{sh}}^4 k/16}$$

we get that

$$\begin{aligned}
\frac{1}{2}\left(\left(\frac{0.9}{(1+\epsilon)^k}\right)^c - 2^{-2^{2-2c}k}\right) &\geq \frac{1}{2}\left(2^{-\epsilon_{\text{sh}}^4 k/25} - 2^{-\epsilon_{\text{sh}}^4 k/16}\right) \\
&\geq 2^{-\epsilon_{\text{sh}}^4 k/16} \\
&\geq 2^{-\epsilon_{\text{sh}} k}
\end{aligned}$$

The second inequality holds for sufficiently large k. Hence $QCC(\tau) \leq c \cdot k \cdot (I+2) + o(k)$ and

$$\Pr[\tau \text{ computes} \geq (1-\epsilon_{\text{sh}})k \text{ coordinates correctly (on any input } (x_1, \ldots, x_k, y_1, \ldots, y_k))] \geq 2^{-\epsilon_{\text{sh}} k}$$

which implies that $QCC(f^k, (1-\epsilon_{\text{sh}})k, 1 - 2^{-\epsilon_{\text{sh}} k}) \leq O(k \cdot (I+2)) + o(k)$. □

Corollary 6.4.8. *For all boolean functions f, $QCC(f, 1/3) \leq 2^{O(QIC(f,1/3)+1)}$.*

Proof. We will use the following folklore result:

$$R(f, 1/3) \leq \left(\frac{1}{\mathrm{disc}(f)}\right)^{O(1)}$$

where $R(f, 1/3)$ is the (public-coin) randomized communication complexity of f with error $1/3$ and $\mathrm{disc}(f) = \min_\mu \mathrm{disc}^\mu(f)$. See, for example, exercise 3.32 in [KN97]. This implies

$$QCC(f, 1/3) \leq R(f, 1/3) \leq \left(\frac{1}{\mathrm{disc}(f)}\right)^{O(1)} \leq 2^{O(GDM_{1/5}(f))} \tag{6.6}$$

Now, by theorem 6.4.7 and theorem 6.3.13, we get that $QIC(f, \eta) \geq \Omega(GDM_{1/5}(f) - O(1))$ for some small constant η. By lemma 6.3.15, we also get that $QIC(f, 1/3) \geq \Omega(GDM_{1/5}(f) - O(1))$, which combined with equation (6.6) completes the proof. □

6.5 From AND to DISJ

In this section, we show that a protocol with low quantum information cost for AND implies a protocol with low quantum information cost for Disjointness

Lemma 6.5.1.

$$\max_\nu QIC(DISJ_n, \nu, 2/n) \leq n \cdot QIC_0^r(AND, 1/n^2) + O(r \cdot \log^5(n)) + o(\sqrt{n}) \tag{6.7}$$

Proof. Let $QIC_0^r(AND, 1/n^2) = I$. Suppose π is a protocol for AND which has error $\leq 1/n^2$ for all inputs and s.t. $\max_{\mu \text{ s.t. } \mu(1,1)=0} QIC(\pi, \mu) \leq I + \delta$, for arbitrary small δ. Using π, we will construct a protocol for $DISJ_n$. The protocol will have low information cost w.r.t. any distribution ν. Suppose τ_k is a quantum protocol for $DISJ_k$ that has worst case error $\leq 1/k^{10}$ and communication cost $O(\sqrt{k}\log(k))$. For example, use the protocol from [AA03]

and amplify the error to $1/k^{10}$. We'll drop the subscript k when it is clear from the context. Consider the protocol π_n described as Protocol 11.

Inputs: $(x, y) \in \{0,1\}^n \times \{0,1\}^n$, $(x, y) \sim \nu$
Goal: check if $DISJ_n(x, y) = 1$ or not.

1. Alice and Bob share a maximally entangled state $\phi_S^{S_A S_B}$ that will serve as shared randomness in order to sample uniformly at random $n/\log^3(n)$ coordinates from $[n]$ (with replacement). Alice has the register S_A and Bob has S_B.
2. On the random coordinates, run τ. Suppose O_A is the output register for Alice and O_B is the output register for Bob. Note that all this can be implemented using unitaries. Also note either $O_A = O_B = 1$ or $O_A = O_B = 0$.
3. If $O_A = O_B = 1$, then run π on each coordinate. If π outputs 1 on any coordinate, then output 0, otherwise output 1. If $O_A = O_B = 0$, Alice and Bob will keep running a dummy protocol (for example keep exchanging a freshly prepared register $|0\rangle$ of dimension same as to be sent in π^n in the corresponding step). In the end they output 0.

Protocol 11: Subsampling Protocol π_n

We'll denote the protocol in which π is run independently on each coordinate by π^n. First lets analyze the error of the protocol π_n. Suppose (x, y) were disjoint. Then probability that we output 0 because of τ is at most $\log^{30}(n)/n^{10} \leq 1/n$. And the probability that we output 0 because of π^n is at most $n/n^2 = 1/n$ because of union bound. So error in this case $\leq 2/n$. If the sets were intersecting, even if we don't output 0 because of τ, we will output 0 because of π^n w.p. at least $1 - 1/n^2$ (because on the intersecting coordinate, $1/n^2$ is the probability of failure). So in both cases, probability of error $\leq 2/n$.

Now lets figure out the information cost of π_n. For running τ, we just bound the information cost by communication cost, which is at most $\sqrt{n}/\sqrt{\log(n)} = o(\sqrt{n})$. The interesting part is what happens after τ. Lets look at the state of Alice and Bob after τ is over. Alice holds the registers A_τ, O_A, S_A, where A_τ is what is left behind with Alice after τ, O_A is Alice's output register for τ and S_A is the entanglement register which acts as shared randomness. Similarly Bob holds B_τ, O_B, S_B. After running i steps of π^n (just before the

$(i+1)^{\text{th}}$ message is transmitted), Alice and Bob hold registers A_{i+1} and B_{i+1} respectively, with C_{i+1} (the register to be sent next) with Alice if i even and with Bob if i odd. Note that the number of rounds of π is r. Then the information cost of step 3 is:

$$
\begin{aligned}
&\frac{1}{2}\cdot\sum_{i=0,i\ even}^{r-1} I(C_{i+1};R|B_{i+1},B_\tau,O_B,S_B)+\frac{1}{2}\cdot\sum_{i=0,i\ odd}^{r-1} I(C_{i+1};R|A_{i+1},A_\tau,O_A,S_A)\\
&\leq\frac{1}{2}\cdot\sum_{i=0,i\ even}^{r-1} I(C_{i+1};R,B_\tau,O_B,S_B|B_{i+1})+\frac{1}{2}\cdot\sum_{i=0,i\ odd}^{r-1} I(C_{i+1};R,A_\tau,O_A,S_A|A_{i+1})\\
&\leq\frac{1}{2}\cdot\sum_{i=0,i\ even}^{r-1} I(C_{i+1};R,B_\tau,O_B,S_B,A_\tau,O_A,S_A|B_{i+1})+\\
&\frac{1}{2}\cdot\sum_{i=0,i\ odd}^{r-1} I(C_{i+1};R,B_\tau,O_B,S_B,A_\tau,O_A,S_A|A_{i+1})\\
&=\frac{1}{2}\cdot\sum_{i=0,i\ even}^{r-1} I(C_{i+1};O_A|B_{i+1})+\frac{1}{2}\cdot\sum_{i=0,i\ even}^{r-1} I(C_{i+1};R,B_\tau,S_B,A_\tau,S_A|B_{i+1},O_A)+\\
&\frac{1}{2}\cdot\sum_{i=0,i\ even}^{r-1} I(C_{i+1};O_B|B_{i+1},R,B_\tau,S_B,A_\tau,S_A,O_A)+\\
&\frac{1}{2}\cdot\sum_{i=0,i\ odd}^{r-1} I(C_{i+1};O_A|A_{i+1})+\sum_{i=0,i\ odd}^{r-1} I(C_{i+1};R,B_\tau,S_B,A_\tau,S_A|A_{i+1},O_A)+\\
&\frac{1}{2}\cdot\sum_{i=0,i\ odd}^{r-1} I(C_{i+1};O_B|A_{i+1},R,B_\tau,S_B,A_\tau,S_A,O_A)\\
&\leq\frac{1}{2}\cdot\sum_{i=0,i\ even}^{r-1} I(C_{i+1};R,B_\tau,S_B,A_\tau,S_A|B_{i+1},O_A)+\\
&\frac{1}{2}\cdot\sum_{i=0,i\ odd}^{r-1} I(C_{i+1};R,B_\tau,S_B,A_\tau,S_A|A_{i+1},O_A)+O(r)\\
&=\frac{1}{2}\cdot\sum_{i=0,i\ even}^{r-1}\Pr[O_A=1]\cdot I(C_{i+1};R,B_\tau,S_B,A_\tau,S_A|B_{i+1},O_A=1)+\\
&\frac{1}{2}\cdot\sum_{i=0,i\ odd}^{r-1}\Pr[O_A=1]\cdot I(C_{i+1};R,B_\tau,S_B,A_\tau,S_A|A_{i+1},O_A=1)+O(r)
\end{aligned}
\tag{6.8}
$$

The first two inequalities are by properties of mutual information. The first equality is just chain rule. Third inequality follows from the fact that O_A, O_B are one dimensional systems. The last equality is true because O_B is just a copy of O_A, so tracing out O_B, O_A becomes a classical system and also conditioned on $O_A = 0$, the mutual information expressions are 0 since in that case the C_{i+1} registers are independent of everything else. Now lets analyze the term:

$$\frac{1}{2} \cdot \sum_{i=0, i \text{ even}}^{r-1} I(C_{i+1}; R, B_\tau, S_B, A_\tau, S_A | B_{i+1}, O_A = 1) + \frac{1}{2} \cdot \sum_{i=0, i \text{ odd}}^{r-1} I(C_{i+1}; R, B_\tau, S_B, A_\tau, S_A | A_{i+1}, O_A = 1)$$

We claim that this is equal to $QIC(\pi^n, \nu')$, where ν' is the distribution $\nu | O_A = 1$. This follows from the following observations:

- Since O_B is just a copy of O_A, for all i, the state of systems $A_{i+1}, B_{i+1}, C_{i+1}, R, B_\tau, S_B, A_\tau, S_A$ conditioned on $O_A = 1$ (the post-measurement state if O_A is measured and the result is 1) is pure.
- For all i, the marginal state of systems $A_{i+1}, B_{i+1}, C_{i+1}$ conditioned on $O_A = 1$ is the same as it would have been if π^n was run starting from the distribution ν'. This is because π^n never touches the registers B_τ, S_B, A_τ, S_A.
- If $|\phi\rangle^{R', A, B, C}$ and $|\psi\rangle^{R, A, B, C}$ are two pure states such that $\mathrm{Tr}_{R'} |\phi\rangle^{R', A, B, C} = \mathrm{Tr}_R |\psi\rangle^{R, A, B, C}$. Then $I(C; R'|B)_\phi = I(C; R|B)_\psi$.

Remark 6.5.2. *The reader might have noticed that the trick of merging stuff with the purification register and then applying the last observation is used at a lot of places in this chapter. This seems to be a very useful trick and seems to replace the classical Proposition 2.9 from [Bra12].*

Putting it all together, we have the following upper bound on information cost of step 3:

$$
\begin{aligned}
&\Pr[O_A = 1] \cdot QIC(\pi^n, \nu') + O(r) \\
&\leq \Pr[O_A = 1] \cdot \left(\sum_{i=1}^{n} QIC(\pi, \nu'_i) \right) + O(r) \\
&\leq \Pr[O_A = 1] \cdot n \cdot QIC\left(\pi, \sum_{i=1}^{n} \nu'_i / n \right) + O(r) \\
&\leq \Pr[O_A = 1] \cdot n \cdot (I + \delta) + O(\Pr[O_A = 1] \cdot n \cdot rH(w)) + O(r) \qquad (6.9)
\end{aligned}
$$

Here ν'_i is the marginal distribution on the i^{th} coordinate and $w = \sum_{i=1}^{n} \nu'_i(1,1)/n$. First inequality is by lemma 6.3.14. Second inequality is just concavity of information cost, lemma 1.6.10. The last inequality follows from corollary 6.3.9. Now we can assume that $\Pr[O_A = 1] \geq 1/n$, otherwise (6.9) is trivially bounded by $O(r)$. Now let us bound w. Suppose (X, Y) are random variables s.t. $(X, Y) \sim \nu$. Also let $N(x, y)$ be the number of intersections in x and y i.e. number of i such that $x_i = y_i = 1$. Then

$$
\begin{aligned}
\Pr[N(X,Y) = d | O_A = 1] &= \frac{\Pr[N(X,Y) = d] \cdot \Pr[O_A = 1 | N(X,Y) = d]}{\Pr[O_A = 1]} \\
&\leq \Pr[N(X,Y) = d] \cdot \Pr[O_A = 1 | N(X,Y) = d] \cdot n \\
&\leq \Pr[N(X,Y) = d] \cdot \left(\left(1 - \frac{d}{n}\right)^{n/\log^3(n)} + \frac{\log^{30}(n)}{n^{10}} \right) \cdot n \\
&\leq e^{-d/\log^3(n)} \cdot n + \frac{\log^{30}(n)}{n^9}
\end{aligned}
$$

The second inequality follows because if there are d intersections, then getting no intersection in $n/\log^3(n)$ uniformly random coordinates is at most the first term. The second term is due to the error of the amplified protocol for disjointness. So for $d \geq 9 \ln(2) \log^4(n)$, $\Pr[N(X,Y) = d | O_A = 1] \leq 1/n^8$. Thus

$$w = \sum_{i=1}^{n} \nu_i'(1,1)/n = \mathbb{E}_{(X,Y)\sim\nu'} N(X,Y)/n \leq O(\log^4(n)/n)$$

Thus we can bound (6.9) as follows:

$$\begin{aligned}
&\Pr[O_A = 1] \cdot n \cdot (I + \delta) + O(\Pr[O_A = 1] \cdot n \cdot rH(w)) + O(r) \\
&\leq n \cdot (I + \delta) + O(n \cdot rH(w)) + O(r) \\
&\leq n \cdot (I + \delta) + O(r \log^5(n))
\end{aligned}$$

Since δ was arbitrary small, this completes the proof.

□

6.6 Proof of the main result

We now put everything together to get a lower bound on $QIC_0^r(AND, 1/3)$.

Lemma 6.6.1. *For all r, it holds that*

$$QIC_0^r(AND, 1/3) \geq \Omega\left(\frac{1}{r \cdot \log^8 r}\right).$$

Proof. We know by theorem 1.5.7 that $GDM_{1/5}(DISJ_n) \geq \Omega(\sqrt{n})$. Hence, by Theorem 6.4.7, we must have that $\max_\mu QIC(DISJ_n, \mu, 2/n) \geq \Omega(\sqrt{n})$. Putting this together with Lemma 6.5.1 and Corollary 6.3.17, and let $r = \Theta\left(\frac{\sqrt{n}}{\log^6 n}\right)$, we have,

$$QIC_0^r(AND, 1/3) = \Omega\left(\frac{1}{\sqrt{n} \cdot \log^2 n}\right) = \Omega\left(\frac{1}{r \cdot \log^8 r}\right).$$

□

Corollary 6.6.2. *Let μ^* be the distribution such that $\mu^*(0,0) = 1/3, \mu^*(0,1) = 1/3, \mu^*(1,0) = 1/3$. Then*

$$\inf_{\Pi \in \mathcal{T}^r(AND,1/3)} QIC(\Pi, \mu^*) = \Omega\left(\frac{1}{r \cdot \log^8 r}\right).$$

Proof. For any distribution μ_0 such that $\mu_0(1,1) = 0$, it is easy to see that μ^* can be written as $\mu^* = \frac{1}{3}\mu_0 + \frac{2}{3}\mu'$ where μ' is some other valid distribution. By Lemma 1.6.10, we have

$$QIC(\Pi, \mu^*) \geq \frac{1}{3}QIC(\Pi, \mu_0) + \frac{2}{3}QIC(\Pi, \mu') \geq \frac{1}{3}QIC(\Pi, \mu_0).$$

Then we have

$$QIC(\Pi, \mu^*) \geq \frac{1}{3} \max_{\mu_0, \mu_0(1,1)=0} QIC(\Pi, \mu_0).$$

Therefore by Lemma 6.6.1, we have

$$\inf_{\Pi \in \mathcal{T}^r(AND,1/3)} QIC(\Pi, \mu^*) \geq \frac{1}{3}QIC_0^r(AND, 1/3) = \Omega\left(\frac{1}{r \cdot \log^8 r}\right).$$

□

Theorem 6.6.3. *For all $r, n \in \mathbb{N}$, $QCC^r(DISJ_n, 1/3) = \Omega\big(\frac{n}{r \cdot \log^8 r}\big)$.*

Proof. Combining Lemma 6.3.20 and Lemma 6.6.1, we get this theorem. □

6.7 Low information protocol for AND

In this section, we exhibit a $\tilde{O}(1/r)$ information $4r$-round protocol for AND (w.r.t. the prior $1/3, 1/3, 1/3, 0$) which computes correctly on all inputs with probability 1. The protocol is due to Jain, Radhakrishnan and Sen. Consider the protocol described in Protocol 12.

Inputs: $(x, y) \in \{0, 1\} \times \{0, 1\}$
Goal: compute $AND(x, y)$

1. Set $\theta = \frac{\pi}{8r}$. Let $|v\rangle$ be the vector $\cos(\theta)|0\rangle + \sin(\theta)|1\rangle$. Let U_v be the unitary operation of reflecting about the vector $|v\rangle$ i.e. $U_v|0\rangle = \cos(2\theta)|0\rangle + \sin(2\theta)|1\rangle$ and $U_v|1\rangle = \sin(2\theta)|0\rangle - \cos(2\theta)|1\rangle$. Also let Z be the unitary operation of reflecting about $|0\rangle$ i.e. $Z|0\rangle = |0\rangle$ and $Z|1\rangle = -|1\rangle$.

2. Alice starts by preparing a qubit C in state $|0\rangle$.

3. If $x = 0$, Alice applies the identity operation on C and sends it to Bob. If $x = 1$, Alice applies the U_v operation on C and sends it to Bob.

4. If $y = 0$, Bob applies the identity operation on C and sends it to Alice. If $y = 1$, Bob applies the Z operation on C and sends it to Alice.

5. After $4r - 1$ rounds, Bob measures the register C. If the result is 1, then he answers 1, otherwise 0. He also sends this to Alice.

Protocol 12: Protocol for AND

First let us see why it computes AND. Let $|\psi_i^{x,y}\rangle = \cos(\phi_i^{x,y})|0\rangle + \sin(\phi_i^{x,y})|1\rangle$ be the state of qubit C after i rounds when the input is (x, y). If the input is $0, 0$, $\phi_i^{0,0}$ is always 0. Also when the input is $0, 1$, $\phi_i^{0,1}$ is always 0. So $|\psi_i^{0,0}\rangle = |\psi_i^{0,1}\rangle = |0\rangle$ always. When the input is $1, 0$, $\phi_i^{1,0}$ follows the trajectory $2\theta \to 2\theta \to 0 \to 0 \to 2\theta \to \cdots$. So $|\psi_{4r-1}^{1,0}\rangle = |0\rangle$ as well. When the input is $1, 1$, $\phi_i^{1,1}$ follows the trajectory $2\theta \to -2\theta \to 4\theta \to -4\theta \to \cdots \to -\pi/2$. So $|\psi_{4r-1}^{1,1}\rangle = -|1\rangle$. Thus the players compute AND correctly.

Now let us analyze the information cost of this protocol. Note that after i rounds the full state can be written as follows:

$$|\psi_i\rangle^{XYCR} = \sum_{x,y \text{ s.t. } x \wedge y = 0} \frac{1}{\sqrt{3}} |x\rangle^X |y\rangle^Y |\psi_i^{x,y}\rangle^C |x, y\rangle^R$$

Then information cost is given by:

$$\frac{1}{2}\cdot\sum_{i=1,odd}^{4r-1} I(C;R|Y)_{\psi_i} + \frac{1}{2}\cdot\sum_{i=1,even}^{4r-1} I(C;R|X)_{\psi_i}$$

Let us look at a particular term:

$$\begin{aligned}
I(C;R|Y)_{\psi_i} &= H(C,Y)_{\psi_i} + H(R,Y)_{\psi_i} - H(C,R,Y)_{\psi_i} - H(Y)_{\psi_i} \\
&= H(C,Y)_{\psi_i} + H(C,X)_{\psi_i} - H(X)_{\psi_i} - H(Y)_{\psi_i} \\
&= H(C|Y)_{\psi_i} + H(C|X)_{\psi_i} \\
&= \frac{2}{3}H(C|Y=0)_{\psi_i} + \frac{1}{3}H(C|Y=1)_{\psi_i} + \frac{2}{3}H(C|X=0)_{\psi_i} + \frac{1}{3}H(C|X=1)_{\psi_i} \\
&= \frac{2}{3}H(C|Y=0)_{\psi_i}
\end{aligned}$$

First equality is by definition. For second equality, we are using the fact that for a pure state on some systems A, B, $H(A) = H(B)$. Third equality is again by definition. For fourth equality, we use the fact that if we trace out R, X, Y become classical. For the fifth equality, we use the fact that conditioned on $Y = 1$, system C is in a pure state, namely $\left|\psi_i^{0,1}\right\rangle$. Similarly conditioned on $X = 1$, it is in state $\left|\psi_i^{1,0}\right\rangle$. Conditioned on $X = 0$, C is in the state $|0\rangle$. Now conditioned on $Y = 0$, C is in the state:

$$\frac{1}{2}\left|\psi_i^{0,0}\right\rangle\left\langle\psi_i^{0,0}\right| + \frac{1}{2}\left|\psi_i^{1,0}\right\rangle\left\langle\psi_i^{1,0}\right|$$

This is $|0\rangle$ if $i \equiv 3 (\text{mod } 4)$ and if $i \equiv 1 (\text{mod } 4)$, the density matrix is given by:

$$\rho = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}\cos^2(2\theta) & \frac{1}{2}\cos(2\theta)\sin(2\theta) \\ \frac{1}{2}\cos(2\theta)\sin(2\theta) & \frac{1}{2}\sin^2(2\theta) \end{bmatrix}$$

Eigenvalue computation shows that $H(\rho) = H(\sin^2(\theta)) = O(\theta^2 \log(1/\theta)) = O(\log(r)/r^2)$. So some of Alice's terms are 0 and some are $O(\log(r)/r^2)$. Similarly some of Bob's terms are 0 and some are $O(\log(r)/r^2)$. So in total we get that the information cost is $O(\log(r)/r)$. Note that from the protocol it might seem that since the roles of Alice and Bob are asymmetric, only Alice is sending information and Bob is not. However this definition of quantum information cost also accounts for sending back information in some sense. For example, in some of the rounds, Alice is sending Bob some information but Bob is sending it back, so that is accounted for. This results in Bob's part of the cost to be non-zero and in fact equal to that of Alice.

Now let us see what happens if we place a small mass w on $(1,1)$ entry. Then the full state can be described as follows:

$$|\psi_i\rangle^{XYCR} = \sum_{x,y \text{ s.t. } x \wedge y = 0} \sqrt{\frac{1-w}{3}} |x\rangle^X |y\rangle^Y |\psi_i^{x,y}\rangle^C |x,y\rangle^R + \sqrt{w} |1\rangle^X |1\rangle^Y |\psi_i^{1,1}\rangle^C |1,1\rangle^R$$

The i^{th} term of the information cost as before is given by:

$$\begin{aligned} &\frac{2(1-w)}{3} H(C|Y=0)_{\psi_i} + \frac{1+2w}{3} H(C|Y=1)_{\psi_i} + \frac{2(1-w)}{3} H(C|X=0)_{\psi_i} \\ &+ \frac{1+2w}{3} H(C|X=1)_{\psi_i} \\ &= \frac{2(1-w)}{3} H(C|Y=0)_{\psi_i} + \frac{1+2w}{3} H(C|Y=1)_{\psi_i} + \frac{1+2w}{3} H(C|X=1)_{\psi_i} \end{aligned}$$

As before $H(C|X=0)_{\psi_i} = 0$. But the other three terms are non-zero. $H(C|Y=0)_{\psi_i}$ is the

same as before. Let us focus on $H(C|Y=1)_{\psi_i}$. State of C conditioned on $Y=1$ is given by:

$$\frac{1-w}{1+2w}\left|\psi_i^{0,1}\right\rangle\left\langle\psi_i^{0,1}\right| + \frac{3w}{1+2w}\left|\psi_i^{1,1}\right\rangle\left\langle\psi_i^{1,1}\right|$$

For i odd, the density matrix is given by:

$$\rho = \begin{bmatrix} \frac{1-w}{1+2w} + \frac{3w}{1+2w}\cos^2((i+1)\theta) & \frac{3w}{1+2w}\cos((i+1)\theta)\sin((i+1)\theta) \\ \frac{3w}{1+2w}\cos((i+1)\theta)\sin((i+1)\theta) & \frac{3w}{1+2w}\sin^2((i+1)\theta) \end{bmatrix}$$

Eigenvalue computation shows that

$$H(\rho) = H\left(\frac{1-\sqrt{1-\frac{12w(1-w)\sin^2((i+1)\theta)}{(1+2w)^2}}}{2}\right)$$

Now assuming $w \leq 1/6$ and considering i such that $\sin^2((i+1)\theta) \geq 4/5$, we get that

$$\begin{aligned} \frac{1-\sqrt{1-\frac{12w(1-w)\sin^2((i+1)\theta)}{(1+2w)^2}}}{2} &\geq \frac{1-\sqrt{1-\frac{8w}{(1+2w)^2}}}{2} \\ &= \frac{1-\frac{1-2w}{1+2w}}{2} \\ &= \frac{2w}{1+2w} \end{aligned}$$

Since other terms involving w either have positive contribution or are of lower order, we get that for a constant fraction of the rounds, the information cost term increases by an additive $\Omega(H(w))$. And hence overall the increase in information cost is at least $\Omega(rH(w))$.

Chapter 7

Communication Complexity Lower Bounds for Statistical Estimation

The results in this chapter are based on joint work with Mark Braverman, Tengyu Ma, Huy Nguyen and David Woodruff [BGM+16]. Results in [BGM+16] subsume the results in [GMN14]. So we only include the results in the paper [BGM+16].

7.1 Introduction

Rapid growth in the size of modern data sets has fueled a lot of interest in solving statistical and machine learning tasks in a distributed environment using multiple machines. Communication between the machines has emerged as an important resource and sometimes the main bottleneck. A lot of recent work has been devoted to design communication-efficient learning algorithms [DAW12, ZDW13, ZX15, KVW14, LBKW14, SSZ14, LSLT15].

In this paper we consider statistical estimation problems in the distributed setting, which can be formalized as follows. There is a family of distributions $\mathcal{P} = \{\mu_\theta : \theta \in \Omega \subset \mathbb{R}^d\}$ that is parameterized by $\theta \in \mathbb{R}^d$. Each of the m machines is given n i.i.d samples drawn from

an unknown distribution $\mu_\theta \in \mathcal{P}$. The machines communicate with each other by message passing, and do computation on their local samples and the messages that they receives from others. Finally one of the machines needs to output an estimator $\hat{\theta}$ and the statistical error is usually measured by the mean-squared loss $\mathbb{E}[\|\hat{\theta} - \theta\|^2]$. We count the communication between the machines in bits.

This paper focuses on understanding the fundamental tradeoff between communication and the statistical error for high-dimensional statistical estimation problems. Modern large datasets are often equipped with a high-dimensional statistical model, while communication of high dimensional vectors could potentially be expensive. It has been shown in [DJWZ14] and [GMN14] that for the linear regression problem, the communication cost must scale with the dimensionality for achieving optimal statistical minimax error – not surprisingly, the machines have to communicate high-dimensional vectors in order to estimate high-dimensional parameters.

These negative results naturally lead to the interest in high-dimensional estimation problems with additional sparse structure on the parameter θ. It has been well understood that the statistical minimax error typically depends on the intrinsic dimension, that is, the sparsity of the parameters, instead of the ambient dimension[1]. Thus it is natural to expect that the same phenomenon also happens for communication.

However, this paper disproves this possibility in the interactive communication model by proving that for the *sparse Gaussian mean estimation* problem (where one estimates the mean of a Gaussian distribution which is promised to be sparse, see Section 7.2 for the formal definition), in order to achieve the statistical minimax error, the communication must scale with the ambient dimension. On the other end of the spectrum, if alternatively the communication only scales with the sparsity, then the statistical error must scale with the ambient dimension (see Theorem 7.4.6). Shamir [Sha14] establishes the same result for the

[1] the dependency on the ambient dimension is typically logarithmic.

1-sparse case under a non-iterative communication model.

Our lower bounds for the Gaussian mean estimation problem imply lower bounds for the *sparse linear regression* problem (Corollary 7.4.9) via the reduction of [ZDJW13]: for a Gaussian design matrix, to achieve the statistical minimax error, the communication cost per machine needs to be $\Omega(\min\{n, d\})$ where d is the ambient dimension and n is the dimension of the observation that each machine receives. This lower bound matches the upper bound in [LSLT15] when n is larger than d. When n is less than d, we note that it is not clear whether $O(n)$ or $O(d)$ should be the minimum communication cost per machine needed. In any case, our contribution here is in proving a lower bound that does not depend on the sparsity. Compared to previous work of Steinhardt and Duchi [SD15], which proves the same lower bounds for a memory-bounded model, our results work for a stronger communication model where multi-round iterative communication is allowed. Moreover, our techniques are possibly simpler and potentially easier to adapt to related problems. For example, we show that the result of Woodruff and Zhang [WZ12] on the information complexity of distributed gap majority can be reproduced by our technique with a cleaner proof (see Theorem 7.8.1).

We complement our lower bounds for this problem in the dense case by providing a new simultaneous protocol, improving the number of rounds of the previous communication-optimal protocol from $O(\log m)$ to 1 (see Theorem 7.4.7). Our protocol is based on a certain combination of many bits from a few Gaussian samples, together with roundings (to a single bit) of the fractional parts of many Gaussian samples.

Our proof techniques are potentially useful for other questions along these lines. We first use a modification of the direct-sum result of [GMN14], which is tailored towards sparse problems, to reduce the estimation problem to a detection problem. Then we prove what we call a *distributed data processing inequality* for bounding from below the cost of the detection problem. The latter is the crux of our proofs. We elaborate more on it in the next subsection.

7.1.1 Distributed Data Processing Inequality

We consider the following distributed detection problem. As we will show in Section 7.4 (by a direct-sum theorem), it suffices to prove a tight lower bound in this setting, in order to prove a lower bound on the communication cost for the sparse linear regression problem.

Distributed detection problem: We have a family of distributions $\mathcal{P}$ that consist of only two distributions $\{\mu_0, \mu_1\}$, and the parameter space $\Omega = \{0, 1\}$. To facilitate the use of tools from information theory, sometimes it is useful to introduce a prior over the parameter space. Let $V \sim B_q$ be a Bernoulli random variable with probability q of being 1. Given $v \in \{0, 1\}$, we draw i.i.d. samples $X_1, \ldots, X_m$ from μ_v and the j-th machine receives one sample X_j, for $j = 1, \ldots, m$. We use $\Pi \in \{0, 1\}^*$ to denote the sequences of messages that are communicated by the machines. We will refer to Π as a "transcript", and the distributed algorithm that the machines execute as a "protocol".

The final goal of the machines is to output an estimator for the hidden parameter v which is as accurate as possible. We formalize the estimator as a (random) function $\hat{v} : \{0, 1\}^* \to \{0, 1\}$ that takes the transcript Π as input. We require that given $V = v$, the estimator is correct with probability at least $3/4$, that is, $\min_{v \in \{0,1\}} \Pr[\hat{v}(\Pi) = v \mid V = v] \geq 3/4$. When $q = 1/2$, this is essentially equivalent to the statement that the transcript Π carries $\Omega(1)$ information about the random variable V. Therefore, the mutual information $\mathrm{I}(V; \Pi)$ is also used as a convenient measure for the quality of the protocol when $q = 1/2$.

Strong data processing inequality: The mutual information viewpoint of the accuracy naturally leads us to the following approach for studying the simple case when $m = 1$ and $q = 1/2$. When $m = 1$, we note that the parameter V, data X, and transcript Π form a simple Markov chain $V \to X \to \Pi$. The channel $V \to X$ is defined as $X \sim \mu_v$, conditioned on $V = v$. The strong data processing inequality (SDPI) captures the relative ratio between $\mathrm{I}(V; \Pi)$ and $\mathrm{I}(X; \Pi)$.

Definition 7.1.1 (Special case of SDPI)**.** Let $V \sim B_{1/2}$ and the channel $V \to X$ be defined as above. Then there exists a constant $\beta \leq 1$ that depends on μ_0 and μ_1, such that for any Π that depends only on X (that is, $V \to X \to \Pi$ forms a Markov Chain), we have

$$\mathrm{I}(V;\Pi) \leq \beta \cdot \mathrm{I}(X;\Pi). \tag{7.1}$$

An inequality of this type is typically referred to as a *strong data processing inequality for mutual information* when $\beta < 1$ [2]. Let $\beta(\mu_0, \mu_1)$ be the infimum over all possible β such that (7.1) is true, which we refer to as the **SDPI constant**.

Observe that the LHS of (7.1) measures how much information Π carries about V, which is closely related to the accuracy of the protocol. The RHS of (7.1) is a lower bound on the expected length of Π, that is, the expected communication cost. Therefore the inequality relates two quantities that we are interested in - the statistical quality of the protocol and the communication cost of the protocol. Concretely, when $q = 1/2$, in order to recover V from Π, we need that $\mathrm{I}(V;\Pi) \geq \Omega(1)$, and therefore inequality (7.1) gives that $\mathrm{I}(X;\Pi) \geq \Omega(\beta^{-1})$. Then it follows from Shannon's source coding theorem that the expected length of Π (denoted by $|\Pi|$) is bounded from below by $\mathbb{E}[|\Pi|] \geq \Omega(\beta^{-1})$. We refer to [Rag16] for a thorough survey of SDPI.

In the multiple machine setting, Duchi et al. [DJWZ14] links the distributed detection problem with SDPI by showing from scratch that for any m, when $q = 1/2$, if β is such that $(1 - \sqrt{\beta})\mu_1 \leq \mu_0 \leq (1 + \sqrt{\beta})\mu_1$, then

$$\mathrm{I}(V;\Pi) \leq \beta \cdot \mathrm{I}(X_1 \dots X_m;\Pi).$$

This results in the bounds for the Gaussian mean estimation problem and the linear regres-

[2] Inequality (7.1) is always true for a Markov chain $V \to X \to \Pi$ with $\beta = 1$ and this is called the data processing inequality.

sion problem. The main limitation of this inequality is that it requires the prior B_q to be unbiased (or close to unbiased). For our target application of high-dimensional problems with sparsity structures, like sparse linear regression, in order to apply this inequality we need to put a very biased prior B_q on V. The proof technique of [DJWZ14] seems also hard to extend to this case with a tight bound[3]. Moreover, the relation between β, μ_0 and μ_1 may not be necessary (or optimal), and indeed for the Gaussian mean estimation problem, the inequality is only tight up to a logarithmic factor, while potentially in other situations the gap is even larger.

Our approach is essentially a prior-free multi-machine SDPI, which has the same SDPI constant β as is required for the single machine one. We prove that, as long as the SDPI (7.1) for a single machine is true with parameter β, and $\mu_0 \le O(1)\mu_1$, then the following prior-free multi-machine SDPI is true with the same constant β (up to a constant factor).

Theorem 7.1.2 (Distributed SDPI)**.** *Suppose $\frac{1}{c} \cdot \mu_0 \le \mu_1 \le c\mu_0$ for some constant $c \ge 1$, and let $\beta(\mu_0, \mu_1)$ be the SDPI constant defined in Definition 7.1.1. Then in the distributed detection problem, we have the following distributed strong data processing inequality,*

$$\mathrm{h}^2(\Pi|_{V=0}, \Pi|_{V=1}) \le Kc\beta(\mu_0, \mu_1) \cdot \min\{\mathrm{I}(X_1 \dots X_m; \Pi \mid V = 0), \mathrm{I}(X_1 \dots X_m; \Pi \mid V = 1)\} \tag{7.2}$$

where K is a universal constant, and $\mathrm{h}(\cdot, \cdot)$ is the Hellinger distance between two distributions and $\Pi|_{V=v}$ denotes the distribution of Π conditioned on $V = v$.

As an immediate consequence, we obtain a lower bound on the communication cost for the distributed detection problem.

Corollary 7.1.3. *Suppose the protocol and estimator $(\Pi, \hat{v})$ are such that for any $v \in \{0, 1\}$, given $V = v$, the estimator $\hat{v}$ (that takes Π as input) can recover v with probability $3/4$.*

[3]We note, though, that it seems possible to extend the proof to the situation where there is only one-round of communication.

Then

$$\max_{v\in\{0,1\}} \mathbb{E}[|\Pi| \mid V = v] \geq \Omega(\beta^{-1}).$$

Our theorem suggests that to bound the communication cost of the multi-machine setting from below, one could simply work in the single machine setting and obtain the right SDPI constant β. Then, a lower bound of $\Omega(\beta^{-1})$ for the multi-machine setting immediately follows. In other words, multi-machines need to communicate a lot to fully exploit the m data points they receive (1 on each single machine) regardless of however complicated their multi-round protocol is.

Organization of the chapter: Section 7.2 formally sets up our model and problems and introduces some preliminaries. Then we prove our main theorem in Section 7.3. In Section 7.4 we state the main applications of our theory to the sparse Gaussian mean estimation problem and to the sparse linear regression problem. The next three sections are devoted to the proofs of results in Section 7.4. In Section 7.5, we prove Theorem 7.4.5. In Section 7.6 we provide tools for proving single machine strong data processing inequality and prove Theorem 7.4.1. In Section 7.7 we present our matching upper bound in the simultaneous communication model. In section 7.8 we give a simple proof of distributed gap majority problems using our machinery.

7.2 Problem Setup, Notations and Preliminaries

7.2.1 Distributed Protocols and Parameter Estimation Problems

Let $\mathcal{P} = \{\mu_\theta : \theta \in \Omega\}$ be a family of distributions over some space $\mathcal{X}$, and $\Omega \subset \mathbb{R}^d$ be the space of all possible parameters. There is an unknown distribution $\mu_\theta \in \mathcal{P}$, and our goal is to estimate a parameter θ using m machines. Machine j receives n i.i.d samples $X_j^{(1)}, \ldots, X_j^{(n)}$ from distribution μ_θ. For simplicity we will use X_j as a shorthand for all

the samples machine j receives, that is, $X_j = (X_j^{(1)}, \ldots, X_j^{(n)})$. Therefore $X_j \sim \mu_\theta^n$, where μ^n denotes the product of n copies of μ. When it is clear from context, we will use X as a shorthand for $(X_1, \ldots, X_m)$. We define the problem of estimating parameter θ in this distributed setting formally as task $T(n, m, \mathcal{P})$. When $\Omega = \{0, 1\}$, we call this a detection problem and refer it to as $T_{det}(n, m, \mathcal{P})$.

The machines communicate via a publicly shown blackboard. That is, when a machine writes a message on the blackboard, all other machines can see the content. The messages that are written on the blackboard are counted as communication between the machines. Note that this model captures both point-to-point communication as well as broadcast communication. Therefore, our lower bounds in this model apply to both the message passing setting and the broadcast setting.

We denote the public randomness and the collection of all the messages written on the blackboard by Π. We will refer to Π as the transcript and note that $\Pi \in \{0, 1\}^*$ is written in bits and the communication cost is defined as the length of Π, denoted by $|\Pi|$. We will call the algorithm that the machines follow to produce Π a protocol. With a slight abuse of notation, we use Π to denote both the protocol and the transcript produced by the protocol.

One of the machines needs to estimate the value of θ using an estimator $\hat{\theta} : \{0, 1\}^* \to \mathbb{R}^d$ which takes Π as input. The accuracy of the estimator on θ is measured by the mean-squared loss:

$$R((\Pi, \hat{\theta}), \theta) = \mathbb{E}\left[\|\hat{\theta}(\Pi) - \theta\|_2^2\right],$$

where the expectation is taken over the randomness of the data X, and the randomness of the protocol. The error of the estimator is the supremum of the loss over all θ,

$$R(\Pi, \hat{\theta}) = \sup_{\theta \in \Omega} \mathbb{E}\left[\|\hat{\theta}(\Pi) - \theta\|_2^2\right]. \tag{7.3}$$

The communication cost of a protocol is measured by the expected length of the transcript

Π, that is, $\mathrm{CC}(\Pi) = \sup_{\theta\in\Omega} \mathbb{E}[|\Pi|]$. The information cost IC of a protocol is defined as the mutual information between transcript Π and the data X,

$$\mathrm{IC}(\Pi) = \sup_{\theta\in\Omega} \mathrm{I}_\theta(\Pi; X \mid R_{\mathrm{pub}}) \tag{7.4}$$

where R_{pub} denotes the random variable for the public coins used by the protocol and $I_\theta(\Pi; X \mid R_{\mathrm{pub}})$ denotes the mutual information between random variable X and Π (conditioned on R_{pub}) when the data X is drawn from distribution μ_θ. We will drop the subscript θ when it is clear from context. [4]

For the detection problem, we need to define minimum information cost, a stronger version of information cost

$$\text{min-IC}(\Pi) = \min_{v\in\{0,1\}} \mathrm{I}_v(\Pi; X \mid R_{\mathrm{pub}}) \tag{7.5}$$

Definition 7.2.1. We say that a protocol and estimator pair $(\Pi, \hat{\theta})$ solves the distributed estimation problem $T(m, n, d, \Omega, \mathcal{P})$ with information cost I, communication cost C, and mean-squared loss R if $\mathrm{IC}(\Pi) \le I$, $\mathrm{CC}(\Pi) \le C$ and $R(\Pi, \hat{\theta}) \le R$.

When $\Omega = \{0, 1\}$, we have a detection problem, and we typically use v to denote the parameter and $\hat{v}$ as the (discrete) estimator for it. We define the communication and information cost the same as above, while defining the error in a more meaningful and convenient way,

$$R_{det}(\Pi, \hat{v}) = \max_{v\in\{0,1\}} \Pr[\hat{v}(\Pi) \neq v \mid V = v]$$

Definition 7.2.2. We say that a protocol and estimator pair $(\Pi, \hat{v})$ solves the distributed detection problem $T_{det}(m, n, d, \Omega, \mathcal{P})$ with information cost I, if $\mathrm{IC}(\Pi) \le I$, $R_{det}(\Pi, \hat{v}) \le 1/4$.

[4] Note that because of the convention that public randomness is part of Π, $I(\Pi; X|R_{\mathrm{pub}}) = I(\Pi; X)$, since R_{pub} is independent of X.

Now we formally define the concrete questions that we are concerned with.

Distributed Gaussian detection problem: We call the problem with $\Omega = \{0, 1\}$ and $\mathcal{P} = \{\mathcal{N}(0, \sigma^2)^n, \mathcal{N}(\delta, \sigma^2)^n\}$ the Gaussian mean detection problem, denoted by $\mathrm{GD}(n, m, \delta, \sigma^2)$.

Distributed (sparse) Gaussian mean estimation problem: The distributed statistical estimation problem defined by $\Omega = \mathbb{R}^d$ and $\mathcal{P} = \{\mathcal{N}(\theta, \sigma^2 I_{d\times d}) : \theta \in \Omega\}$ is called the distributed Gaussian mean estimation problem, abbreviated $\mathrm{GME}(n, m, d, \sigma^2)$. When $\Omega = \{\theta \in \mathbb{R}^d : |\theta|_0 \leq k\}$, the corresponding problem is referred to as distributed sparse Gaussian mean estimation, abbreviated $\mathrm{SGME}(n, m, d, k, \sigma^2)$.

Distributed sparse linear regression: For simplicity and the purpose of lower bounds, we only consider sparse linear regression with a random design matrix. To fit into our framework, we can also regard the design matrix as part of the data. We have a parameter space $\Omega = \{\theta \in \mathbb{R}^d : |\theta|_0 \leq k\}$. The j-th data point consists of a row of design matrix A_j and the observation $y_j = \langle A_j, \theta\rangle + w_j$ where $w_j \sim \mathcal{N}(0, \sigma^2)$ for $j = 1, \ldots, mn$, and each machine receives n data points among them[5]. Formally, let μ_θ denote the joint distribution of (A_j, y_j) here, and let $\mathcal{P} = \{\mu_\theta : \theta \in \Omega\}$. We use $\mathrm{SLR}(n, m, d, k, \sigma^2)$ as shorthand for this problem.

7.2.2 Hellinger distance and cut-paste property

In this subsection, we introduce Hellinger distance, and the key property of protocols that we exploit here, the so-called "cut-paste" property developed by [BYJKS04] for proving lower bounds for set-disjointness and other problems. We also introduce some notation that will be used later in the proofs.

Definition 7.2.3 (Hellinger distance)**.** Consider two distributions with probability density

[5]We note that here for convenience, we use subscripts for samples, which is different from the notation convention used for previous problems.

functions $f, g : \Omega \to \mathbb{R}$. The square of the Hellinger distance between f and g is defined as

$$\mathrm{h}^2(f,g) := \frac{1}{2} \cdot \int_\Omega \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx$$

A key observation regarding the property of a protocol by [BYJKS04, Lemma 16] is the following: fixing $X_1 = x_1, \ldots, X_m = x_m$, the distribution of $\Pi|_{X=x}$ can be factored in the following form,

$$\Pr[\Pi = \pi \mid X = x] = p_{1,\pi}(x_1) \ldots p_{m,\pi}(x_m) \tag{7.6}$$

where $p_{i,\pi}(\cdot)$ is a function that only depends on i and the entire transcript π . To see this, one could simply write the density of π as a products of density of each messages of the machines and group the terms properly according to machines (and note that $p_{i,\pi}(\cdot)$ is allowed to depend on the entire transcript π).

We extend equation (7.6) to the situation where the inputs are from product distributions. For any vector $\boldsymbol{b} \in \{0,1\}^m$, let $\mu_{\boldsymbol{b}} := \mu_{b_1} \times \cdots \times \mu_{b_m}$ be a distribution over $\mathcal{X}^m$. We denote by $\Pi_{\boldsymbol{b}}$ the distribution of $\Pi(X_1, \ldots, X_m)$ when $(X_1, \ldots, X_m) \sim \mu_{\boldsymbol{b}}$.

Therefore if $X \sim \mu_{\boldsymbol{b}}$, using the fact that $\mu_{\boldsymbol{b}}$ is a product measure, we can marginalize over X and obtain the marginal distribution of Π when $X \sim \mu_{\boldsymbol{b}}$,

$$\Pr_{X \sim \mu_{\boldsymbol{b}}}[\Pi = \pi] = q_{1,\pi}(b_1) \ldots q_{m,\pi}(b_m), \tag{7.7}$$

where $q_{j,\pi}(b_j)$ is the marginalization of $p_{j,\pi}(x)$ over $x \sim \mu_{b_j}$, that is, $q_{j,\pi}(b_j) = \int_x p_{j,\pi}(x) d\mu_{b_j}$.

Let $\Pi_{\boldsymbol{b}}$ denote the distribution of Π when $X \sim \mu_{\boldsymbol{b}}$. Then by the decomposition (7.7) of $\Pi_{\boldsymbol{b}}(\pi)$ above, we have the following cut-paste property for $\Pi_{\boldsymbol{b}}$ which will be the key property of a protocol that we exploit.

Proposition 7.2.4 (Cut-paste property of a protocol)**.** *For any $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{c}, \boldsymbol{d}$ with $\{a_i, b_i\} =$*

$\{c_i, d_i\}$ *(in a multi-set sense) for every* $i \in [m]$,

$$\Pi_{\boldsymbol{a}}(\pi) \cdot \Pi_{\boldsymbol{b}}(\pi) = \Pi_{\boldsymbol{c}}(\pi) \cdot \Pi_{\boldsymbol{d}}(\pi) \tag{7.8}$$

and therefore,

$$\mathrm{h}^2(\Pi_{\boldsymbol{a}}, \Pi_{\boldsymbol{b}}) = \mathrm{h}^2(\Pi_{\boldsymbol{c}}, \Pi_{\boldsymbol{d}}) \tag{7.9}$$

Lemma 7.2.5 (Hellinger v.s. total variation)**.** *For any two distribution* P, Q*, we have*

$$\mathrm{h}^2(P, Q) \leq \|P - Q\|_{TV} \leq \sqrt{2}\mathrm{h}(P, Q)$$

Lemma 7.2.6. *Let* $\phi(z_1)$ *and* $\phi(z_2)$ *be two random variables. Let* Z *denote a random variable with uniform distribution in* $\{z_1, z_2\}$*: Suppose* $\phi(z)$ *is independent of* Z *for each* $z \in \{z_1, z_2\}$*: Then,*

$$2\mathrm{h}^2(\phi_{z_1}, \phi_{z_2}) \geq \mathrm{I}(Z; \phi(Z)) \geq \mathrm{h}^2(\phi_{z_1}, \phi_{z_2})$$

Proof. The lower bound of the mutual information follows from Lemma 6.2 of [BYJKS04]. For the upper bound, we assume that for simplicity ϕ has discrete support $\mathcal{X}$, though the proof extends continuous random variable directly. We have

$$\begin{aligned}
\mathrm{I}(Z; \phi(Z)) &= \frac{1}{2}\mathrm{D}_{\mathrm{kl}}(\phi_1 \| (\phi_1 + \phi_2)/2) + \frac{1}{2}\mathrm{D}_{\mathrm{kl}}(\phi_2 \| (\phi_1 + \phi_2)/2) \\
&\leq \frac{1}{2}\chi^2(\phi_1 \| (\phi_1 + \phi_2)/2) + \frac{1}{2}\chi^2(\phi_2 \| (\phi_1 + \phi_2)/2) \\
&= \frac{1}{4}\sum_{x \in \mathcal{X}} \frac{(\phi_1(x) - \phi_2(x))^2}{\phi_1(x) + \phi_2(x)} + \frac{1}{4}\sum_{x \in \mathcal{X}} \frac{(\phi_1(x) - \phi_2(x))^2}{\phi_1(x) + \phi_2(x)} \\
&\leq \sum_{x \in \mathcal{X}} \frac{(\phi_1(x) - \phi_2(x))^2}{(\sqrt{\phi_1(x)} + \sqrt{\phi_2(x)})^2} \\
&= 2\mathrm{h}^2(\phi_1, \phi_2)
\end{aligned}$$

where the first inequality uses that KL-divergence is less than χ^2 distance and the second

one uses the inequality $a^2 + b^2 \geq \frac{(a+b)^2}{2}$. □

Theorem 7.2.7 (Corollary of Theorem 7 of [Jay09b])**.** *Suppose a family of distribution* $\{P_{\boldsymbol{b}} : \boldsymbol{b} \in \{0,1\}^m\}$ *satisfies the cut-paste property: for any for any* $\boldsymbol{a}, \boldsymbol{b}$ *and* $\boldsymbol{c}, \boldsymbol{d}$ *with* $\{a_i, b_i\} = \{c_i, d_i\}$ *(in a multi-set sense) for every* $i \in [m]$, $\mathrm{h}^2(\Pi_{\boldsymbol{a}}, \Pi_{\boldsymbol{b}}) = \mathrm{h}^2(\Pi_{\boldsymbol{c}}, \Pi_{\boldsymbol{d}})$. *Then we have*

$$\sum_{i=1}^{m} \mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{e_i}}) \geq \Omega(1) \cdot \mathrm{h}^2(P_{\mathbf{0}}, P_{\mathbf{1}}) \tag{7.10}$$

where $\mathbf{0}$ *and* $\mathbf{1}$ *are all 0's and all 1's vectors respectively, and* $\boldsymbol{e_i}$ *is the unit vector that only takes 1 in the* i*th entry.*

Proof. Theorem 7 of [Jay09b] already proves a stronger version of this theorem for the $m = 2^t$ case. Suppose on the other hand $m = 2^t + \ell$ for $\ell < 2^t$. We divide $[m] = \{1, \ldots, m\}$ into a collection of 2^t subsets $A_1, \ldots, A_{2^t}$, each of which contains at most 2 elements. Let $\boldsymbol{f_i}$ be the indicator vector of the subset A_i. For example, if $A_i = \{p, q\}$, then $\boldsymbol{f_i} = \boldsymbol{e_p} + \boldsymbol{e_q}$. We claim that $\sum_{j \in A_i} \mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{e_j}}) \geq \Omega(1)\mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{f_i}})$. This is trivial when $|A_i| = 1$ and when $A_i = \{p, q\}$, we have that by Cauchy–Schwarz inequality and the cut-paste property

$$\mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{e_p}}) + \mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{e_q}}) \geq \frac{1}{2}\mathrm{h}^2(P_{\boldsymbol{e_p}}, P_{\boldsymbol{e_q}}) = \frac{1}{2}\mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{e_q}+\boldsymbol{e_q}}).$$

Therefore, we can lowerbound LHS as

$$\sum_{i=1}^{m} \mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{e_i}}) \geq \frac{1}{2} \sum_{i=1}^{2^t} \mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{f_i}}).$$

Then applying Theorem 7 of [Jay09b] on the RHS of the inequality above we have

$$\frac{1}{2} \sum_{i=1}^{2^t} \mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{f_i}}) \geq \Omega(1) \cdot \mathrm{h}^2(P_{\mathbf{0}}, P_{\mathbf{1}}),$$

and the theorem follows. □

7.3 Distributed Strong Data Processing Inequalities

In this section we prove our main Theorem 7.1.2. We state a slightly weaker looking version here but in fact it implies Theorem 7.1.2 by symmetry. The same proof also goes through for the case when the RHS is conditioned on $V = 1$.

Theorem 7.3.1. *Suppose $\mu_1 \le c \cdot \mu_0$, and $\beta(\mu_0, \mu_1) = \beta$, we have*

$$\mathrm{h}^2(\Pi|_{V=0}, \Pi|_{V=1}) \le K(c+1)\beta \cdot \mathrm{I}(X; \Pi \mid V = 0)\,. \tag{7.11}$$

where K is an absolute constant.

Note that the RHS of (7.11) naturally tensorizes (by Lemma 7.3.2 that appears below) in the sense that

$$\sum_{i=1}^{m} \mathrm{I}(X_i; \Pi \mid V = 0) \le \mathrm{I}(X; \Pi \mid V = 0), \tag{7.12}$$

since conditioned on $V = 0$, the X_i's are independent. Our main idea consists of the following two steps a) We tensorize the LHS of (7.11) so that the target inequality (7.11) can be written as a sum of m inequalities. b) We prove each of these m inequalities using the single machine SDPI.

To this end, we do the following thought experiment: Suppose W is a random variable that takes value from $\{0, 1\}$ uniformly. Suppose data X' is generated as follows: $X'_j \sim \mu_W$, and for any $j \neq i$, $X'_j \sim \mu_0$. We apply the protocol on the input X', and view the resulting transcript Π' as communication between the i-th machine and the remaining machines. Then we are in the situation of a single machine case, that is, $W \to X'_i \to \Pi'$ forms a Markov Chain. Applying the data processing inequality (7.1), we obtain that

$$\mathrm{I}(W; \Pi') \le \beta \mathrm{I}(X'_i; \Pi')\cdot \tag{7.13}$$

Using Lemma 7.2.6, we can lower bound the LHS of (7.13) by the Hellinger distance and obtain

$$\mathrm{h}^2(\Pi'|_{W=0}, \Pi'|_{W=1}) \leq \beta \cdot \mathrm{I}(X'_i; \Pi')$$

Let $\boldsymbol{e_i} = (0, 0, \ldots, 1, \ldots, 0)$ be the unit vector that only takes 1 in the ith entry, and $\mathbf{0}$ the all zero vector. Using the notation defined in Section 7.2.2, we observe that $\Pi'|_{W=0}$ has distribution $\Pi_{\mathbf{0}}$ while $\Pi'|_{W=1}$ has distribution $\Pi_{\boldsymbol{e}_i}$. Then we can rewrite the equation above as

$$\mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\boldsymbol{e}_i}) \leq \beta \cdot \mathrm{I}(X'_i; \Pi') \tag{7.14}$$

Observe that the RHS of (7.14) is close to the first entry of the LHS of (7.12) since the joint distribution of (X'_1, Π') is not very far from $X, \Pi \mid V = 0$. (The only difference is that X'_1 is drawn from a mixture of μ_0 and μ_1, and note that μ_0 is not too far from μ_1). On the other hand, the sum of LHS of (7.14) over $i \in [m]$ is lower-bounded by the LHS of (7.11). Therefore, we can tensorize equation (7.11) into inequality (7.14) which can be proved by the single machine SDPI.

We formalize the intuition above by the following two lemmas,

Lemma 7.3.2. *Suppose $\mu_1 \leq c \cdot \mu_0$, and $\beta(\mu_0, \mu_1) = \beta$, then*

$$\mathrm{h}^2(\Pi_{\boldsymbol{e_i}}, \Pi_{\mathbf{0}}) \leq \frac{(c+1)\beta}{2} \cdot \mathrm{I}(X_i; \Pi \mid V = 0) \tag{7.15}$$

Lemma 7.3.3. *Let $\mathbf{0}$ be the m-dimensional all 0's vector, and $\mathbf{1}$ the all 1's vector, we have that*

$$\mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{1}}) \leq O(1) \cdot \sum_{i=1}^{m} \mathrm{h}^2(\Pi_{\boldsymbol{e_i}}, \Pi_{\mathbf{0}}) \tag{7.16}$$

Using Lemma 7.3.2 and Lemma 7.3.3, we obtain Theorem 7.3.1 straightforwardly by combining inequalities (7.12), (7.15) and (7.16)[6].

[6]Note that $\Pi_{\mathbf{0}}$ is the same distribution as $\Pi|_{V=0}$ under the notation introduced in Section 7.2.2.

Finally we provide the proof of Lemma 7.3.2. Lemma 7.3.3 is a direct corollary of Theorem 7.2.7 (which is in turn a direct corollary of Theorem 7 of [Jay09b]) and Proposition 7.2.4.

Proof of Lemma 7.3.2. Let W be uniform Bernoulli random variable and define X' and Π' as follows: Conditioned on $W = 0$, $X' \sim \mu_{\mathbf{0}}$ and conditioned on $W = 1$, $X' \sim \mu_{\boldsymbol{e_i}}$. We run protocol on X' and get transcript Π'.

Note that $V \to X' \to \Pi'$ is a Markov chain and so is $V \to X'_i \to \Pi'$. Also by definition, the conditional random variable $X'|V$ has the same distribution as the random variable $X|V$ in Definition 7.1.1. Therefore by Definition 7.1.1, we have that

$$\beta \cdot \mathrm{I}(X'_i; \Pi') \geq \mathrm{I}(V; \Pi'). \tag{7.17}$$

It is known that mutual information can be expressed as the expectation of KL divergence, which in turn is lower-bounded by Hellinger distance. We invoke a technical variant of this argument, Lemma 6.2 of [BYJKS04], restated as Lemma 7.2.6, to lower bound the right hand side. Note that Z in Lemma 7.2.6 corresponds to V here and ϕ_{z_1}, ϕ_{z_2} corresponds to $\Pi_{\boldsymbol{e_i}}$ and $\Pi_{\mathbf{0}}$. Therefore,

$$\mathrm{I}(V; \Pi') \geq \mathrm{h}^2(\Pi_{\boldsymbol{e_i}}, \Pi_{\mathbf{0}}). \tag{7.18}$$

It remains to relate $\mathrm{I}(X'_i; \Pi')$ to $\mathrm{I}(X_i; \Pi \mid V = 0)$. Note that the difference between joint distributions of (X'_i, Π') and $(X_i, \Pi)|_{V=0}$ is that $X'_i \sim \frac{1}{2}(\mu_0 + \mu_1)$ and $X_i|_{V=0} \sim \mu_0$. We claim (by Lemma 1.2.24) that since $\mu_0 \geq \frac{2}{c+1}(\frac{\mu_0+\mu_1}{2})$, we have

$$\mathrm{I}(X_i; \Pi \mid V = 0) \geq \frac{2}{c+1} \cdot \mathrm{I}(X'_i; \Pi'). \tag{7.19}$$

Combining equations (7.17), (7.18) and (7.19), we obtain the desired inequality.

□

7.4 Applications to Parameter Estimation Problems

7.4.1 Warm-up: Distributed Gaussian mean detection

In this section we apply our main technical Theorem 7.3.1 to the situation when $\mu_0 = \mathcal{N}(0, \sigma^2)$ and $\mu_1 = \mathcal{N}(\delta, \sigma^2)$. We are also interested in the case when each machine receives n samples from either μ_0 or μ_1. We will denote the product of n i.i.d copies of μ_v by μ_v^n, for $v \in \{0, 1\}$.

Theorem 7.3.1 requires that a) $\beta = \beta(\mu_0, \mu_1)$ can be calculated/estimated b) the densities of distributions μ_0 and μ_1 are within a constant factor with each other at every point.

Certainly b) is not true for any two Gaussian distributions. To this end, we consider μ_0', μ_1', the truncation of μ_0 and μ_1 on some support $[-\tau, \tau]$, and argue that the probability mass outside $[-\tau, \tau]$ is too small to make a difference.

For a), we use tools provided by Raginsky [Rag16] to estimate the SDPI constant β. [Rag16] proves that Gaussian distributions μ_0 and μ_1 have SDPI constant $\beta(\mu_0, \mu_1) \leq O(\delta^2/\sigma^2)$, and more generally it connects the SDPI constants to transportation inequalities. We use the framework established by [Rag16] and apply it to the truncated Gaussian distributions μ_0' and μ_1'. Our proof essentially uses the fact that $(\mu_0' + \mu_1')/2$ is a log-concacve distribution and therefore it satisfies the log-Sobolev inequality, and equivalently it also satisfies the transportation inequality. The details and connections to concentration of measures are provided in Section 7.6.3.

Theorem 7.4.1. *Let μ_0' and μ_1' be the distributions obtained by truncating μ_0 and μ_1 on support $[-\tau, \tau]$ for some $\tau > 0$. If $\delta \leq \sigma$, we have $\beta(\mu_0', \mu_1') \leq \delta^2/\sigma^2$.*

As a corollary, the SDPI constant between n copies of μ_0' and μ_1' is bounded by $n\delta^2/\sigma^2$.

Corollary 7.4.2. *Let $\tilde{\mu}_0$ and $\tilde{\mu}_1$ be the distributions over $\mathbb{R}^n$ that are obtained by truncating*

μ_0^n and μ_1^n outside the ball $\mathcal{B} = \{x \in \mathbb{R}^n : |x_1 + \cdots + x_n| \leq \tau\}$. Then when $\sqrt{n}\delta \leq \sigma$, we have

$$\beta(\tilde{\mu}_0, \tilde{\mu}_1) \leq n\delta^2/\sigma^2$$

Applying our distributed data processing inequality (Theorem 7.3.1) on $\tilde{\mu}_0$ and $\tilde{\mu}_1$, we obtain directly that to distinguish $\tilde{\mu}_0$ and $\tilde{\mu}_1$ in the distributed setting, $\Omega\left(\frac{\sigma^2}{n\delta^2}\right)$ communication is required. By properly handling the truncation of the support, we can prove that it is also true with the true Gaussian distribution.

Theorem 7.4.3. *Any protocol estimator pair $(\Pi, \hat{v})$ that solves the distributed Gaussian mean detection problem $\mathrm{GD}(n, m, \delta, \sigma^2)$ with $\delta \leq \sigma/\sqrt{n}$ requires communication cost and minimum information cost at least,*

$$\mathbb{E}[|\Pi|] \geq \text{min-IC}(\Pi) \geq \Omega\left(\frac{\sigma^2}{n\delta^2}\right).$$

Remark 7.4.4. *The condition $\delta \leq \sigma/\sqrt{n}$ captures the interesting regime. When $\delta \gg \sigma/\sqrt{n}$, a single machine can even distinguish μ_0 and μ_1 by its local n samples.*

Proof of Theorem 7.4.3. Let $\Pi_{\mathbf{0}}$ and $\Pi_{\mathbf{1}}$ be the distribution of $\Pi|V = 0$ and $\Pi|V = 1$ as defined in Section 7.2.2. Since $\hat{v}$ solves the detection problem, we have that $\|\Pi_{\mathbf{0}} - \Pi_{\mathbf{1}}\|_{\mathrm{TV}} \geq 1/4$. It follows from Lemma 7.2.5 that $\mathrm{h}(\Pi_{\mathbf{0}}, \Pi_{\mathbf{1}}) \geq \Omega(1)$.

We pick a threshold $\tau = 20\sigma$, and let $\mathcal{B} = \{z \in \mathbb{R}^n : |z_1 + \cdots + z_n| \leq \sqrt{n}\tau\}$. Let $F = 1$ denote the event that $X = (X_1, \ldots, X_n) \in \mathcal{B}$, and otherwise $F = 0$. Note that $\Pr[F = 1] \geq 0.95$ and therefore even if we conditioned on the event that $F = 1$, the protocol estimator pair should still be able to recover v with good probability in the sense that

$$\Pr[\hat{v}(\Pi(X)) = v \mid V = v, F = 1] \geq 0.6 \tag{7.20}$$

We run our whole argument conditioning on the event $F = 1$. First note that for any

Markov chain $V \to X \to \Pi$, and any random variable F that only depends on X, the chain $V|_{F=1} \to X|_{F=1} \to \Pi|_{F=1}$ is also a Markov Chain. Second, the channel from V to $X|_{F=1}$ satisfies that random variable $X|_{V=v,F=1}$ has the distribution $\tilde{\mu}_v$ as defined in the statement of Corollary 7.4.2. Note that by Corollary 7.4.2, we have that $\beta(\tilde{\mu}_0, \tilde{\mu}_1) \leq n\delta^2/\sigma^2$. Also note that by the choice of τ and the fact that $\delta \leq O(\sigma/\sqrt{n})$, we have that for any $z \in \mathcal{B}$, $\tilde{\mu}_0(z) \leq O(1) \cdot \tilde{\mu}_1(z)$.

Therefore we are ready to apply Theorem 7.3.1 and conclude that

$$\mathrm{I}(X; \Pi \mid V = 0, F = 1) \geq \Omega(\beta(\tilde{\mu}_0, \tilde{\mu}_1)^{-1}) = \Omega\left(\frac{\sigma^2}{n\delta^2}\right)$$

Note that Π is independent of F conditioned on X and $V = 0$. Therefore we have that

$$\mathrm{I}(X; \Pi \mid V = 0) \geq \mathrm{I}(X; \Pi \mid F, V = 0) \geq \mathrm{I}(X; \Pi | F = 1, V = 0) \Pr[F = 1 \mid V = 0] = \Omega\left(\frac{\sigma^2}{n\delta^2}\right)$$

Note that by construction, it is also true that $\tilde{\mu}_0 \leq O(1)\tilde{\mu}_1$, and therefore if we switch the position of $\tilde{\mu}_0, \tilde{\mu}_1$ and run the argument above we will have

$$\mathrm{I}(X; \Pi \mid V = 1) = \Omega\left(\frac{\sigma^2}{n\delta^2}\right)$$

Hence the proof is complete.

□

7.4.2 Sparse Gaussian mean estimation

In this subsection, we prove our lower bound for the sparse Gaussian mean estimation problem via a variant of the direct-sum theorem of [GMN14] tailored towards sparse mean estimation.

Our general idea is to make the following reduction argument: Given a protocol Π' for d-dimensional k-sparse estimation problem with information cost I and loss R, we can construct a protocol Π' for the detection problem with information cost roughly I/d and loss R/k. The protocol Π' embeds the detection problem into one random coordinate of the d-dimensional problem, prepares fake data on the remaining coordinates, and then runs the protocol Π on the high dimensional problem. It then extracts information about the true data from the corresponding coordinate of the high-dimensional estimator.

The key distinction from the construction of [GMN14] is that here we are not able to show that Π' has small information cost, but only able to show that Π' has a small minimum information cost [7]. This is the reason why in Theorem 7.4.3 we needed to bound the minimum information cost instead of the information cost.

To formalize the intuition, let $\mathcal{P} = \{\mu_0, \mu_1\}$ define the detection problem. Let $\Omega_{d,k,\delta} = \{\theta : \theta \in \{0, \delta\}^d, |\theta|_0 \leq k\}$ and $\mathcal{Q}_{d,k,\delta} = \{\mu_\theta = \mu_{\theta_1/\delta} \times \cdots \times \mu_{\theta_d/\delta} : \theta \in \Omega_{d,k,\delta}\}$. Therefore $\mathcal{Q}$ is a special case of the general k-sparse high-dimensional problem. We have that

Theorem 7.4.5 (Direct-sum for sparse parameters)**.** *Let $d \geq 2k$, and $\mathcal{P}$ and $\mathcal{Q}$ defined as above. If there exists a protocol estimator pair $(\Pi, \hat{\theta})$ that solves the detection task $T(n, m, \mathcal{Q})$ with information cost I and mean-squared loss $R \leq \frac{1}{16}k\delta^2$, then there exists a protocol estimator pair $(\Pi', \hat{v}')$ (shown in Protocol 13 in Section 7.5) that solves the task $T_{det}(n, m, \mathcal{P})$ with minimum information cost $\frac{I}{d-k+1}$.*

The proof of the theorem is deferred to Section 7.5. Combining Theorem 7.4.3 and Theorem 7.4.5, we get the following theorem:

Theorem 7.4.6. *Suppose $d \geq 2k$. Any protocol estimator pair $(\Pi, \hat{v})$ that solves the k-sparse Gaussian mean problem $\mathrm{SGME}(n, m, d, k, \sigma^2)$ with mean-squared loss R and information cost*

[7] This might be inevitable because protocol Π might reveal a lot information for the nonzero coordinate of θ but since there are very few non-zeros, the total information revealed is still not too much.

I and communication cost C satisfy that

$$R \geq \Omega\left(\min\left\{\frac{\sigma^2 k}{n}, \max\left\{\frac{\sigma^2 dk}{nI}, \frac{\sigma^2 k}{nm}\right\}\right\}\right) \geq \Omega\left(\min\left\{\frac{\sigma^2 k}{n}, \max\left\{\frac{\sigma^2 dk}{nC}, \frac{\sigma^2 k}{nm}\right\}\right\}\right). \quad (7.21)$$

Intuitively, to parse equation (7.21), we remark that the term $\frac{\sigma^2 k}{n}$ comes from the fact that any local machine can achieve this error $O(\frac{\sigma^2 k}{n})$ using only its local samples, and the term $\frac{\sigma^2 k}{nm}$ is the minimax error that the machines can achieve with infinite amount of communication. When the target error is between these two quantities, equation (7.21) predicts that the minimum communication C should scale inverse linearly in the error R.

Our theorem gives a tight tradeoff between C and R up to logarithmic factor, since it is known [GMN14] that for any communication budget C, there exists protocol which uses C bits and has error $R \leq O\left(\min\left\{\frac{\sigma^2 k}{n}, \max\left\{\frac{\sigma^2 dk}{nC}, \frac{\sigma^2 k}{nm}\right\}\right\} \cdot \log d\right)$.

As a side product, in the case when $k = d/2$, our lower bound improves previous works [DJWZ14] and [GMN14] by a logarithmic factor, and turns out to match the upper bound in [GMN14] up to a constant factor.

Proof of Theorem 7.4.6. If $R \leq \frac{1}{16}\frac{k\sigma^2}{n}$ then we are done. Otherwise, let $\delta := \sqrt{16R/k} \leq \sigma/\sqrt{n}$. Let $\mu_0 = \mathcal{N}(0, \sigma^2)$ and $\mu_1 = \mathcal{N}(\delta, \sigma^2)$ and $\mathcal{P} = \{\mu_0, \mu_1\}$. Let $\mathcal{Q}_{d,k,\delta} = \{\mu_\theta = \mu_{\theta_1/\delta} \times \cdots \times \mu_{\theta_d/\delta} : \theta \in \Omega_{d,k,\delta}\}$. Then $T(n, m, \mathcal{Q})$ is just a special case of sparse Gaussian mean estimation problem $\mathrm{SGME}(n, m, d, k, \sigma^2)$, and $T(n, m, \mathcal{P})$ is the distributed Gaussian mean detection problem $\mathrm{GD}(n, m, \delta, \sigma^2)$. Therefore, by Theorem 7.4.5, there exists $(\Pi', \hat{v}')$ that solves $\mathrm{GD}(n, m, \delta, \sigma^2)$ with minimum information cost $I' = O(I/d)$. Since $\delta \leq O(\sigma/\sqrt{n})$, by Theorem 7.4.3 we have that $I' \geq \Omega(\sigma^2/(n\delta^2))$. It follows that $I \geq \Omega(d\sigma^2/(n\delta^2)) = \Omega(kd\sigma^2/(nR))$. To derive (7.21), we observe that $\Omega(\sigma^2 k/nm)$ is the minimax lower bound for R, which completes the proof. □

To complement our lower bounds, we also give a new protocol for the Gaussian mean estimation problem achieving communication optimal up to a constant factor in any number

of dimensions in the dense case. Our protocol is a *simultaneous protocol*, whereas the only previous protocol achieving optimal communication requires $\Omega(\log m)$ rounds [GMN14]. This resolves an open question in Remark 2 of [GMN14], improving the trivial protocol in which each player sends its truncated Gaussian to the coordinator by an $O(\log m)$ factor.

Theorem 7.4.7. *For any $0 \leq \alpha \leq 1$, there exists a protocol that uses one round of communication for the Gaussian mean estimation problem* $\mathrm{GME}(n, m, d, \sigma^2)$ *with communication cost $C = \alpha dm$ and mean-squared loss $R = O\left(\frac{\sigma^2 d}{\alpha mn}\right)$.*

The protocol and proof of this theorem are deferred to Section 7.7, though we mention a few aspects here. We first give a protocol under the assumption that $|\theta|_\infty \leq \frac{\sigma}{\sqrt{n}}$. The protocol trivially generalizes to d dimensions so we focus on 1 dimension. The protocol coincides with the first round of the multi-round protocol in [GMN14], yet we can extract all necessary information in only one round, by having each machine send a single bit indicating if its input Gaussian is positive or negative. Since the mean is on the same order as the standard deviation, one can bound the variance and give an estimator based on the Gaussian density function. In Section 7.7.1 the mean of the Gaussian is allowed to be much larger than the variance, and this no longer works. Instead, a few machines send their truncated inputs so the coordinator learns a crude approximation. To refine this approximation, in parallel the remaining machines each send a bit which is 1 with probability $x - \lfloor x \rfloor$, where x is the machine's input Gaussian. This can be viewed as rounding a sample of the "sawtooth wave function" h applied to a Gaussian. For technical reasons each machine needs to send two bits, another which is 1 with probability $(x+1/5) - \lfloor (x+1/5) \rfloor$. We give an estimator based on an analysis using the Fourier series of h.

Sparse Gaussian estimation with signal strength lower bound Our techniques can also be used to study the optimal rate-communication tradeoffs in the presence of a strong signal in the non-zero coordinates, which is sometimes assumed for sparse signals. That is,

suppose the machines are promised that the mean $\theta \in \mathcal{R}^d$ is k-sparse and also if $\theta_i \neq 0$, then $|\theta_i| \geq \eta$, where η is a parameter called the signal strength. We get tight lower bounds for this case as well.

Theorem 7.4.8. *For $d \geq 2k$ and $\eta^2 \geq 16R/k$, any protocol estimator pair $(\Pi, \hat{v})$ that solves the k-sparse Gaussian mean problem $\mathrm{SGME}(n, m, d, k, \sigma^2)$ with signal strength η and mean-squared loss R requires information cost (and hence expected communication cost) at least $\Omega\left(\frac{\sigma^2 d}{n\eta^2}\right)$.*

Note that there is a protocol for $\mathrm{SGME}(n, m, d, k, \sigma^2)$ with signal strength η and mean-squared loss R that has communication cost $\tilde{O}\left(\min\left\{\frac{\sigma^2 d}{n\eta^2} + \frac{\sigma^2 k^2}{nR}, \frac{\sigma^2 dk}{nR}\right\}\right)$. In the regime where $\eta^2 \geq 16R/k$, the first term dominates and by Theorem 7.4.8, and the fact that $\frac{\sigma^2 k^2}{nR}$ is a lower bound even when the machines know the support [GMN14], we also get a matching lower bound. In the regime where $\eta^2 \leq 16R/k$, second term dominates and it is a lower bound by Theorem 7.4.6.

Proof of Theorem 7.4.8. The proof is very similar to the proof of Theorem 7.4.5. Given a protocol estimator pair $(\Pi, \hat{v})$ that solves $\mathrm{SGME}(n, m, d, k, \sigma^2)$ with signal strength η, mean-squared loss R and information cost I (where $\eta^2 \geq 16R/k$), we can find a protocol Π' that solves the Gaussian mean detection problem $\mathrm{GD}(n, m, \eta, \sigma^2)$ with information cost $\leq O(I/d)$ (as usual the information cost is measured when the mean is 0). Π' would be exactly the same as Protocol 13 but with μ_0 replaced by $\mathcal{N}(0, \sigma^2)$, μ_1 replaced by $\mathcal{N}(\eta, \sigma^2)$ and δ replaced by η. We leave the details to the reader. □

7.4.3 Lower bound for Sparse Linear Regression

In this section we consider the sparse linear regression problem $\mathrm{SLR}(n, m, d, k, \sigma^2)$ in the distributed setting as defined in Section 7.2. Suppose the i-th machine receives a subset S_i of the mn data points, and we use $A_{S_i} \in \mathbb{R}^{n \times d}$ to denote the design matrix that the i-th

machine receives and y_{S_i} to denote the observed vector. That is, $y_{S_i} = A_{S_i}\theta + w_{S_i}$, where $w_{S_i} \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$ is Gaussian noise.

This problem can be reduced from the sparse Gaussian mean problem, and thus its communication can be lower-bounded. It follows straightforwardly from our Theorem 7.4.6 and the reduction in Corollary 2 of [DJWZ14]. Suppose $\lambda = \max_{1\le i\le m} \|A_{S_i}\|/\sqrt{n}$.

Corollary 7.4.9. *Suppose machines receive data from the sparse linear regression model. Let λ be as defined above. If there exists a protocol under which the machines can output an estimator $\hat{\theta}$ with mean squared loss $R = \mathbb{E}[\|\hat{\theta} - \theta\|^2]$ with communication C, then $R \cdot C \ge \Omega\left(\frac{\sigma^2 kd}{\lambda^2 n}\right)$.*

When A_{S_i} is a Gaussian design matrix, that is, the rows of A_{S_i} are i.i.d drawn from distribution $\mathcal{N}(0, I_{d\times d})$, we have $\lambda = O\left(\max\{\sqrt{d/n}, 1\}\right)$ and Corollary 7.4.9 implies that to achieve the statistical minimax rate $R = O(\frac{k\sigma^2}{nm})$, the algorithm has to communicate $\Omega(m\cdot\min\{n, d\})$ bits. The point is that we get a lower bound that doesn't depend on k– that is, with sparsity assumptions, it is impossible to improve both the loss and communication so that they depend on the intrinsic dimension k instead of the ambient dimension d. Moreover, in the regime when $d/n \to c$ for a constant c, our lower bound matches the upper bound of [LSLT15] up to a logarithmic factor. The proof follows from Theorem 7.4.6 and the reduction from Gaussian mean estimation to sparse linear regression of [ZDJW13] straightforwardly.

Proof of Corollary 7.4.9. Suppose there exists such a protocol with mean-squared loss R and communication cost C for sparse linear regression problem $\mathrm{SLR}(n, m, k, d, \sigma^2)$. We are going to use it to solve the sparse linear regression problem $\mathrm{SGME}(m, 1, d, k, \sigma_0)$ as follows. Suppose the i^{th} machine has data $X_i \sim \mathcal{N}(\theta, \sigma_0^2 I_{d\times d})$ with $\sigma_0 = \frac{\sigma}{\lambda\sqrt{n}}$. Then the machines can prepare

$$y_{S_i} = A_{S_i} X_i + b_i$$

where $b_i \sim \mathcal{N}(0, \sigma^2 I - \sigma_0^2 A_{S_i} A_{S_i}^T)$. Note that by the bound $\|A_{S_i}\| \leq \lambda/\sqrt{n}$, we have that $\sigma^2 I - \sigma_0^2 A_{S_i} A_{S_i}^T$ is positive semidefinite. Note that then y_{S_i} can written in the form

$$y_{S_i} = A_{S_i}\theta + \xi_i$$

where ξ_i's are independent distributed according to $\mathcal{N}(0, \sigma^2 I_{n\times n})$

Then the machines call the protocol for the sparse linear regression problem with data (y_{S_i}, A_{S_i}). Therefore we obtain a protocol that solves SGME$(m, 1, d, k, \sigma_0)$ with communication R and C. Then by Theorem 7.4.6, we know that

$$R \cdot C \geq \Omega(\sigma_0^2 kd) = \Omega\left(\frac{\sigma^2 kd}{\lambda^2 n}\right)$$

□

7.5 Direct-sum Theorem for Sparse Parameters

Unknown parameter: $v \in \{0, 1\}$
Inputs: Machine j gets n samples $X_j = (X_j^{(1)}, \ldots, X_j^{(n)})$, where X_j is distributed according to μ_v^n.

1. All machines publicly sample k independent coordinates $I_1, \ldots, I_k \subset [d]$ (without replacement).
2. Each machine j locally prepares data $\widetilde{X}_j = \left(\widetilde{X}_{j,1}, \ldots, \widetilde{X}_{j,d}\right)$ as follows: The I_1-th coordinate is embedded with the true data, $\widetilde{X}_{j,I_1} = X_j$. For $r = 2, \ldots, k$, j-th the machine draws $\widetilde{X}_{j,I_r}$ privately from distribution μ_1^n. For any coordinate $i \in [d]\backslash\{I_1, \ldots, I_k\}$, the j-th machine draws privately $\widetilde{X}_{j,i}$ from the distribution μ_0^n.
3. The machines run protocol Π with input data $\widetilde{X}$.
4. If $|\hat{\theta}(\Pi)_{I_1}| \geq \delta/2$, then the machines output 1, otherwise they output 0.

Protocol 13: direct-sum reduction for sparse parameter

We prove Theorem 7.4.5 in this section. Let Π' be the protocol described in Protocol 13. Let $\theta \in \mathbb{R}^d$ be such that $\theta_{I_1} = v\delta$ and $\theta_{I_r} = \delta$ for $r = 2, \ldots, k$, and $\theta_i = 0$ for $i \in [d] \backslash \{I_1, \ldots, I_k\}$. We can see that by our construction, the distribution of $\widetilde{X}_j$ is the same as μ_θ^n, and all X_j's are independent. Also note that θ is k-sparse. Therefore when Π' invokes Π on data $\widetilde{X}$, Π will have loss R and information cost I with respect to $\widetilde{X}$.

We first verify that the protocol Π does distinguish between $v = 0$ and $v = 1$.

Proposition 7.5.1. *Under the assumption of Theorem 7.4.5, when $v = 1$, we have that*

$$\mathbb{E}\left[|\hat{\theta}(\Pi)_{I_1} - \delta|^2\right] \leq \frac{R}{k} \tag{7.22}$$

and when $v = 0$, we have

$$\mathbb{E}\left[|\hat{\theta}(\Pi)_{I_1}|^2\right] \leq \frac{R}{d-k+1} \tag{7.23}$$

Moreover, with probability at least 3/4, Π' outputs the correct answer v.

Proof. We know that Π has mean-squared loss R, that is,

$$\begin{aligned} R((\Pi, \hat{\theta}), \theta) &= \mathbb{E}\left[||\hat{\theta}(\Pi) - \theta||_2^2\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{d} |\hat{\theta}(\Pi)_i - \theta_i|^2\right] \end{aligned}$$

Here the expectation is over the randomness of the protocol Π and randomness of the samples $\widetilde{X}_1, \ldots, \widetilde{X}_m$. We first prove equation (7.23), that is

$$\mathbb{E}\left[|\hat{\theta}(\Pi)_{I_1}|^2\right] \leq \frac{R}{d-k+1}$$

Here the expectation is over $I_1, \ldots, I_k$ in addition to being over the randomness of Π and the samples $\widetilde{X}_1, \ldots, \widetilde{X}_m$. We will in fact prove this claim for any fixing of $I_2, \ldots, I_k$ to some

$i_2, \ldots, i_k$. Then I_1 is a random coordinate in $[d] \backslash \{i_2, \ldots, i_k\}$. Then

$$\begin{aligned}
\mathbb{E}\left[|\hat{\theta}(\Pi)_{I_1}|^2 \mid I_r = i_r, r \geq 2\right] &= \frac{1}{d-k+1} \sum_{i \in [d] \backslash \{i_2, \ldots, i_k\}} \mathbb{E}\left[|\hat{\theta}(\Pi)_i|^2 \mid I_r = i_r, r \geq 2\right] \\
&\leq \frac{1}{d-k+1} \left(\sum_{i \in [d] \backslash \{i_2, \ldots, i_k\}} \mathbb{E}\left[|\hat{\theta}(\Pi)_i|^2 \mid I_r = i_r, r \geq 2\right] \right. \\
&\quad \left. + \sum_{i \in \{i_2, \ldots, i_k\}} \mathbb{E}\left[|\hat{\theta}(\Pi)_i - \delta|^2 \mid I_r = i_r, r \geq 2\right] \right)
\end{aligned}$$

Taking expectation over $I_2, \ldots, I_r$ we obtain

$$\begin{aligned}
\mathbb{E}\left[|\hat{\theta}(\Pi)_{I_1}|^2\right] \leq \frac{1}{d-k+1} \sum_{i=1}^{d} \mathbb{E}\left[|\hat{\theta}(\Pi)_i - \theta|^2\right] &= \frac{1}{d-k+1} R((\Pi, \hat{\theta}), \theta) \\
&\leq \frac{R}{d-k+1}
\end{aligned}$$

In order to prove equation (7.22), we prove the statement for every fixing of $\{I_1, \ldots, I_k\}$ to some $S \subset [d]$.

$$
\begin{aligned}
&\mathbb{E}\left[|\hat{\theta}(\Pi)_{I_1} - \delta|^2 \mid \{I_1, \ldots, I_k\} = S\right] \\
&= \frac{1}{k}\sum_{i\in S} \mathbb{E}\left[|\hat{\theta}(\Pi)_i - \delta|^2 \mid \{I_1, \ldots, I_k\} = S\right] \\
&\leq \frac{1}{k}\left(\sum_{i\in S} \mathbb{E}\left[|\hat{\theta}(\Pi)_i - \delta|^2 \mid \{I_1, \ldots, I_k\} = S\right] + \sum_{i\notin S} \mathbb{E}\left[|\hat{\theta}(\Pi)_i|^2 \mid \{I_1, \ldots, I_k\} = S\right]\right) \\
&= \frac{1}{k}\sum_{i=1}^{d} \mathbb{E}\left[|\hat{\theta}(\Pi)_i - \delta|^2 \mid \{I_1, \ldots, I_k\} = S\right]
\end{aligned}
$$

Taking expectation over $I_1, \ldots, I_k$ we obtain,

$$
\mathbb{E}\left[\mathbb{E}\left[|\hat{\theta}(\Pi)_{I_1} - \delta|^2\right] \mid \{I_1, \ldots, I_k\} = S\right] = \frac{1}{k} R((\Pi, \hat{\theta}), \theta) \leq \frac{R}{k}
$$

The last statement of proposition follows easily from Markov's inequality and the assumption that $R \leq k\delta^2/16$. □

Now we prove the information cost of the protocol Π' under the case $v = 0$ is small.

Proposition 7.5.2. *Under the assumption of Theorem 7.4.5, we have*

$$
\text{min-IC}(\Pi') \leq \mathrm{I}_0(\Pi'; X_1, \ldots, X_m \mid R'_{\text{pub}}) \leq \frac{I}{d-k+1}
$$

where $X_j \sim \mu_0^n$ and R'_{pub} is the public coin used by Π'.

Proof. Let us denote $\left(\widetilde{X}_{j,i}^{(1)}, \ldots, \widetilde{X}_{j,i}^{(n)}\right)$ by $\widetilde{X}_{j,i}$, that is, $\widetilde{X}_{j,i}$ is the collection of i-th coordinates of the samples on machine j. Let R_{pub} be the public coins used by protocol Π. Note that

R'_{pub} are just $I_1, \ldots, I_k$ and R_{pub}, therefore, the information cost of Π' is

$$\begin{aligned}\mathrm{I}_0(\Pi'; X_1, \ldots, X_m \mid R'_{\text{pub}}) &= \mathrm{I}(\Pi; \widetilde{X}_{1,I_1}, \ldots, \widetilde{X}_{m,I_1} | I_1, \ldots, I_k, R_{\text{pub}}) \\ &= \mathop{\mathbb{E}}_{i_2,\ldots,i_k} \left[I(\Pi; \widetilde{X}_{1,I_1}, \ldots, \widetilde{X}_{m,I_1} | I_1, I_2 = i_2, \ldots, I_k = i_k, R_{\text{pub}}) \right]\end{aligned} \tag{7.24}$$

For each $i_2, \ldots, i_k$, we will prove that $\mathrm{I}(\Pi; \widetilde{X}_{1,I_1}, \ldots, \widetilde{X}_{m,I_1} | I_1, I_2 = i_2, \ldots, I_k = i_k, R_{\text{pub}}) \leq I/(d-k+1)$. Note that conditioned on $I_r = i_r$ for $r \geq 2$, I_1 is uniform over $[d] \backslash \{i_2, \ldots, i_k\}$

$$\begin{aligned}&\mathrm{I}(\Pi; \widetilde{X}_{1,I_1}, \ldots, \widetilde{X}_{m,I_1} | I_1, I_2 = i_2, \ldots, I_k = i_k, R_{\text{pub}}) \qquad (7.25) \\ &= \frac{1}{d-k+1} \sum_{i \in [d] \backslash \{i_2,\ldots,i_k\}} \mathrm{I}(\Pi; \widetilde{X}_{1,i}, \ldots, \widetilde{X}_{m,i} | I_1 = i, I_2 = i_2, \ldots, I_k = i_k, R_{\text{pub}}) \\ &= \frac{1}{d-k+1} \sum_{i \in [d] \backslash \{i_2,\ldots,i_k\}} \mathrm{I}(\Pi; \widetilde{X}_{1,i}, \ldots, \widetilde{X}_{m,i} | I_2 = i_2, \ldots, I_k = i_k, R_{\text{pub}}) \\ &\leq \frac{1}{d-k+1} \mathrm{I}\left(\Pi; \left(\widetilde{X}_{1,i}, \ldots, \widetilde{X}_{m,i}\right)_{i \in [d] \backslash \{i_2,\ldots,i_k\}} | I_2 = i_2, \ldots, I_k = i_k, R_{\text{pub}}\right) \\ &\leq \frac{1}{d-k+1} \mathrm{I}(\Pi; \widetilde{X}_1, \ldots, \widetilde{X}_m | I_2 = i_2, \ldots, I_k = i_k, R_{\text{pub}}) \qquad (7.26)\end{aligned}$$

The second equality follows from the fact that the distribution of $\widetilde{X}_{1,i}, \ldots, \widetilde{X}_{m,i}$ for $i \in [d] \backslash \{i_2, \ldots, i_k\}$ does not depend on i and the protocol Π is also oblivious of I_1 and hence we can remove the conditioning on $I_1 = i$. First inequality follows from lemma 1.2.25 and the fact that $\widetilde{X}_{1,i}, \ldots, \widetilde{X}_{m,i}$ are independent across i. The second inequality follows from the fact that $I(A; B) \leq I(A; B, C)$.

Finally, note that Π performs the task $T(n, m, \mathcal{Q})$ with information cost $I = \sup_\theta \mathrm{I}_\theta(\Pi; \tilde{X} \mid R_{\text{pub}})$. Note that conditioned on $I_r = i_r$ and $I_1 = i$, $\widetilde{X}$ are drawn from some valid μ_θ with a k-sparse θ. Therefore by the definition of information cost, we have that

$$\mathrm{I}(\Pi; \widetilde{X}_1, \ldots, \widetilde{X}_m | I_1 = i, I_2 = i_2, \ldots, I_k = i_k, R_{\text{pub}}) \leq I \tag{7.27}$$

Hence it follows from equations (7.24) and (7.26) and (7.27), we have that

$$\mathrm{I}_0(\Pi'; X_1, \ldots, X_m \mid R'_{\mathrm{pub}}) \leq \frac{I}{d-k+1} \tag{7.28}$$

ant it follows by definition that min-IC$(\Pi') \leq \frac{I}{d-k+1}$. □

7.6 Data Processing Inequality for Truncated Gaussian

In this section, we prove Theorem 7.4.1, the SDPI for truncated gaussian distributions. We first survey the connection between SDPI and transportation inequalities established by Raginsky [Rag16] in Section 7.6.1. Then we prove in Section 7.6.2 that when a distribution has log-concave density function on a finite interval, it satisfies the transportation inequalities. These preparations imply straightforwardly Theorem 7.4.1, which is proved in Section 7.6.3.

7.6.1 SDPI Constant and Transportation Inequality

Usually in literature, the inequality (7.1) is referred to SDPI for mutual information. Here we introduce the more common version of strong data processing inequality, which turns out to be generally equivalent to SDPI for mutual information. In this section, it will be more convenient to work with the definition of KL divergence with natural log. We will denote this by D_{kl}. Note that we will mostly be working with ratios of divergence terms for which it doesn't matter which log to take.

Lemma 7.6.1 (Special case of Theorem 4 in [AGKN13])**.** *Consider the joint distribution of (V, X) where $V \sim B_{1/2}$ and conditioned on $V = v$, we have $X \sim \mu_v$. Note that X is distributed according to the distribution $\mu = (\mu_0 + \mu_1)/2$. By Bayes' rule, we can define the reverse channel $K : X \to V$ with transition probabilities $\{K(v|x) : v \in \{0,1\}, x \in \mathbb{R}\}$ the*

same as the conditional probabilities $P_{V|X}$ of the above joint distribution. For any distribution ν over $\mathbb{R}$, let νK denote the distribution of the output v of K if the input x is distributed according to ν. Then

$$\beta(\mu_0, \mu_1) = \sup_{\nu \neq \mu} \frac{\mathrm{D}_{\mathrm{kl}}(\nu K \| \mu K)}{\mathrm{D}_{\mathrm{kl}}(\nu \| \mu)} \tag{7.29}$$

Thus, it suffices to bound from above the RHS of (7.29). We use the technique developed in Theorem 3.7 of [Rag16], which relates the strong data processing inequality with the concentration of measure and specifically the transportation inequality.

To state the transportation inequality, we define the Wasserstein distance $w_1(\cdot, \cdot)$ between two probability measures,

Definition 7.6.2. The w_1 distance between two probability measure μ, ν over $\mathbb{R}$ is defined as

$$w_1(\nu, \mu) = \sup_{f: f \text{ is 1-Lipschitz}} \left| \int f d\nu - \int f d\mu \right| \tag{7.30}$$

We will review transportation inequalities in section 7.6.2, which relate the cost of transporting ν to μ in Wasserstein distance w_1 with the KL-divergence between ν and μ,

$$w_1(\nu, \mu)^2 \leq \alpha \mathrm{D}_{\mathrm{kl}}(\nu \| \mu). \tag{7.31}$$

For a complete survey of transportation inequalities with other cost functions, please see the survey of Gozlan and Léonard [GL10]. However, before moving to transportation inequalities, we show how to use it to derive a bound on $\beta(\mu_0, \mu_1)$.

Lemma 7.6.3 (A special case of Theorem 3.7 [Rag16])**.** *Suppose for any $v \in \{0, 1\}$, $f_v(x) = \Pr[V = v \mid X = x]$ is L-Lipschitz, and transportation inequality* (7.31) *is true for $\mu = (\mu_0 + \mu_1)/2$ and any measure ν, then*

$$\beta(\mu_0, \mu_1) = \sup_{\nu \neq \mu} \frac{\mathrm{D}_{\mathrm{kl}}(\nu K \| \mu K)}{\mathrm{D}_{\mathrm{kl}}(\nu \| \mu)} \leq 4\alpha L^2 \tag{7.32}$$

Proof of Lemma 7.6.3. We basically follow the proof of Theorem 3.7 of [Rag16] with some simplifications and modifications. Note μK is the unbiased Bernoulli distribution and by the fact that KL divergence is not greater than χ^2 divergence (easy consequence of concavity of log), we have

$$\begin{aligned} \mathrm{D}_{\mathrm{kl}}(\nu K \| \mu K) \leq \chi^2(\nu K \| \mu K) &= \sum_{v \in \{0,1\}} \frac{(\mu K(v) - \nu K(v))^2}{\mu K(v)} \\ &= 2 \sum_{v \in \{0,1\}} (\mu K(v) - \nu K(v))^2 \end{aligned} \tag{7.33}$$

Fixing any $v \in \{0, 1\}$, we have that

$$\begin{aligned} |\mu K(v) - \nu K(v)| &= \left| \int \Pr[V = v \mid X = x] d\mu - \int \Pr[V = v \mid X = x] d\nu \right| \\ &= \left| \int f_v(x) d\mu - \int f_v(x) d\nu \right| \\ &\leq L w_1(\nu, \mu) \end{aligned} \tag{7.34}$$

where the last inequality is by the definition of Wasserstein distance and the fact that $f_v(x)$ is L-Lipschitz.

It follows from (7.34) and (7.33) that

$$\mathrm{D}_{\mathrm{kl}}(\nu K \| \mu K) \leq 4L^2 w_1^2(\nu, \mu).$$

Then by transportation inequality (7.31) we have that

$$\mathrm{D}_{\mathrm{kl}}(\nu K \| \mu K) \leq 4L^2 w_1^2(\nu, \mu) \leq 4\alpha L^2 D(\nu \| \mu).$$

□

7.6.2 Proving transportation inequality via concentration of measure

In this subsection, we show that if μ is log-concave then it satisfies transportation inequality (7.31). To obtain the following theorem, we use a series of tools from the theory of concentration of measures in a straightforward way, albeit that in our setting, μ has only support on a finite interval and therefore we need to take some additional care.

Theorem 7.6.4. *Suppose μ is a measure defined on $[a, b]$ with $d\mu = \exp(-u(x))dx$, and $\nabla^2 u(x) \geq c$, then for any measure ν we have*

$$w_1(\nu, \mu)^2 \leq \frac{2}{c} \cdot \mathrm{D}_{\mathrm{kl}}(\nu \| \mu). \tag{7.35}$$

In addition, it can be proved by direct calculation that if both μ_0 and μ_1 are log-concave and μ_0 and μ_1 are not too far away in some sense, then $\mu = (\mu_0 + \mu_1)/2$ is also log-concave with similar parameters.

Lemma 7.6.5. *Suppose distribution μ_0 and μ_1 has supports on $[a, b]$ with $d\mu_0 = \exp(-u_0(x))dx$ and $d\mu_1 = \exp(-u_1(x))dx$. Suppose $\nabla^2 u_0(x) \geq c$, and $\nabla^2 u_1(x) \geq c$, and $|\nabla u_0(x) - \nabla u_1(x)| \leq \sqrt{c}$ then $\mu = \frac{1}{2}(\mu_0 + \mu_1)$ satisfies that $d\mu = \exp(-u(x))dx$ with $\nabla^2 u(x) \geq \frac{c}{2}$.*

Proof of Lemma 7.6.5. Let $u(x)$ be such that $d\mu = \exp(-u(x))dx$, that is,

$$u(x) = -\ln\left(\frac{\exp(-u_0(x)) + \exp(-u_1(x))}{2}\right)$$

We calculate $u''(x)$ as follows:

We can simply calculate the derivatives of u. For simplicity of notation, let $h(x) =$

$\exp(-u_0(x)) + \exp(-u_1(x))$. We have that

$$h' = -u_0' \exp(-u_0) - u_1' \exp(-u_1),$$

and

$$h'' = (u_0'^2 - u_0'') \exp(-u_0) + (u_1'^2 - u_1'') \exp(-u_1).$$

Therefore we have

$$\begin{aligned} u'' &= \frac{-hh'' + h'^2}{h^2} \\ &= \frac{u_0'' \exp(-2u_0) + u_1'' \exp(-2u_1) + (u_0'' + u_1'' - (u_0' - u_1')^2) \exp(-u_1 - u_2)}{(\exp(-u_0) + \exp(-u_1))^2} \end{aligned}$$

With some simple algebraic manipulations we have that $u'' \geq t$ (for $t \leq \min\{\mu_0'', \mu_1''\}$) is equivalent to

$$\begin{aligned} &\left(\sqrt{\mu_0'' - t} \exp(-u_0) - \sqrt{\mu_1'' - t} \exp(-u_1)\right)^2 + \\ &\left(\left(\sqrt{\mu_0'' - t} + \sqrt{\mu_1'' - t}\right)^2 - (u_0' - u_1')^2\right) \exp(-u_0 - u_1) \geq 0 \end{aligned}$$

Therefore, taking $t = \frac{c}{2}$ and under our assumptions that $|\mu_0'(x) - \mu_1'(x)| \leq \sqrt{c}$ for any $x \in [a, b]$, we have that $u'' \geq \frac{c}{2}$ as desired. □

To prove Theorem 7.6.4, we exploit the well-established connections between transportation inequality, concentration of measure and log-Sobolev inequalities. First of all, transportation inequality (7.35) with Wasserstein w_1 and KL-divergence ties closely to the concentration of probability measure μ. The theorem of Bobkov-Gotze established the exact connection:

Theorem 7.6.6 (Bobkov-Gotze [BG99] Theorem 3.1)**.** *Let $\mu \in \mathbb{P}_1$ be a probability measure on a metric space $(\mathbb{X}, d)$. Then the following two are equivalent for $X \sim \mu$.*

1. *$w_1(\nu, \mu) \le \sqrt{2\sigma^2 \mathrm{D}_{\mathrm{kl}}(\nu \| \mu)}$ for all ν.*
2. *$f(X)$ is σ^2-subgaussian for every 1-Lipschitz function f.*

Using Theorem 7.6.6, in order to prove Theorem 7.6.4, it suffices to prove the concentration of measure for $f(X)$ when $X \sim \mu$, and f is 1-Lipschitz. Although one might prove $f(X)$ is subgaussian directly by definition, we use the log-Sobolev inequality to get around the tedious calculation. We begin by defining the entropy of a nonnegative random variable.

Definition 7.6.7. The entropy of the a nonnegative random variable Z is defined as

$$\mathrm{Ent}[Z] := \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z] \tag{7.36}$$

Entropy is very useful for proving concentration of measure. As illustrated in the following lemma, to prove X is subgaussian we only need to bound $\mathrm{Ent}[e^{\lambda X}]$ by $\mathbb{E}[e^{\lambda X}]$.

Lemma 7.6.8 (Herbst, c.f. [Led01])**.** *Suppose that for some random variable X, we have*

$$\mathrm{Ent}[e^{\lambda X}] \le \frac{\lambda^2 \sigma^2}{2} \mathbb{E}[e^{\lambda X}], \qquad \text{for all } \lambda \ge 0 \tag{7.37}$$

Then

$$\psi(\lambda) := \log \mathbb{E}[e^{\lambda(X - \mathbb{E}X)}] \le \frac{\lambda^2 \sigma^2}{2}, \qquad \text{for all } \lambda \ge 0$$

and as an immediate consequences, X is a σ^2-subgaussian random variable.

Therefore by Theorem 7.6.6 and Lemma 7.6.8, in order to prove transportation inequality, it suffices to to upper bound $\mathrm{Ent}_\mu[e^{\lambda f}]$ by $\mathbb{E}[e^{\lambda f}]$. It turns out that as long as the measure μ is log-concave, we get the concentration inequality for $f(X)$ with 1-Lipschitz function f.

Theorem 7.6.9 (Theorem 5.2 of [Led01]). *Let $d\mu = e^{-u}dx$ where for some $c > 0$, $\nabla^2 u(x) \geq c$ for all $x \in \mathbb{R}$. Then for all smooth functions f on $\mathbb{R}$,*

$$\mathrm{Ent}_\mu(f^2) \leq \frac{2}{c}\int |\nabla f|^2 d\mu$$

As a direct corollary, we obtain inequality (7.37) that we are interested in.

Corollary 7.6.10. *Let $d\mu = e^{-u}dx$ where for some $c > 0$, $\nabla^2 u(x) \geq c$ for all $x \in \mathbb{R}$. Then for all 1-Lipschitz and smooth function f on $\mathbb{R}$, and any $\lambda \geq 0$, we have*

$$\mathrm{Ent}_\mu(e^{\lambda f}) \leq \frac{\lambda^2}{2c}\,\mathbb{E}[e^{\lambda f}]$$

Proof of Corollary 7.6.10. Applying directly Theorem 7.6.9 on $e^{\lambda f/2}$ we obtain,

$$\mathrm{Ent}_\mu[e^{\lambda f}] \leq \frac{2}{c}\int |\nabla e^{\lambda f/2}|^2 d\mu = \frac{2}{c}\int |e^{\lambda f/2}\cdot \lambda \nabla f/2|^2 d\mu$$

Note that if f is 1-Lipschitz, we have $|\nabla e^{\lambda f/2}| \leq |\frac{1}{2}\lambda e^{\lambda f/2}|$, and therefore

$$\mathrm{Ent}_\mu[e^{\lambda f}] \leq \frac{\lambda^2}{2c}\int e^{\lambda f} d\mu = \frac{\lambda^2}{2c}\mathbb{E}_\mu[e^{\lambda f}]$$

□

The distributions that we are interested has continuous density function on a finite support and 0 elsewhere. Therefore we need to use a non-continuous version of the Corollary above to be rigorous.

Corollary 7.6.11. *Let $S = [a, b]$ be a finite interval in $\mathbb{R}$. Let $d\mu = e^{-u}dx$ for $x \in S$ and $d\mu = 0$ for $x \notin S$. Suppose for some $c > 0$, we have $\nabla^2 u(x) \geq c$ for all $x \in S$.Then the conclusion of Corollary 7.6.10 is still true.*

Proof of Corollary 7.6.11. We first extend Theorem 7.6.9 to the finite support case. Let g be an extension of f to $\mathbb{R}$, such that g is nonnegative and bounded above by some constant C, and ∇g is also bounded by C. Let u_n be a series of extensions of u to $\mathbb{R}$ such that the following happens: a) u_n is twice-differentiable b) $\nabla^2 u_n(x) \geq c$ for all $x \in \mathbb{R}$ c) $\mu_n = e^{-u_n} dx$ approaches to μ in TV norm as n tends to infinity. The following choice will work, for example,

$$\begin{aligned} u_n(x) &= u(x) \\ &+ \mathbf{1}_{x>b} \cdot \left(\nabla u(b)(x-b) + \nabla^2 u(b)(x-b)^2 + \exp(n(x-b)^4)\right) \\ &+ \mathbf{1}_{x<a} \cdot \left(\nabla u(b)(x-a) + \nabla^2 u(b)(x-a)^2 + \exp(n(x-a)^4)\right) \end{aligned}$$

Since g and ∇g are bounded, we have that $|\mathbb{E}_{\mu_n}(g^2) - \mathbb{E}_{\mu}(g^2)| = \int g^2(d\mu_n - d\mu) \leq C^2 \|\mu_n - \mu\|_{\mathrm{TV}} \to 0$ as n tends to infinity. Similarly we have that $\mathrm{Ent}_{\mu_n}(g^2) \to \mathrm{Ent}_{\mu}(g^2)$ and $\mathbb{E}_{\mu_n}[|\nabla g|^2] \to \mathbb{E}_{\mu}[|\nabla g|^2]$. Note that under μ, g agrees with f and therefore we have that $\mathrm{Ent}_{\mu_n}(g^2) \to \mathrm{Ent}_{\mu}(f^2)$ and $\mathbb{E}_{\mu_n}[|\nabla g|^2] \to \mathbb{E}_{\mu}[|\nabla f|^2]$.

Also note that μ_n satisfies the condition of Theorem 7.6.9, therefore

$$\mathrm{Ent}_{\mu_n}(g^2) \leq \frac{2}{c} \int |\nabla g|^2 d\mu_n$$

and the desired result follows by taking n to infinity. □

7.6.3 SDPI for truncated Gaussian

We first compute the Lipschitz constants for $f_v(x) = \Pr[V = 0 \mid X = x]$ as defined in Lemma 7.6.3. Here $V \sim B_{1/2}$. $X|V = 0 \sim \mu_0'$ and $X|V = 1 \sim \mu_1'$, where μ_0' is the truncation of $\mu_0 = \mathcal{N}(0, \sigma^2)$ to the interval $[-\tau, \tau]$ and μ_1' is the truncation of $\mu_1 = \mathcal{N}(\delta, \sigma^2)$ to the interval $[-\tau, \tau]$.

Lemma 7.6.12. *When X is generated by $X \sim \mu_v$ conditioned on $V = v$, let $f_v(x) = \Pr[V = 0 \mid X = x]$, we have that $f_v(x)$ is $\delta/4\sigma^2$-Lipschitz for any $v \in \{0, 1\}$.*

Proof. The proof is by direct calculation. Note that by definition, on support $[-\tau, \tau]$, $d\mu'_0 = \gamma_0 \exp(-u_0(x))dx$, and $d\mu'_1 = \gamma_1 \exp(-u_1(x))dx$ with $u_0(x) = \frac{x^2}{2\sigma^2}$ and $u_1(x) = \frac{(x-\delta)^2}{2\sigma^2}$, where γ_0 and γ_1 are scaling constants. Note that by the definition of the reverse channel K,

$$f_0(x) = \Pr[V = 0 \mid X = x] = \frac{\gamma_0 e^{-\frac{x^2}{2\sigma^2}}}{\gamma_0 e^{-\frac{x^2}{2\sigma^2}} + \gamma_1 e^{-\frac{(x-\delta)^2}{2\sigma^2}}}$$

Therefore

$$f'_0(x) = -\left(\gamma_0 + \gamma_1 \exp\left(\frac{2x\delta - \delta^2}{2\sigma^2}\right)\right)^{-2} \cdot \gamma_0\gamma_1 \cdot \left(\frac{\delta}{\sigma^2}\right) \cdot \exp\left(\frac{2x\delta - \delta^2}{2\sigma^2}\right)$$

By AM-GM inequality we have

$$f'_0(x) \geq -\left(4\gamma_0\gamma_1 \exp\left(\frac{2x\delta - \delta^2}{2\sigma^2}\right)\right)^{-1} \cdot \gamma_0\gamma_1 \cdot \left(\frac{\delta}{\sigma^2}\right) \cdot \exp\left(\frac{2x\delta - \delta^2}{2\sigma^2}\right) = -\frac{\delta}{4\sigma^2}$$

Similarly for $f_1(x)$ we have

$$f_1(x) = \frac{\gamma_1 e^{-\frac{(x-\delta)^2}{2\sigma^2}}}{\gamma_0 e^{-\frac{x^2}{2\sigma^2}} + \gamma_1 e^{-\frac{(x-\delta)^2}{2\sigma^2}}}$$

and

$$f'_1(x) = \left(\gamma_1 + \gamma_0 \exp\left(\frac{-2x\delta + \delta^2}{2\sigma^2}\right)\right)^{-2} \cdot \gamma_0\gamma_1 \cdot \left(\frac{\delta}{\sigma^2}\right) \cdot \exp\left(\frac{-2x\delta + \delta^2}{2\sigma^2}\right) \leq \frac{\delta}{4\sigma^2}$$

Also note that $f'_0 \leq 0$ and $f'_1 \geq 0$. Therefore for any v, f_v is $\frac{\delta}{4\sigma^2}$-Lipschitz. □

Proof of Theorem 7.4.1. Note that by definition on support $[-\tau, \tau]$, $d\mu'_0 = \gamma_0 \exp(-u_0(x))dx$, and $d\mu'_1 = \gamma_1 \exp(-u_1(x))dx$ with $u_0(x) = \frac{x^2}{2\sigma^2}$ and $u_1(x) = \frac{(x-\delta)^2}{2\sigma^2}$. By Lemma 7.6.5, we have

that $\mu = (\mu_0' + \mu_1')/2$ is $1/2\sigma^2$ log-concave, and therefore by Theorem 7.6.4, we have

$$w_1(\nu, \mu)^2 \leq 4\sigma^2 \cdot \mathrm{D}_{\mathrm{kl}}(\nu \| \mu).$$

By Lemma 7.6.12, we have that f_v's are $\delta/4\sigma^2$-Lipschitz and therefore by Lemma 7.6.3, we have that

$$\beta(\mu_0, \mu_1) \leq \delta^2/\sigma^2$$

□

Then we present the proof of Corollary 7.4.2, which relies on the following observation:

Lemma 7.6.13. *Suppose $V \to (X_1, \ldots, X_n) \to \Pi$ forms a Markov Chain, where conditioned on $V = v$, $(X_1, \ldots, X_n)$ are distributed according to μ_v'. Then $V \to X_1 + \cdots + X_n \to (X_1, \ldots, X_n) \to \Pi$ also forms a Markov Chain.*

Proof. Let us look at the density of $(X_1, \ldots, X_n)$ conditioned on $X_1 + \cdots + X_n = l \leq \tau$ and $V = v$. Suppose $x_1, \cdots, x_n$ be such that $\sum_i x_i = l$, then for some normalizing constant C

$$\begin{aligned} p(x_1, \cdots, x_n | l, v) &= C \frac{e^{-(x_1 - v\delta)^2/2\sigma^2} \cdots e^{-(x_n - v\delta)^2/2\sigma^2}}{e^{-(l - nv\delta)^2/2n\sigma^2}} \\ &= C e^{(l-nv\delta)^2/2n\sigma^2 - \sum_i (x_i - v\delta)^2/2\sigma^2} \\ &= C e^{\frac{(l-nv\delta)^2 - n\sum_i (x_i - v\delta)^2}{2n\sigma^2}} \\ &= C e^{\frac{l^2 - n\sum_i x_i^2}{2n\sigma^2}} \end{aligned}$$

which is independent of v and that proves the lemma. Note that we used the fact that $\sum_i x_i = l$ to simplify the expression. □

Now we are ready to prove Corollary 7.4.2.

Proof. (Of corollary 7.4.2) Let us restate what we want to prove. Suppose $V \sim B_{1/2}$,

$(X_1, \ldots, X_n)|V = 0 \sim \tilde{\mu}_0$ and $(X_1, \ldots, X_n)|V = 1 \sim \tilde{\mu}_1$ and $V \to (X_1, \ldots, X_n) \to \Pi$ be a Markov chain. Then

$$I(\Pi; V) \leq \frac{n\delta^2}{\sigma^2} I(\Pi; X_1, \ldots, X_n)$$

By lemma 7.6.13, $V \to X_1 + \cdots + X_n \to (X_1, \ldots, X_n) \to \Pi$ also forms a Markov chain. Then

$$I(\Pi; V) \leq \frac{n\delta^2}{\sigma^2} I(\Pi; X_1 + \cdots + X_n) \leq \frac{n\delta^2}{\sigma^2} I(\Pi; X_1, \ldots, X_n)$$

where the first inequality follows from Theorem 7.4.1 and the fact that the distribution of $X_1 + \cdots + X_n|V = 0$ is the Gaussian $\mathcal{N}(0, n\sigma^2)$ truncated to $[-\tau, \tau]$ and the distribution of $X_1 + \cdots + X_n|V = 1$ is the Gaussian $\mathcal{N}(n\delta, n\sigma^2)$ truncated to $[-\tau, \tau]$. The second inequality follows from data processing. □

7.7 Tight Upper Bound with One-way Communication

In this section, we describe a one-way communication protocol achieving the tight minimal communication for Gaussian mean estimation problem $\mathrm{GME}(n, m, d, \sigma^2)$ with the assumption that $|\theta|_\infty \leq \frac{\sigma}{\sqrt{n}}$. Note that for the design of protocol, it suffices to consider a one-dimensional problem. Protocol 14 solves the one-dimensional Gaussian mean estimation problem, with each machine sending exactly 1 bit, and therefore the total communication is m bits. To get a d-dimensional protocol, we just need to apply Protocol 14 to each dimension. In order to obtain the tradeoff as stated in Theorem 7.4.7, one needs to run Protocol 14 on the first αm machines, and let the other machines be idle.

The correctness of the protocol follows from the following theorem.

Theorem 7.7.1. *The algorithm described in Protocol 14 uses m bits of communication and*

Unknown parameter $\theta \in [-\sigma/\sqrt{n}, \sigma/\sqrt{n}]$
Inputs: Machine i gets n samples $(X_i^{(1)}, \ldots, X_i^{(n)})$ where $X_i^{(j)} \sim \mathcal{N}(\theta, \sigma)$.

- Simultaneously, each machine i
  1. Computes $X_i = \frac{1}{\sigma\sqrt{n}} \sum_{j=1}^{n} X_i^{(j)}$
  2. Sends B_i

$$B_i = \begin{cases} 1 & \text{if } X_i \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

- Machine 1 computes

$$T = \sqrt{2} \cdot \text{erf}^{-1}\left(\frac{1}{m}\sum_{i=1}^{m} B_i\right)$$

  where erf^{-1} is the inverse of the Gauss error function.

- It returns the estimate $\hat{\theta} = \frac{\sigma}{\sqrt{n}}\hat{\theta}'$ where $\hat{\theta}' = \max(\min(T, 1), -1)$ is obtained by truncating T to the interval $[-1, 1]$.

Protocol 14: A simultaneous algorithm for estimating the mean of a normal distribution in the distributed setting.

achieves the following mean squared loss.

$$\mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = O\left(\frac{\sigma^2}{mn}\right)$$

where the expectation is over the random samples and the random coin tosses of the machines.

Proof. Let $\bar{\theta} = \theta\sqrt{n}/\sigma$.

Notice that X_i is distributed according to $\mathcal{N}(\bar{\theta}, 1)$. Our goal is to estimate $\bar{\theta}$ from the X_i's. By our assumption on θ, we have $\bar{\theta} \in [-1, 1]$.

The random variables B_i are independent with each other. We consider the mean and variance of B_i's. For the mean we have that,

$$\mathbb{E}\left[B_i\right] \;=\; \mathbb{E}\left[2 \cdot \Pr[0 \le X_i] - 1\right]$$

For any $i \in [m]$,

$\Pr[0 \le X_i] = \Pr[-X_i \le 0] = \Phi_{-\bar{\theta},1}(0)$, where Φ_{μ,σ^2} is the CDF of normal distribution $\mathcal{N}(\mu, \sigma^2)$. Note the following relation between the error function and the CDF of a normal random variable

$$\Phi_{\mu,\sigma^2}(x) = \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)$$

Hence,

$$\mathbb{E}\left[B_i\right] = \operatorname{erf}(\bar{\theta}/\sqrt{2}).$$

Let $B = \frac{1}{m}\sum_{i=1}^{m} B_i$, then we have that $\mathbb{E}[B] = \operatorname{erf}(\bar{\theta}/\sqrt{2}) \le \operatorname{erf}(1/\sqrt{2})$ and therefore by a Chernoff bound,

the probability that $B > \operatorname{erf}(1)$ or $B \le \operatorname{erf}(-1)$ is $\exp(-\Omega(m))$. Thus, with probability at least $1 - \exp(-\Omega(m))$, we have $\operatorname{erf}(-1) \le B \le \operatorname{erf}(1)$ and therefore $|T| \le \sqrt{2}$.

Let $\mathcal{E}$ be the event that $|T| \le \sqrt{2}$, then we have that the error of $\bar{\theta}$ is bounded by

$$\begin{aligned}
\mathbb{E}[|\hat{\theta}' - \bar{\theta}|^2] &= \mathbb{E}[|\hat{\theta}' - \bar{\theta}|^2 \mid \mathcal{E}]\Pr[\mathcal{E}] + \mathbb{E}[|\hat{\theta}' - \bar{\theta}|^2 \mid \bar{\mathcal{E}}]\Pr[\bar{\mathcal{E}}] \\
&\le \mathbb{E}[|\sqrt{2}\operatorname{erf}^{-1}(B) - \sqrt{2}\operatorname{erf}^{-1}(\mathbb{E}[B])|^2 \mid \mathcal{E}]\Pr[\mathcal{E}] + 2\Pr[\bar{\mathcal{E}}] \\
&= \mathbb{E}[|\sqrt{2}\operatorname{erf}^{-1}(B) - \sqrt{2}\operatorname{erf}^{-1}(\mathbb{E}[B])|^2 \mid \mathcal{E}]\Pr[\mathcal{E}] + 2\exp(-\Omega(m))
\end{aligned}$$

Let $M = \max_{\operatorname{erf}^{-1}(x)\in[-1,1]} \frac{d\operatorname{erf}^{-1}(x)}{dx} < 3$. Then we have that $|\operatorname{erf}^{-1}(x) - \operatorname{erf}^{-1}(y)| \le M|x-y| \le$

$O(1) \cdot |x - y|$ for any $x, y \in [-1, 1]$. Therefore it follows that

$$\begin{aligned}
\mathbb{E}[|\hat{\theta}' - \bar{\theta}|^2] &\leq \mathbb{E}[|\sqrt{2}\operatorname{erf}^{-1}(B) - \sqrt{2}\operatorname{erf}^{-1}(\mathbb{E}[B])|^2 \mid \mathcal{E}] \Pr[\mathcal{E}] + 2\exp(-\Omega(m)) \\
&\leq \mathbb{E}[2M^2|B - \mathbb{E}[B]|^2 \mid \mathcal{E}] \Pr[\mathcal{E}] + 2\exp(-\Omega(m)) \\
&\leq \mathbb{E}[2M^2|B - \mathbb{E}[B]|^2] + 2\exp(-\Omega(m)) \\
&\leq O\left(\frac{1}{m}\right) + 2\exp(-\Omega(m)) \\
&\leq O\left(\frac{1}{m}\right)
\end{aligned}$$

Hence we have that

$$\mathbb{E}\left[|\hat{\theta} - \theta|^2\right] = \frac{\sigma^2}{n}\,\mathbb{E}\left[|\hat{\theta}' - \bar{\theta}|^2\right] = O\left(\frac{\sigma^2}{mn}\right)$$

□

7.7.1 Extension to general θ

Now we do not assume that $\theta_\ell \in [-\sigma/\sqrt{n}, \sigma/\sqrt{n}]$ for each dimension $\ell \in [d]$, and still show how to achieve a 1-round protocol with $O(md)$ bits of communication, up to low order terms. We will make the simplifying and standard assumptions though, that $|\theta_\ell| \leq U = \text{poly}(md)$ for each $\ell \in [d]$, as well as $\log(mdn/\sigma) = o(m)$ and $mdn/\sigma \geq (mdn)^c$ for a constant $c > 0$.

As before, it suffices to consider a one-dimensional problem. Protocol 15 solves the one-dimensional Gaussian mean estimation problem using $O(m + \log^2(mdn/\sigma))$ bits of communication. To solve the d-dimensional problem, we run the protocol independently on each coordinate. The total communication will be $O(md + d\log^2(mdn/\sigma))$ bits. We fix $\ell \in [d]$ and let $\theta = \theta_\ell$. Let $\bar{\theta} = \theta\sqrt{n}/\sigma$, where now we no longer assume $\bar{\theta} \leq 1$. We will show the

output $\hat{\theta}$ satisfies:

$$\mathbb{E}[|\hat{\theta} - \bar{\theta}|^2] = O\left(\frac{1}{m}\right),$$

from which it follows that

$$\mathbb{E}[|\frac{\sigma}{\sqrt{n}}\hat{\theta} - \theta|^2] = O\left(\frac{\sigma^2}{mn}\right).$$

We now describe the one-dimensional problem for a given unknown mean $\bar{\theta}$. The first $r = O(\log(mdn/\sigma))$ machines i send the first $O(\log(mdn/\sigma))$ bits of their (averaged) input Gaussians $X_i = \frac{1}{\sigma\sqrt{n}} \sum_{j=1}^{n} X_i^{(j)}$ to the coordinator. Note that the random variables X_i are distributed according to $\mathcal{N}(\bar{\theta}, 1)$.

Since $O(\log(mdn/\sigma))$ bits of each X_i are communicated to the coordinator, since $\bar{\theta} \leq \text{poly}(md) \cdot \sqrt{n}/\sigma$ (here we use our assumption that $|\theta_\ell| \leq \text{poly}(md)$ for each $\ell \in [d]$), and since each X_i has variance 1, it follows by standard Chernoff bounds that the median γ of $X_1, \ldots, X_r$ is within an additive $\frac{1}{100}$ of $\bar{\theta}$ with probability $1 - \frac{1}{(mdn/\sigma)^\alpha}$ for an arbitrarily large constant $\alpha > 0$ depending on the value $r = O(\log(mdn/\sigma))$. We call this event $\mathcal{E}$, so $\Pr[\mathcal{E}] \geq 1 - \frac{1}{(mdn/\sigma)^\alpha}$.

In parallel, machines $r+1, r+2, \ldots, m$ do the following. Let $R_i \in [0, 1)$ be such that $R_i = X_i - \lfloor X_i \rfloor$. Similarly, let $R_i' \in [0, 1)$ be such that $R_i' = X_i + 1/5 - \lfloor X_i + 1/5 \rfloor$.

For $i = r+1, \ldots, m$, the i-th machine sends a bit $B_i \in \{0, 1\}$, where

$$\Pr[B_i = 1] = R_i,$$

and the i-th matchine also sends a bit $B_i' \in \{0, 1\}$ where

$$\Pr[B_i' = 1] = R_i'.$$

We describe the output of the coordinator in the proof of correctness below. Observe that the overall communication is $O(m + \log^2(mdn/\sigma))$, as desired.

Correctness. Consider the "sawtooth" wave $f(x)$, which for a parameter L, satisfies $f(x) = x/(2L)$ for $x \in [0, 2L)$, and is periodic with period $2L$. Its Fourier series[8] is given by

$$f(x) = \frac{1}{2} - \frac{1}{\pi}\sum_{k=1}^{\infty}\frac{1}{k}\sin\left(\frac{k\pi x}{L}\right).$$

We set $L = 1/2$ and note that $f(X_i) = R_i$. Then, for $X \sim N(\bar{\theta}, 1)$, using a standard transformation of the Gaussian distribution,

$$\mathbf{E}[\sin(tX)] = e^{-t^2/2}\sin(t\bar{\theta}),$$

we have

$$\begin{aligned}
\mathbf{E}[B_i] &= \mathbf{E}[R_i] \\
&= \mathbf{E}[f(X_i)] \\
&= \frac{1}{2} - \frac{1}{\pi}\sum_{k=1}^{\infty}\frac{1}{k}e^{-(k\pi/L)^2/2}\sin(k\pi\bar{\theta}/L) \\
&= \frac{1}{2} - \frac{1}{\pi}\sum_{k=1}^{\infty}\frac{1}{k}e^{-2k^2\pi^2}\sin(2k\pi\bar{\theta}).
\end{aligned}$$

Let $B = \frac{1}{m}\sum_{i=r+1}^{m} B_i$, so that $\mathbf{E}[B] = \mathbf{E}[B_i]$. Since the B_i are Bernoulli random variables,

$$\mathbf{E}[|B - \mathbf{E}[B]|^2] \leq \frac{1}{m-r} \leq \frac{2}{m}, \tag{7.38}$$

where the second inequality uses that $r = O(\log(mdn/\sigma))$ is at most $m/2$ under our assumption that $\log(mdn/\sigma) = o(m)$.

In an analogous fashion the coordinator computes a B' using the B'_i.

If event $\mathcal{E}$ occurs, then the coordinator knows γ satisfying $|\gamma - \bar{\theta}| < \frac{1}{100}$, and using γ

[8] See, e.g., `http://mathworld.wolfram.com/FourierSeriesSawtoothWave.html`

together with B, will output its estimate to $\bar{\theta}$ as follows. Let $\{x\} = x - \lfloor x \rfloor$. The coordinator checks which of the two conditions γ satisfies:

1. $1/50 < \{\gamma\} < 49/50$ and $|\{\gamma\} - 1/4| \geq 3/100$ and $|\{\gamma\} - 3/4| \geq 3/100$
2. $1/50 < \{\gamma + 1/5\} < 49/50$ and $|\{\gamma + 1/5\} - 1/4| \geq 3/100$ and $|\{\gamma + 1/5\} - 3/4| \geq 3/100$.

We note that one of these two conditions must be satisfied. To see this, suppose the first condition is not satisfied. If it is not satisfied because $\{\gamma\} < 1/50$, then $\{\gamma + 1/5\} \in [1/5, 1/5 + 1/50]$, which satisfies the second of the two conditions. If it is not satisfied because $\{\gamma\} > 49/50$, then $\{\gamma + 1/5\} \in [1/5 - 1/50, 1/5]$, which satisfies the second of the two conditions. If the first condition is not satisfied because $\{\gamma\} \in [1/4 - 1/50, 1/4 + 1/50]$, then $\{\gamma + 1/5\} \in [9/20 - 1/50, 9/20 + 1/50]$ and the second condition is satisfied. If the first condition is not satisfied because $\{\gamma\} \in [3/4 - 1/50, 3/4 + 1/50]$, then $\{\gamma + 1/5\} \in [19/20 - 1/50, 19/20 + 1/50]$, which satisfies the second condition.

If the first condition holds, the coordinator will use B and estimate $\bar{\theta}$ below, otherwise it will use B' and estimate $\bar{\theta} + 1/5$ below. We will analyze the first case; the second case is analogous. Note that since $\{\gamma\} > 1/50$, and $|\gamma - \bar{\theta}| < \frac{1}{100}$, the coordinator learns $Z = \lfloor \bar{\theta} \rfloor$. Its estimate $\hat{\theta}$ for $\bar{\theta}$ is then $Z + g(B)$, for a function $g(B)$ to be specified (in the other case the coordinator would have learned $\{\bar{\theta} + 1/5\}$ and $\hat{\theta}$ would have been $\{\bar{\theta} + 1/5\} + g(B') - 1/5$).

To define $g(B)$, we need the following claim. Note that in the first case $|\{\gamma\} - 1/4| \geq 3/100$ and so by the triangle inequality $|\{\bar{\theta}\} - 1/4| \geq 3/100 - \gamma = 1/50$. Similarly, $|\{\bar{\theta}\} - 3/4| \geq 1/50$, so the conditions of the following claim hold for $\{\bar{\theta}\}$.

Claim 7.7.2. *Define* $h(x) = \sum_{k=1}^{\infty} \frac{1}{k} e^{-2k^2\pi^2} \sin(2k\pi x)$. *There exists a constant* $C > 0$ *with the following guarantee. If* $|\{\bar{\theta}\} - 1/4| \geq 1/50$ *and* $|\{\bar{\theta}\} - 3/4| \geq 1/50$ *then for any number* $x \in [\{\bar{\theta}\} - 1/100, \{\bar{\theta}\} + 1/100]$,

$$C \leq h'(x) \leq 1.$$

Before proving the claim, we conclude the correctness proof. The coordinator guesses $\frac{i}{\sqrt{m}}$ for each integer i for which $|Z + \frac{i}{\sqrt{m}} - \gamma| < \frac{1}{100}$. For each guess $\frac{i}{\sqrt{m}}$, the coordinator checks if

$$|\sum_{k=1}^{\infty} \frac{1}{k} e^{-2k^2\pi^2} \sin(2k\pi \frac{i}{\sqrt{m}}) - \pi(\frac{1}{2} - B)| \leq \frac{1}{\sqrt{m}} \tag{7.39}$$

Note that, since the above Fourier series is periodic between succesive integers, we need not add Z to $\frac{i}{\sqrt{m}}$ in (7.39). Let $g(B)$ be the first guess which passes the check. The coordinator outputs $\hat{\theta} = Z + g(B)$ as its estimate to $\bar{\theta}$ (the second case is analogous, in which Z corresponds to $\lfloor \bar{\theta} + 1/5 \rfloor$ and $g(B')$ is defined in the same way). If there is no such $g(B)$ the coordinator just outputs γ. Note also that if its output ever exceeds our assumed upper bound $U = \text{poly}(mnd/\sigma)$ on the magnitude of $\bar{\theta}$, then we instead output U.

Then

$$\begin{aligned}
\mathbf{E}[|\hat{\theta} - \bar{\theta}|^2] &= \mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E}] \Pr[\mathcal{E}] + \mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \neg\mathcal{E}] \Pr[\neg\mathcal{E}] \\
&= \mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E}](1 - \frac{1}{(mdn/\sigma)^{\alpha}}) + 4U^2 \cdot \frac{1}{(nmd/\sigma)^{\alpha}} \\
&\leq \mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E}](1 - \frac{1}{(mdn)^{c\alpha}}) + 4U^2 \cdot \frac{1}{(mdn)^{c\alpha}} \\
&\leq \mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E}] + \frac{1}{m},
\end{aligned} \tag{7.40}$$

where the first inequality uses our assumption that $(mdn/\sigma) \geq (mdn)^c$ for a constant $c > 0$, and the second inequality holds for a sufficiently large constant $\alpha > 0$.

Conditioned on $\mathcal{E}$, we have $\hat{\theta} - \bar{\theta} = g(B) - \{\theta\}$. If (7.39) holds for a given $\frac{i}{\sqrt{m}}$, then

$$|\sum_{k=1}^{\infty} \frac{1}{k} e^{-2k^2\pi^2} \sin(2k\pi \frac{i}{\sqrt{m}}) - \pi(\frac{1}{2} - B)| \leq \frac{1}{\sqrt{m}}.$$

Let $\mathcal{F}$ be the event that the coordinator finds such an $\frac{i}{\sqrt{m}}$ for which (7.39) holds. We use

the shorthand $h(z)$ to denote $\sum_{k=1}^{\infty} \frac{1}{k} e^{-2k^2\pi^2} \sin(2k\pi z)$.

$$\begin{aligned}
\mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E} \wedge \mathcal{F}] &= \mathbf{E}[|\frac{i}{\sqrt{m}} - \{\bar{\theta}\}|^2 \mid \mathcal{E} \wedge \mathcal{F}] \\
&\leq \mathbf{E}[|h(\frac{i}{\sqrt{m}}) - h(\{\bar{\theta}\})|^2 \mid \mathcal{E} \wedge \mathcal{F}] \\
&\leq \mathbf{E}[(|h(\frac{i}{\sqrt{m}}) - \pi(\frac{1}{2} - B)| + |\pi(\frac{1}{2} - B) - h(\{\bar{\theta}\})|)^2 \mid \mathcal{E} \wedge \mathcal{F}] \\
&\leq \mathbf{E}[(\frac{1}{\sqrt{m}} + |\pi(\frac{1}{2} - B) - \pi(\frac{1}{2} - \mathbf{E}[B])|)^2 \mid \mathcal{E} \wedge \mathcal{F}] \\
&\leq \mathbf{E}[(\frac{1}{\sqrt{m}} + \pi|B - \mathbf{E}[B]|)^2 \mid \mathcal{E} \wedge \mathcal{F}] \\
&\leq \frac{2}{m} + 2\pi^2 \mathbf{E}[|B - \mathbf{E}[B]|^2 \mid \mathcal{E} \wedge \mathcal{F}]
\end{aligned}$$

where the first equality follows from $\hat{\theta} - \bar{\theta} = g(B) - \{\theta\}$, the first inequality uses the fact that the algorithm ensures $|\frac{i}{\sqrt{m}} - \{\bar{\theta}\}| \leq \frac{1}{100}$ given that $\mathcal{E}$ occurs and therefore one can apply Claim 7.7.2 with $x = \frac{i}{\sqrt{m}}$ to conclude that $|h(\frac{i}{\sqrt{m}}) - h(\{\bar{\theta}\})| \leq |\frac{i}{\sqrt{m}} - \{\bar{\theta}\}|$, the second inequality is the triangle inequality, the third inequality uses the guarantee on the value $\frac{i}{\sqrt{m}}$ chosen by the coordinator and the definition of $\mathbf{E}[B]$, the fourth inequality rearranges terms, and the fifth inequality uses $(a+b)^2 \leq 2a^2 + 2b^2$.

If there is no value $\frac{i}{\sqrt{m}}$ for which (7.39) holds, then since $\mathcal{E}$ occurs it means there is no integer multiple of $\frac{1}{\sqrt{m}}$, call it x, with $|x - \{\bar{\theta}\}| \leq \frac{1}{100}$ for which $|h(x) - \pi(\frac{1}{2} - B)| \leq \frac{1}{\sqrt{m}}$. If it were the case that $|\mathbf{E}[B] - B| < \frac{C}{100\pi}$, where $C > 0$ is the constant of Claim 7.7.2, then $|\frac{1}{2} - \frac{1}{\pi}h(\bar{\theta}) - B| < \frac{C}{100\pi}$, or equivalently, $|\pi(\frac{1}{2} - B) - h(\bar{\theta})| < \frac{C}{100}$. By Claim 7.7.2, though, we can find an x which is an integer multiple of $\frac{1}{\sqrt{m}}$ which is within $\frac{1}{\sqrt{m}}$ of y, where $h(y) = \pi(\frac{1}{2} - B)$. This follows since the derivative on $[\{\bar{\theta}\} - 1/100, \{\bar{\theta}\} + 1/100]$ is at least C. But then $|h(x) - h(y)| \leq |x - y| \leq \frac{1}{\sqrt{m}}$, contradicting that (7.39) did not hold. It follows that in this case $|\mathbf{E}[B] - B| \geq \frac{C}{100\pi}$. Now in this case, we obtain an additive $\frac{1}{100}$ approximation,

and so $|\hat{\theta} - \bar{\theta}|^2 \leq \frac{\pi^2}{C^2}|B - \mathbf{E}[B]|^2$. Hence,

$$\mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E} \wedge \neg\mathcal{F}] \leq O(1) \cdot \mathbf{E}[|B - \mathbf{E}[B]|^2 \mid \mathcal{E} \wedge \neg\mathcal{F}],$$

and so

$$\begin{aligned}
\mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E}] &\leq \mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E}, \mathcal{F}] \Pr[\mathcal{F}] + \mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E}, \neg\mathcal{F}] \Pr[\neg\mathcal{F}] \\
&\leq \frac{2}{m} + 2\pi^2 \mathbf{E}[|B - \mathbf{E}[B]|^2 \mid \mathcal{E} \wedge \mathcal{F}] \Pr[\mathcal{F}] + O(1) \cdot \mathbf{E}[|B - \mathbf{E}[B]|^2 \mid \mathcal{E} \wedge \neg\mathcal{F}] \Pr[\neg\mathcal{F}] \\
&\leq O\left(\frac{1}{m}\right) + O(1) \cdot \mathbf{E}[|B - \mathbf{E}[B]|^2 \mid \mathcal{E}] \\
&\leq O\left(\frac{1}{m}\right),
\end{aligned}$$

where the final inequality uses $\mathbf{E}[|B - \mathbf{E}[B]|^2 \mid \mathcal{E}] \leq \frac{\mathbf{E}[|B-\mathbf{E}[B]|^2]}{\Pr[\mathcal{E}]} \leq 2\mathbf{E}[|B - \mathbf{E}[B]|^2]$, and (7.38).

Combining this with (7.40) completes the proof that $\mathbf{E}[|\hat{\theta} - \bar{\theta}|^2] = O(1/m)$.

Proof of Claim. We need to understand the derivative, with respect to x, of the function

$$h(x) = \sum_{k=1}^{\infty} \frac{1}{k} e^{-2k^2\pi^2} \sin(2k\pi x),$$

which is equal to

$$h'(x) = \sum_{k=1}^{\infty} 2\pi e^{-2k^2\pi^2} \cos(2k\pi x).$$

Note that the function is periodic in x with period 1, so we can restrict to $x \in [0, 1)$. Consider $z = 2\pi x$. Suppose first that $|z - \pi/2| > \epsilon$ and $|z - 3\pi/2| > \epsilon$ for a constant $\epsilon > 0$ to be determined. Then,

$$|\cos(2\pi z)| \geq \cos(\pi/2 - \epsilon) = \sin(\epsilon) \geq 2\epsilon/\pi,$$

using that $\cos(\pi/2 - \epsilon) = \sin(\epsilon)$ and that $\sin(x)/x \geq 2/\pi$ for $0 < x < \pi/2$. In this case, it

follows that

$$|h'(x)| \;\geq\; (2\pi)e^{-2\pi^2}2\epsilon/\pi - \sum_{k>1} 2\pi e^{-2k^2\pi^2} \geq 4e^{-2\pi^2}\epsilon - 4\pi e^{-8\pi^2},$$

using that the summation is dominated by a geometric series. Note that this expression is at least $4e^{-2\pi^2}(\epsilon - \pi e^{-6\pi^2})$, and so setting $\epsilon = 2\pi e^{-6\pi^2}$ shows that $|h'(x)| = \Omega(1)$. Notice that x satisfies $|2\pi x - \pi/2| > \epsilon$ provided $|x - 1/4| \geq 1/100 > \epsilon/(2\pi)$ and that x satisfies $|2\pi x - 3\pi/2| > \epsilon$ provided that $|x - 3/4| \geq 1/100 > \epsilon/(2\pi)$. As $|\{\bar{\theta}\} - 1/4| \geq 1/50$ and $|\{\bar{\theta}\} - 3/4| \geq 1/50$, it follows that $x \in [\{\bar{\theta}\} - 1/100, \{\bar{\theta}\} + 1/100]$. Hence, $|h'(x)| = \Omega(1)$ for such x, as desired.

On the other hand, it is clear that $h'(x) \leq 1$, by upper bounding $\cos(2k\pi x)$ by 1 and using a geometric series to bound $h'(x)$. □

7.8 Distributed Gap Majority

Our techniques can also be used to obtain a cleaner proof of the lower bound on the information complexity of distributed gap majority due to Woodruff and Zhang [WZ12]. In this problem, there are k parties/machines and the i^{th} machine receives a bit z_i. The machines communicate via a shared blackboard and their goal is to decide whether $\sum_{i=1}^{k} z_i \leq k/2 - \sqrt{k}$ or $\sum_{i=1}^{k} z_i \geq k/2 + \sqrt{k}$. In [WZ12], it was proven that the information complexity of this problem is $\Omega(k)$. We give a different proof using strong data processing inequalities.

The distribution we will consider is the following: let $B \sim B_{1/2}$. Denote $B_{1/2+10/\sqrt{k}}$ by μ_1 and $B_{1/2-10/\sqrt{k}}$ by μ_0. If $B = 1$, sample $Z_1, \ldots, Z_k$ according to μ_1^k. If $B = 0$, sample $Z_1, \ldots, Z_k$ according to μ_0^k.

Theorem 7.8.1. *Suppose π is a k-party protocol (with inputs $Z_1, \ldots, Z_k$) and π solves the gap majority problem (up to some error). Then $I(\Pi; Z_1, \ldots, Z_k | B = 0) \geq \Omega(k)$.*

Π is the random variable for the transcript of the protocol π. The intuition for the proof is pretty simple. It is not hard to verify that since π solves the gap majority problem, it should be able to estimate B as well i.e. $I(\Pi; B) \geq \Omega(1)$. However since each Z_i has only $\Theta(1/k)$ information about B, the protocol needs to gather information about $\Omega(k)$ of the Z_i's. It is satisfying that this intuition can indeed be formalized! Perhaps worth noting that similar intuition can be drawn for the two-party gap hamming distance problem but there we don't have a completely information theoretic proof of the linear lower bound [CR11]. We will be using the strong data processing inequality for the binary symmetric channel first proven by [AG76]. It studies how information decays on a binary symmetric channel. Suppose X be a bit distributed according to $B_{1/2}$. Y be another bit obtained from X by passing it through a binary symmetric channel with error $1/2 - \epsilon$ (i.e. Y remains X w.p. $1/2 + \epsilon$ and gets flipped w.p. $1/2 - \epsilon$). Then for any random variable U s.t. $U - X - Y$ is a Markov chain, $I(U; Y) \leq 4\epsilon^2 I(U; X)$.

Proof. We will denote by $\Pi_{b_1,\ldots,b_k}$ the transcript of the protocol π when the inputs to π are sampled according to $\mu_{b_1} \otimes \mu_{b_2} \otimes \cdots \otimes \mu_{b_k}$. Since $I(\Pi; B) \geq \Omega(1)$, we know that $h^2(\Pi_{0^k}, \Pi_{1^k}) \geq \Omega(1)$. Now

$$I(\Pi; Z_1, \ldots, Z_k | B = 0) \geq \sum_{i=1}^{k} I(\Pi; Z_i | B = 0)$$

Lets denote our distribution of $\Pi, Z_1, \ldots, Z_k, B$ by ρ. We will tweak this distribution a little bit. Take an independent $B' \sim B_{1/2}$. All the variables are distributed the same as ρ except Z_i which is taken to be independently distributed as $\mu_{B'}$. Denote the new distribution as ρ'. It is easy to verify that

$$I(\Pi; Z_i | B = 0)_\rho \geq I(\Pi; Z_i | B = 0)_{\rho'}/2$$

This is true since in ρ, conditioned on $B = 0$, Z_i has the distribution $B_{1/2-10/\sqrt{k}}$ and in ρ' it is $B_{1/2}$ (and hence use Lemma 1.2.24). We can also see that

$$\begin{aligned} I(\Pi; Z_i | B = 0)_{\rho'} &\geq \Omega\left(k \cdot I(\Pi; B' | B = 0)_{\rho'}\right) \\ &\geq \Omega\left(k \cdot h^2(\Pi_{e_i}, \Pi_{0^k})\right) \end{aligned}$$

The first inequality is by strong data processing inequality for the binary symmetric channel and the second by Lemma 7.2.6. Now

$$\begin{aligned} I(\Pi; Z_1, \ldots, Z_k | B = 0) &\geq \sum_{i=1}^{k} I(\Pi; Z_i | B = 0) \\ &\geq \sum_{i=1}^{k} \Omega\left(k \cdot h^2(\Pi_{e_i}, \Pi_{0^k})\right) \\ &\geq \Omega\left(k \cdot h^2(\Pi_{0^k}, \Pi_{1^k})\right) \\ &\geq \Omega(k) \end{aligned}$$

The third inequality is by noting that $\Pi_{b_1,\ldots,b_k}$ satisfies a cut-and-paste property because π is a k-party protocol and hence Theorem 7.2.7 applies. □

Unknown parameter θ
Inputs: Machine i gets n samples $(X_i^{(1)}, \ldots, X_i^{(n)})$ where $X_i^{(j)} \sim \mathcal{N}(\theta, \sigma)$.

- Simultaneously, each machine i
  1. Computes $X_i = \frac{1}{\sigma\sqrt{n}} \sum_{j=1}^{n} X_i^{(j)}$
  2. If $i \leq r = O(\log(mdn/\sigma))$, machine i sends its first $O(\log(mdn/\sigma))$ bits of X_i to the coordinator (Machine 1)
  3. Else if $i > r$, machine i
     (a) Computes $R_i = X_i - \lfloor X_i \rfloor$, $R_i' = X_i + 1/5 - \lfloor X_i + 1/5 \rfloor$
     (b) Sends B_i and B_i'

$$B_i = \begin{cases} 1 & \text{with probability } R_i \\ 0 & \text{with probability } 1 - R_i \end{cases}$$

$$B_i' = \begin{cases} 1 & \text{with probability } R_i' \\ 0 & \text{with probability } 1 - R_i' \end{cases}$$

- Machine 1
  1. Computes an estimate $\gamma = \frac{\sqrt{n}}{\sigma}$ times the median of X_i's sent by the first r machines.
  2. Computes
$$T = \frac{1}{m-r} \sum_{i=r+1}^{m} B_i, T' = \frac{1}{m-r} \sum_{i=r+1}^{m} B_i'$$
  3. Returns $\frac{\sigma}{\sqrt{n}}\hat{\theta}$ where $\hat{\theta}$ is a multiple of $1/\sqrt{m-r}$ satisfying $|\gamma - \hat{\theta}| < 1/100$ and certain agreement conditions with T, T' described in the text.

Protocol 15: A simultaneous algorithm for estimating the mean of a normal distribution in the distributed setting without assuming $|\theta| \leq \sigma/\sqrt{n}$.

Bibliography

[AA03] Scott Aaronson and Andris Ambainis. Quantum search of spatial regions. In *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*, pages 200–209. IEEE, 2003.

[Abl93] Farid Ablayev. Lower bounds for one-way probabilistic communication complexity. In Andrzej Lingas, Rolf Karlsson, and Svante Carlsson, editors, *Proceedings of the 20th International Colloquium on Automata, Languages, and Programming*, volume 700 of *LNCS*, pages 241–252. Springer-Verlag, 1993.

[AG76] R. Ahlswede and P. Gacs. Spreading of sets in product spaces and hypercontraction of the markov operator. *Annals of Probability*, 4:925–939, 1976.

[AGKN13] Venkat Anantharam, Amin Gohari, Sudeep Kamath, and Chandra Nair. On maximal correlation, hypercontractivity, and the data processing inequality studied by erkip and cover. *http://arxiv.org/abs/1304.6133*, 2013.

[AS92] Noga Alon and Joel Spencer. *The Probabilistic Method*. John Wiley, 1992.

[BBCR10] Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. How to compress interactive communication. In *STOC*, pages 67–76, 2010.

[BBHT98] Michel Boyer, Gilles Brassard, Peter Hoyer, and Alain Tapp. Tight bounds on quantum searching. *Fortschritte der Physik*, 46:493–506, 1998.

[BBK+13] Joshua Brody, Harry Buhrman, Michal Kouckỳ, Bruno Loff, Florian Speelman, and Nikolay Vereshchagin. Towards a reverse newman's theorem in interactive information complexity. *Conference on Computational Complexity*, pages 24 – 33, 2013.

[BCK14] Joshua Brody, Amit Chakrabarti, and Ranganath Kondapally. Certifying equality with limited interaction. *APPROX-RANDOM*, pages 545–581, 2014.

[BCW98] Harry Buhrman, Richard Cleve, and Avi Wigderson. Quantum vs. classical communication and computation. In *STOC*, 1998.

[BG99] Sergej G Bobkov and Friedrich Götze. Exponential integrability and transportation cost related to logarithmic sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28, 1999.

[BG14] Mark Braverman and Ankit Garg. Public vs private coin in bounded-round information. *41st International Colloquium on Automata, Languages and Programming*, 2014.

[BG15] Mark Braverman and Ankit Garg. Small value parallel repetition for general games. *STOC*, pages 335–340, 2015.

[BGK+15] Mark Braverman, Ankit Garg, Young Kun Ko, Jieming Mao, and Dave Touchette. Near-optimal bounds on bounded-round quantum communication complexity of disjointness. *FOCS*, 2015.

[BGM+16] Mark Braverman, Ankit Garg, Tengyu Ma, Huy Nguyen, and David Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. *STOC*, 2016.

[BGPW13a] Mark Braverman, Ankit Garg, Denis Pankratov, and Omri Weinstein. From information to exact communication. *STOC*, 2013.

[BGPW13b] Mark Braverman, Ankit Garg, Denis Pankratov, and Omri Weinstein. From information to exact communication. *http://eccc.hpi-web.de/report/2012/171/*, 2013.

[BGPW13c] Mark Braverman, Ankit Garg, Denis Pankratov, and Omri Weinstein. Information lower bounds via self-reducibility. In *Computer Science–Theory and Applications*, pages 183–194. Springer, 2013.

[BHOS14] Fernando G.S.L. Brandao, Aram W. Harrow, Jonathan Oppenheim, and Sergii Strelchuk. Quantum conditional mutual information, reconstructed states, and state redistribution. *http://arxiv.org/abs/1411.4921*, 2014.

[BM13] Mark Braverman and Ankur Moitra. An information complexity approach to extended formulations. *STOC*, 2013.

[BMY15] Balthazar Bauer, Shay Moran, and Amir Yehudayoff. Internal compression of protocols to entropy. *RANDOM*, 2015.

[BP13] Gabor Braun and Sebastian Pokutta. Common information and unique disjointness. *FOCS*, 2013.

[BR11] Mark Braverman and Anup Rao. Information equals amortized communication. *FOCS*, 2011.

[Bra12] Mark Braverman. Interactive information complexity. In *STOC*, pages 505–524, 2012.

[BRR+09] Boaz Barak, Anup Rao, Ran Raz, Ricky Rosen, and Ronen Shaltiel. Strong parallel repetition theorem for free projection games. *RANDOM*, 2009.

[BRWY13a] Mark Braverman, Anup Rao, Omri Weinstein, and Amir Yehudayoff. Direct product via round-preserving compression. *ECCC*, 20(35), 2013.

[BRWY13b] Mark Braverman, Anup Rao, Omri Weinstein, and Amir Yehudayoff. Direct products in communication complexity. *FOCS*, 2013.

[BRWY13c] Mark Braverman, Anup Rao, Omri Weinstein, and Amir Yehudayoff. Direct products in communication complexity. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 746–755. IEEE, 2013.

[BS16] Mark Braverman and Jon Schneider. Information complexity is computable. *ICALP*, 2016.

[BSST99] Charles Bennett, Peter Shor, John Smolin, and Ashish Thapliyal. Entanglement-assisted classical capacity of noisy quantum channels. *Physical Review Letters*, 83(15):3081–3084, 1999.

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

[BT15] Mario Berta and Marco Tomamichel. The fidelity of recovery is multiplicative. *http://arxiv.org/abs/1502.07973*, 2015.

[BW12] Mark Braverman and Omri Weinstein. A discrepancy lower bound for information complexity. In *RANDOM*, pages 459–470. Springer, 2012.

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

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

[CG88] Benny Chor and Oded Goldreich. Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM Journal on Computing*, 17(2):230–261, 1988.

[CKS03] Amit Chakrabarti, Subhash Khot, and Xiaodong Sun. Near-optimal lower bounds on the multi-party communication complexity of set disjointness. *CCC*, pages 107–117, 2003.

[CKW12] Amit Chakrabarti, Ranganath Kondapally, and Zhenghui Wang. Information complexity versus corruption and applications to orthogonality and gap-hamming. *CoRR*, abs/1205.0968, 2012.

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

[CS14] André Chailloux and Giannicola Scarpa. Parallel repetition of entangled games with exponential decay via the superposed information cost. *41st International Colloquium on Automata, Languages and Programming*, 2014.

[CSUU08] Richard Cleve, William Slofstra, Falk Unger, and Sarvagya Upadhyay. Perfect parallel repetition theorem for quantum xor proof systems. *Journal of Computational Complexity*, 17(2):282–299, May 2008.

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

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

[DAW12] John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: convergence analysis and network scaling. *Automatic Control, IEEE Transactions on*, 57(3):592–606, 2012.

[DJWZ14] John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Yuchen Zhang. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. *CoRR*, abs/1405.0782, 2014.

[DM11] Irit Dinur and Or Meir. Derandomized parallel repetition via structured pcps. *IEEE Conference on Computational Complexity*, 2011.

[DP] Devdatt P. Dubhashi and Alessandro Panconesi. Concentration of measure for the analysis of randomised algorithms.

[DS14] Irit Dinur and David Steurer. Analytical approach to parallel repetition. *46th Annual Symposium on the Theory of Computing*, 2014.

[DSV14] Irit Dinur, David Steurer, and Thomas Vidick. A parallel repetition theorem for entangled projection games. *IEEE Conference on Computational Complexity*, 2014.

[EV93] Avshalom C Elitzur and Lev Vaidman. Quantum mechanical interaction-free measurements. *Foundations of Physics*, 23(7):987–997, 1993.

[FR14] Omar Fawzi and Renato Renner. Quantum conditional mutual information and approximate markov chains. *http://arxiv.org/abs/1410.0664*, 2014.

[FV02] Uriel Feige and Oleg Verbitsky. Error reduction by parallel repetition–A negative result. *Combinatorica*, 22, 2002.

[GKR14a] Anat Ganor, Gillat Kol, and Ran Raz. Exponential separation of information and communication. *FOCS*, 2014.

[GKR14b] Anat Ganor, Gillat Kol, and Ran Raz. Exponential separation of information and communication. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 176–185. IEEE, 2014.

[GKR15] Anat Ganor, Gillat Kol, and Ran Raz. Exponential separation of information and communication for boolean functions. *STOC*, 2015.

[GL10] Nathael Gozlan and Christian Léonard. Transport inequalities. a survey. *arXiv preprint arXiv:1003.3852*, 2010.

[GMN14] Ankit Garg, Tengyu Ma, and Huy Nguyen. On communication cost of distributed statistical estimation and dimensionality. *28th Annual Conference on Neural Information Processing Systems (NIPS)*, 2014.

[GPW05] Berry Groisman, Sandu Popescu, and Andreas Winter. Quantum, classical, and total amount of correlations in a quantum state. *Physical Review A*, 72, 2005.

[Gro96] Lov Kumar Grover. A fast quantum mechanical algorithm for database search. *STOC*, 1996.

[Gro09] André Gronemeier. Asymptotically optimal lower bounds on the nih-multiparty information complexity of the and-function and disjointness. *STACS*, pages 505–516, 2009.

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

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

[HOW07] Michał Horodecki, Jonathan Oppenheim, and Andreas Winter. Quantum state merging and negative information. *Communications in Mathematical Physics*, 269(1):107–136, 2007.

[HW07] Johan Håstad and Avi Wigderson. The randomized communication complexity of set disjointness. *Theory of Computing*, 3:211–219, 2007.

[Jay09a] T. S. Jayram. Hellinger strikes back: A note on the multi-party information complexity of and. *RANDOM*, pages 562 – 573, 2009.

[Jay09b] T.S. Jayram. Hellinger strikes back: A note on the multi-party information complexity of and. In Irit Dinur, Klaus Jansen, Joseph Naor, and José Rolim, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, volume 5687 of *Lecture Notes in Computer Science*, pages 562–573. Springer Berlin Heidelberg, 2009.

[JPY14] Rahul Jain, Attila Pereszlényi, and Penghui Yao. A parallel repetition theorem for entangled two-player one-round games under product distributions. *IEEE Conference on Computational Complexity*, 2014.

[JRS03] Rahul Jain, Jaikumar Radhakrishnan, and Pranab Sen. A lower bound for bounded round quantum communication complexity of set disjointness. *FOCS*, pages 220 – 229, 2003.

[JSWZ13] Rahul Jain, Yaoyun Shi, Zhaohui Wei, and Shengyu Zhang. Efficient protocols for generating bipartite classical distributions and quantum states. *IEEE Transactions of Information Theory*, 59(8):5171–5178, 2013.

[Kla98] Hartmut Klauck. Lower bounds for computation with limited nondeterminism. *Computational Complexity*, pages 141–152, 1998.

[KLL+12a] 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.

[KLL+12b] Iordanis Kerenidis, Sophie Laplante, Virginie Lerays, Jérémie Roland, and David Xiao. Lower bounds on information complexity via zero-communication protocols and applications. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 500–509. IEEE, 2012.

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

[KNTSZ01] Hartmut Klauck, Ashwin Nayak, Amnon Ta-Shma, and David Zuckerman. Interaction in quantum communication and the complexity of set disjointness. *STOC*, 2001.

[KRW95] 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.

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

[KSDW04] H. Klauck, R. Spalek, and R. De Wolf. Quantum and classical strong direct product theorems and optimal time-space tradeoffs. In *Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on*, pages 12–21. IEEE, 2004.

[KV11] Julia Kempe and Thomas Vidick. Parallel repetition of entangled games. *43rd annual ACM symposium on Theory of computing*, 2011.

[KVW14] Ravi Kannan, Santosh Vempala, and David P. Woodruff. Principal component analysis and higher correlations for distributed data. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 1040–1057, 2014.

[KWHZ95] Paul Kwiat, Harald Weinfurter, Thomas Herzog, and Anton Zeilinger. Interaction-free measurement. *Physical Review Letters*, 74(24), 1995.

[LBKW14] Yingyu Liang, Maria-Florina Balcan, Vandana Kanchanapally, and David P. Woodruff. Improved distributed principal component analysis. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3113–3121, 2014.

[Led01] Michel Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2001.

[LR73] Elliot Lieb and Mary Ruskai. Proof of the strong subadditivity of quantum mechanichal entropy. *Journal of Mathematical Physics*, 14:1938–1941, 1973.

[LRS15] James R. Lee, Prasad Raghavendra, and David Steurer. Lower bounds on the size of semidefinite programming relaxations. *STOC*, 2015.

[LSLT15] Jason D Lee, Yuekai Sun, Qiang Liu, and Jonathan E Taylor. Communication-efficient sparse regression: a one-shot approach. *arXiv preprint arXiv:1503.04337*, 2015.

[MI] Nan Ma and Prakash Ishwar. Personal communication.

[MI08] Nan Ma and Prakash Ishwar. Two-terminal distributed source coding with alternating messages for function computation. *IEEE International Symposium on Information Theory*, pages 51–55, 2008.

[MI09] Nan Ma and Prakash Ishwar. Infinite-message distributed source coding for two-terminal interactive computing. *Proceedings of the 47th annual Allerton Conference on Communication, Control and Computing*, pages 1510–1517, 2009.

[MI11] Nan Ma and Prakash Ishwar. Some results on distributed source coding for interactive function computation. *IEEE Transactions on Information Theory*, 57(9):6180–6195, 2011.

[Mos14] Dana Moshkovitz. Parallel repetition of fortified games. *Electronic Colloquium on Computational Complexity (ECCC)*, 2014.

[New91] Ilan Newman. Private vs. common random bits in communication complexity. *Information Processing Letters*, 39(2):67–71, 31 July 1991.

[NW93a] N. Nisan and A. Wigderson. Rounds in communication complexity revisited. *SIAM Journal on Computing*, 22(1):211–219, 1993.

[NW93b] Noam Nisan and Avi Wigderson. Rounds in communication complexity revisited. *SIAM Journal on Computing*, 22(211-219), 1993.

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

[Orl91] A. Orlitsky. Worst-case interactive communication. ii. two messages are not optimal. *Information Theory, IEEE Transactions on*, 37(4):995–1005, 1991.

[Pan12] Denis Pankratov. *Direct sum questions in classical communication complexity.* PhD thesis, 2012.

[PRV01] Stephen J. Ponzio, Jaikumar Radhakrishnan, and S. Venkatesh. The communication complexity of pointer chasing. *Journal of Computer and System Sciences*, 62:323–355, 2001.

[Rag16] Maxim Raginsky. Strong data processing inequalities and Φ-sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 2016.

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

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

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

[Raz02] Alexander A. Razborov. Quantum communication complexity of symmetric predicates. *Izvestiya of the Russian Academy of Science, Mathematics*, 67, 2002.

[Raz08] Ran Raz. A counterexample to strong parallel repetition. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*. IEEE Computer Society, 2008.

[Raz11] Ran Raz. A counterexample to strong parallel repetition. *SIAM Journal on Computing*, 40(3):771–777, 2011.

[RR12] Ran Raz and Ricky Rosen. A strong parallel repetition theorem for projection games on ex- panders. *IEEE Conference on Computational Complexity*, pages 247–257, 2012.

[Sch95] Benjamin Schumacher. Quantum coding. *Physical Review A*, 51(2738 - 2747), 1995.

[SD15] Jacob Steinhardt and John C. Duchi. Minimax rates for memory-bounded sparse linear regression. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, pages 1564–1587, 2015.

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

[Sha13] Ronen Shaltiel. Derandomized parallel repetition theorems for free games. *Complexity*, 22(3):565–594, 2013.

[Sha14] Ohad Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 163–171. Curran Associates, Inc., 2014.

[She07] Alexander A. Sherstov. The pattern matrix method for lower bounds on quantum communication. *STOC*, 2007.

[She12] Alexander A. Sherstov. Strong direct product theorems for quantum communication and query complexity. *SIAM Journal on Computing*, 41(5):1122–1165, 2012.

[She14] Alexander A. Sherstov. Communication lower bounds using directional derivatives. *Journal of the ACM*, 61(6):1–71, 2014.

[SSZ14] Ohad Shamir, Nathan Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *Proceed-*

ings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, pages 1000–1008, 2014.

[ST13] Mert Saglam and Gábor Tardos. On the communication complexity of sparse set disjointness and exists-equal problems. *FOCS*, 2013.

[Ter72] Frode Terkelsen. Some minimax theorems. *Mathematica Scandinavica*, 31:405–413, 1972.

[TL] Dave Touchette and Mathieu Laurière. Personal communication.

[Tou15] Dave Touchette. A new, fully quantum notion of information complexity, and an application to direct sum for bounded round quantum communication complexity. *STOC*, 2015.

[TWZ14] Madhur Tulsiani, John Wright, and Yuan Zhou. Optimal strong parallel repetition for projection games on low threshold rank graphs. *ICALP*, 2014.

[Ver94] Oleg Verbitsky. Towards the parallel repetition conjecture. In *Structure in Complexity Theory Conference*, pages 304–307, 1994.

[Wat13] John Watrous. Theory of quantum information. *Lecture notes*, 2013.

[Wil13] Mark Wilde. Quantum information theory. *Cambridge University Press*, June 10, 2013.

[WZ12] David P. Woodruff and Qin Zhang. Tight bounds for distributed functional monitoring. *STOC*, 2012.

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

[Yao93] Andrew Chi-Chih Yao. Quantum circuit complexity. *FOCS*, 1993.

[YD09] Jon Yard and Igor Devetak. Optimal quantum source coding with quantum information at the encoder and decoder. *IEEE Transactions on Information Theory*, 55(11):5339–5351, 2009.

[Zal99] Christof Zalka. Grover's quantum searching algorithm is optimal. *Physical Review A*, 60(4), 1999.

[ZDJW13] Yuchen Zhang, John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *NIPS*, pages 2328–2336, 2013.

[ZDW13] Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14(1):3321–3363, 2013.

[ZX15] Yuchen Zhang and Lin Xiao. Communication-efficient distributed optimization of self-concordant empirical loss. *CoRR*, abs/1501.00263, 2015.