Lower Bounds for Interactive Compression and Linear Programs

Makrand Sinha

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Anup Rao, Chair

Paul Beame

Thomas Rothvoß

Program Authorized to Offer Degree:
Computer Science & Engineering

University of Washington

Abstract

Lower Bounds for Interactive Compression and Linear Programs

Makrand Sinha

Chair of the Supervisory Committee:
Associate Professor Anup Rao
Paul G. Allen School of Computer Science & Engineering

This thesis proves concrete lower bounds related to two computational settings — interactive compression and linear programs.

The first part of this thesis studies data compression in interactive models. In the classical setting of data compression, we have a sender that wishes to convey a message to a receiver by transmitting as few bits as possible. Since the pioneering work of Shannon (*The Bell System Technical Journal*, 27 (4), 1948), it has been known how to encode the messages so that the number of bits sent is equal to the amount of information contained in the message, which turns out to be optimal. The setting of interactive compression is a generalization of the classical setting to the case when there are multiple parties interacting with each other and they possess partial knowledge about each other's messages. In this thesis we further the understanding of interactive compression in the settings of communication complexity and streaming algorithms. The first part of this thesis proves the following lower bounds for interactive compression:

Exponential Separation between Information and Communication. Given two parties with correlated inputs X and Y who are communicating to compute a function $f(X, Y)$, can we compress the communication to the amount of information exchanged? This thesis introduces a new technique, the notion of *fooling distributions*, to prove that information can be exponentially smaller than communication.

Compression and Direct-sum for Streaming. We consider a direct-sum question for streaming

algorithms: suppose a streaming algorithm processes k input streams that arrive in parallel; does it need k times more memory to perform the computation? This thesis introduces a notion of information cost for streaming algorithms and shows that compressing the memory of a streaming algorithm to its information cost would imply that no streaming algorithm can perform significantly better than using k times more memory on k parallel streams.

The second part of this thesis studies linear programs for combinatorial optimization problems. Many combinatorial optimization problems can naturally be expressed as a linear program, albeit with exponentially many inequalities. It turns out that in certain instances, we can significantly reduce the size of an exact (or approximate) linear program for an optimization problem by using some auxiliary variables. Such linear programs are known as *extended formulations* or *higher dimensional lifts*. They also capture the best known algorithms for many combinatorial optimization problems, as well as the LP hierarchies (Sherali-Adams, Lovász-Schrijver). In recent years, strong lower bounds have been proven on the size of extended formulations for many well-known combinatorial optimization problem. This thesis proves the following result about extended formulations:

Approximate Extended Formulations for the Matching Polytope. We prove a tight lower bound for approximate extended formulations for the Matching polytope. In the course of proving this lower bound, we also obtain a tight non-negative rank lower bound for a family of matrices known as lopsided unique disjointness.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This journey has not been easy, and at some points perhaps exceedingly difficult, but maybe that is why finishing it makes it so much more meaningful. When I started, I was naive about a lot of things — what it meant to do research, what research directions are important, how to enjoy the process, and my whole worldview has shifted completely in that regard. At the end of this journey, what matters most to me is the desire to understand. Perhaps this was already an innate desire, but I have had to find it again through this journey and I hope I can carry it with me into the future.

Many people have contributed so much to this journey in innumerable ways. They have helped me when I was struggling, given advice when I needed it, cared for me in ways that I could not have imagined. Without them I don't know how I would have made it here.

I have a lot of gratitude to my advisor Anup for teaching me a lot, for imparting his zest for explaining and simplifying things onto me, for teaching me what is important in research, for giving me unfettered freedom and for believing in me.

Other faculty members at UW CSE helped me a lot both directly and indirectly. Paul gave me the most useful and practical advice about grad school and research, served on all of my committees, read this thesis thoroughly and vouched for me. Thomas listened to me when I was stuck, served on my committee and wrote a letter on my behalf. Anna, James, Shayan and other faculty members of the theory group gave me useful advice and taught me a lot over the years. Dan, Chris and Sreeram helped me out by being on my committee.

My fellow grad students provided me with countless hours of learning, moral support and com-

panionship among the ups and downs of grad school. I spent countless hours with Siva discussing problems, writing papers, ruminating, gossiping, climbing and I'm really glad for him being there. Widad, Cyrus, Becca, Xin, Rahul and all the other theory group students made it fun to learn and think. Ray, Dimi, Rahul were lifting bros and lifelong friends.

Improv, theater and dancing were a big part of my life in Seattle and will remain with me forever. My improv and theater friends, Jamie, Hillary, Sid and countless others, made me a part of the family and taught me countless things about life. My dance instructors Gabi, Gwen and Josh not only taught me how to dance, but how to learn.

I made a lot of friends in Seattle and I will dearly miss them being a part of my life. William and Ka were there to share my love for food. My roommates Dave, Lexa, Amanda, David, Kristin, Sush, Samrat, Kevin, Agraj and YG tolerated my idiosyncrasies and gave me a home worth coming back to. My friends Carla and Adam supported me through difficult times, shared my embarrassing moments and were true friends. Megan and Huck cared about me and helped me discover life. Mary, Kavita, Jack, Maria, Megha, Erin, Michelle and Adam helped me find my place in the world. Mike listened and shared his curiosity about the whole world with me. Garima and Vanessa helped me become a better person. Allison listened to me, motivated me and climbed with me. There were many others who I could not list here, but they showed me what kindness really meant and I will always remember them warmly.

Looking back, I cherish the memories I have made living in Seattle. I have a lot to be grateful for and I know that I will carry this gratitude with me forever.

Pour mes amis, pour ma famille et
pour moi quand j'etais petit garçon

Nur für Verrückte! Der Eintritt kostet den Verstand!

— Hermann Hesse, *Der Steppenwolf*

Prologue

Human beings have known how to compute for millennia — the ancient Sumerians wrote multiplication tables on clay tablets and solved geometric and arithmetic problems 5000 years ago, mathematicians in Egypt knew how to solve quadratic equations in 1800 BC; history abounds in many such extraordinary computational feats — but what does it really mean to compute? We all know intuitively from early schooling experiences, that the task of addition seems much easier than long division, but how do we explain that certain tasks are easier computationally than others?

In 1936, Alan Turing introduced a way of modeling computation mathematically by defining what is now called a *Turing Machine* [Tur36]. Not only did this lay the foundations of modern computing but it also gave us a way to answer questions about the process of computation itself, such as the ones posed before. A Turing Machine reads an input of length n from a tape, such as two numbers of $n/2$ digits that we want to add, runs for a certain number of steps that depends on n and outputs the result of the computation. One can think of a computational task as more challenging if the Turing Machine takes more steps to compute the answer.

A Turing Machine can simulate any computation that could be implemented on a physical computer. Thus, studying computational tasks in the model of Turing Machines proves fundamental limitations on the time required to perform computation in the real world. If the Turing Machine took 2^n steps to perform a computation on an input of length n, then when the input becomes too large, such as $n = 1000$, the Turing Machine would take 2^{1000} steps, an astronomical number. Were we to implement such a computation on a physical computer, even assuming we could achieve the physical limits of hardware, running 2^{1000} steps would still take trillions of years, making it completely

intractable.

Since the time of Turing, the landscape of computing has expanded vastly. Time is not the only bottleneck anymore, other resources such as the amount of memory or bandwidth available, may also render a computation intractable for all practical purposes, as data is often spread across millions of servers which are communicating with each other. It is becoming ever more important to understand what computational problems are intractable in these new models of computing.

This thesis tries to answer some such fundamental questions about limitations of computation.

1 Introduction

A fundamental question in computer science is to characterize which computational problems are intractable. This question came to the forefront in the 1960s and in the last 60 years since then, a rich theory has been developed to understand the relative difficulty of computational problems. Such a way of classifying problems has brought us a lot of understanding. For many problems of practical interest, we are now able to make *conditional* statements — we can say that if SAT is intractable, then the problem of interest is intractable as well. On the other hand, all the evidence so far suggests that many of these fundamental problems ought to take exponential computational time. However, despite our best efforts, there has been very little progress towards proving *unconditional* limitations on algorithms for many problems we care about. For example, in the model of boolean circuits, at this point we are even unable to show that SAT, or any explicit function for that matter, requires more than $4n$ size.

This thesis is concerned with proving concrete limitations on algorithms. However, given that we often do not know how to prove unconditional lower bounds for many problems of interest in the most general models of computation, we focus our efforts on proving limitations on different computational models, which still capture, or have applications, to a broad class of interesting algorithms. In this thesis, we focus on two different computational settings — *interactive models* and *linear programming*. The common theme uniting these two seemingly disparate models in this thesis is the use of information theoretic methods for proving lower bounds.

The first part of this thesis looks at the setting of interactive computational models where the inputs for the computational tasks are distributed across multiple parties. Such computational models

capture new types of algorithms that have arisen due to recent technological developments. Companies like Google and Facebook have massive amounts of data which is stored in data centres spread across the world. In these settings, we not only care about the running time of the algorithm but we also care about other resources such as communication between different machines, or the memory used to carry out the computation, since they are often significant bottlenecks. In some cases, such models also have applications to proving lower bounds for an interesting class of classical algorithms (see Chapter 8 in the textbook [RY18]). We consider the question of *data compression* in these interactive models and prove some fundamental lower bounds.

In the second part of this thesis, we look at linear programs for combinatorial optimization problems. A combinatorial optimization problem can often be written as a linear program of the following form:

$$\begin{aligned} &\text{maximize } w^T x \\ &\text{subject to } x \in P, \end{aligned}$$

where w is a cost function and P is a polytope that captures the convex hull of the feasible integer solutions to the problem. The complexity of finding an optimum solution to a linear program usually depends on the size of the linear program, which for linear programs of the above form is characterized by the number of inequalities needed to describe the polytope P. For many combinatorial optimization problems, like finding the shortest Hamiltonian tour in a graph or finding the maximum weight matching in a graph, representing them with linear programs of the above form often requires exponential sized linear programs in terms of the instance size. A very natural idea for reducing the size of such linear programs is to employ what are called *linear extended formulations* or *higher dimensional lifts* of polytopes. We prove lower bounds on the size of such higher dimensional lifts in the second part of this thesis.

1.1 Contributions of this Thesis

1.1.1 Interactive Compression

The first part of this thesis deals with data compression in an interactive setting. In the classical setting of data compression, first introduced by Shannon [Sha48], we have a sender who wants to convey a message to a receiver by sending as few bits as possible. This setting is not interactive, as the receiver does not send messages back to the sender. In this setting, it is known how to compress messages optimally [Sha48, Huf52] — we know how to encode messages so that their length is essentially the same as the amount of information they carry. The amount of information that a message carries is the limit of what is achievable, as Shannon already showed in his seminal paper [Sha48]. For example, if the message consists of n random bits, then the sender cannot send a message containing fewer than n bits to convey the message. On the other hand, if the message consists of one random bit followed by $n - 1$ zeros, then the sender only needs to send the first bit of the message.

In the setting of interactive compression, we may hope to achieve similar compression guarantees even when there are multiple parties who are interacting with each other. Unfortunately, for these interactive models, it is often not clear how to even define the proper notion of information contained in the messages since the parties may have partial knowledge about each others' messages. In some interactive settings, even if one can define a natural information measure, the techniques for classical data compression are usually not applicable. These are significant roadblocks in trying to generalize Shannon's theory to interactive settings.

In this thesis we further the understanding of the interactive compression question in two different computational settings — two-party communication complexity and streaming algorithms. Apart from generalizing classical data compression, the interactive compression question in both these models is also closely connected to the question of proving lower bounds for algorithms in these models and the results proved here also translate to such lower bounds. We prove the following about interactive compression:

Exponential Separation of Information and Communication

In the setting of two-party communication complexity, there are two parties who hold inputs x and y and want to compute a boolean function $f(x, y)$. In order to achieve this, they have to communicate with each other by exchanging messages through a pre-agreed protocol. The computational resource we care about in this setting is the amount of communication required to compute the function f. In Chapter 4 of this thesis, we look at the interactive compression question for two-party communication complexity.

To generalize Shannon's theory to this setting, we need a notion that captures the amount of "information contained in the messages". Developing the ideas introduced in [CSWY01, BYJKS02], Barak, Braverman, Chen and Rao [BBCR13] suggested that the right measure for the information contained in the messages in this model is the amount of information that the messages reveal to the parties about each other's input. This is termed as the *internal information cost* of the communication protocol, and a central question in this area is whether any protocol with communication C and internal information cost I can be *simulated* with communication $O(I)$. Alternately stated, whether protocols with internal information cost I can be compressed to communication $O(I)$.

For the most general scenario, the best compression known is with $O(\sqrt{IC}\log C)$ bits of communication [BBCR13], and with communication $2^{O(I)}$ if we do not want a dependence on C [Bra12]. However, optimally one could hope for compression with $O(I)$ bits. To prove that optimal compression is impossible in this model was also a challenge since it turned out that all known methods for proving communication lower bounds also gave lower bounds on the internal information cost [KLL$^+$12]. In a breakthrough paper, Ganor, Kol and Raz [GKR16] introduced a new method for proving communication lower bounds and proved that compression with $2^{O(I)}$ bits of communication is the best possible, if we do not want a dependence on the communication C.

In Chapter 4, we introduce the notion of *fooling distributions*, another method for proving communication lower bounds that allows us to prove a separation between information and communication. Our method allows us to give a simpler and more intuitive proof of the result of [GKR16]. A technical lemma we prove here also allows us to give a simpler proof of the lower bound on the

randomized communication complexity of the Greater-Than function.

It turns out that the interactive compression question for two-party communication is also equivalent to the *direct sum* question in communication complexity which asks if k independent copies of a communication task requires k times the communication (see [BR11]). As such the lower bound for compression proved in Chapter 4 as well the work [GKR16] also gives us a communication task where one can compute t copies in a way that is significantly cheaper than expending k times the communication required to compute a single copy of the same task.

Compression and Direct Sum for Streaming Algorithms

In the setting of streaming algorithms, the algorithm receives its input as a sequence of updates $x_1, \ldots, x_n$ arriving sequentially in time. The complexity measure of interest is the memory needed to carry out the computation, where the memory used at time t is the number of bits stored by the algorithm after reading the inputs $x_1, \ldots, x_t$. One can view this setting as an interactive setting by viewing the update at any time as a message sent by an interacting party.

In the setting of communication complexity above, we tried to capture the information cost of a protocol with respect to the amount of information that the messages contain about each party's input. For the case of streaming algorithms, it is natural to ask if information cost can be defined in terms of the memory contents of the algorithm, and whether it is possible to compress the memory used by streaming algorithms to its information cost.

In Chapter 5, we introduce several analogues of information cost for streaming algorithms. We also consider a direct-sum question for streaming algorithms where there are k independent input streams arriving in parallel and we want to know if it is possible to carry out this computation with less than k times the memory required for a single stream. In Chapter 5, we also show that if one can compress the memory used by a streaming algorithm to its information cost, then an optimal direct-sum statement would hold.

1.1.2 Linear Programs for Combinatorial Optimization Problems

The second part of this thesis explores linear programs for combinatorial optimization problems. As discussed before, many combinatorial optimization problems can be naturally modeled exactly or approximately with a linear program of the form

$$\begin{aligned} &\text{maximize } w^T x \\ &\text{subject to } x \in P, \end{aligned}$$

where w is a cost function and P is a polytope that captures the convex hull of the feasible integer solutions to the problem. For example, for the Traveling Salesman Problem on undirected graphs with n vertices, one wants to minimize a linear function (the length of the tour) over the polytope $P \subseteq \mathbb{R}^{\binom{[n]}{2}}$ that is the convex hull of all Hamiltonian cycles in the complete graph K_n.

Even though an optimal solution of a linear program can be found in time that is polynomial in the number of bits needed to describe the linear program, in general, the time to solve it scales with the number of inequalities in the system. Given a linear program, a very general idea to reduce the number of inequalities is to use auxiliary variables so that the new system of inequalities, when projected on the original variables, is equivalent. This is what is called an *extended formulation*. In geometric terms, adding auxiliary variables corresponds to a higher dimensional polytope that projects to the original polytope under some linear map. For this reason, extended formulations are also sometimes referred to as *higher dimensional lifts*. Adding auxiliary variables, or equivalently lifting to higher dimensions, helps in reducing the size of the linear program because the higher dimensional polytope could have fewer *facets* and hence to describe it as a linear program one needs to write down fewer inequalities (see Figure 1.1.1).

Extended formulations can sometimes lead to an exponential reduction in the size of the original linear program and so the algorithms that optimize over an extended formulation capture a powerful class of linear programming based algorithms, including linear programming hierarchies such as Sherali-Adams or Lovász-Schrijver. If we can unconditionally prove that extended formulations for natural polytopes P for NP-hard problems require exponential size, this is a step towards proving

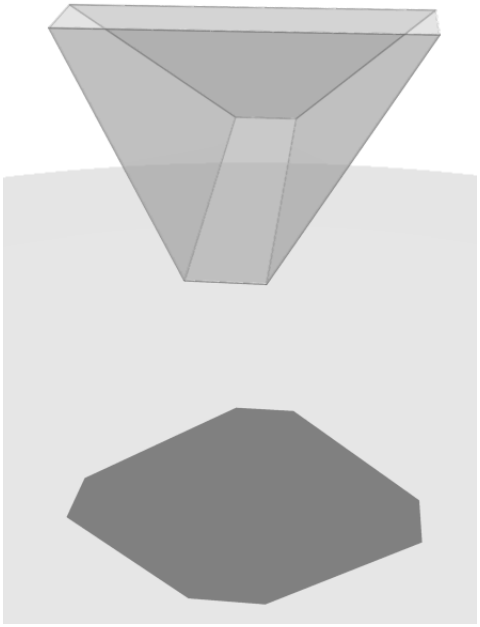


Figure 1.1.1: An example of an extended formulation

$\mathsf{P} \neq \mathsf{NP}$ by ruling out an important class of linear programming based algorithms.

Recently, using ideas from communication complexity and information theory, exponential lower bounds have been proven for the size of exact and even approximate extended formulations for various natural polytopes for well-known combinatorial optimization problems. In this thesis, we look at extended formulations for the Matching Polytope.

Approximate Extended Formulations for the Matching Polytope

A matching in a graph is a set of edges that have no common vertices. The matching problem asks to find a maximum weighted matching in a given graph G that has weights w on the edges. Denoting by $\mathbb{1}_M \in \mathbb{R}^{\binom{[n]}{2}}$ the indicator vector for a matching M in the complete graph K_n, the matching problem

can be naturally written as the following linear program

$$\text{maximize } w^T x$$
$$\text{subject to } x \in P_{MAT}(n),$$

where $P_{MAT}(n)$ is the matching polytope defined as

$$P_{MAT}(n) = \text{conv}\left\{ \mathbb{1}_M \in \mathbb{R}^{\binom{[n]}{2}} \;\middle|\; M \subseteq \binom{[n]}{2} \text{ is a matching in } K_n \right\}.$$

The Matching Polytope is a fundamental object in the field of combinatorial optimization and it is well known (see [Edm65]) that to describe the Matching Polytope by a linear program, one needs $2^{\Theta(n)}$ many inequalities. It was a long-standing open problem to show that arbitrary extended formulations for Matching Polytope must also be of exponential size. In a breakthrough work, Rothvoß [Rot17] showed that any exact extended formulation for $P_{MAT}(n)$ must have $2^{\Omega(n)}$ size.

Building on this, Braun and Pokutta [BP15] showed that any extended formulation that gives a $(1+\frac{1}{n})$ approximation requires $2^{\Omega(n)}$ size, which as noted in [Rot17] implies a $2^{\Omega(1/\epsilon)}$ bound for a $(1+\epsilon)$ approximation when $\epsilon \geq \frac{2}{n}$. This rules out the possibility of an $FPTAS$-style extended formulation (size polynomial in both n and $\frac{1}{\epsilon}$) for the matching problem, but for large values of ϵ, these lower bounds are trivial.

In Chapter 7 of this thesis, we prove a lower bound for the Matching Polytope by proving that all $(1+\epsilon)$ approximating extended formulations must be of $\approx \binom{n}{1/\epsilon}$ size for all $\epsilon \geq \frac{2}{n}$. This lower bound is tight for all parameters $\epsilon > 0$ as there is a simple linear program that matches this. In the course of proving a lower bound for the Matching Polytope, we also prove a tight lower bound on the *non-negative rank* of a family of matrices known as lopsided unique disjointness.

1.2 Organization

Most of this thesis relies on tools from information theory. In Chapter 2, we collect the basic information theoretic facts that will be used throughout this thesis.

Part I of this thesis considers the setting of interactive models of computation. In Chapter 3, we motivate the question of interactive compression and provide an extensive background. Chapter 4

proves the results separating information and communication. Chapter 5 considers the question of compression and direct sum for streaming algorithms. This part of the thesis is based on the following works:

- Anup Rao and Makrand Sinha. Simplified Separation of Information and Communication. *Electronic Colloquium on Computational Complexity (ECCC)*, 22:57, 2015
- Sivaramakrishnan Natarajan Ramamoorthy and Makrand Sinha. On the Communication Complexity of Greater-Than. In *53rd Annual Allerton Conference on Communication, Control and Computing*, pages 442–444, 2015
- Anup Rao and Makrand Sinha. A Direct-Sum Theorem for Read-Once Branching Programs. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2016, September 7-9, 2016, Paris, France*, pages 44:1–44:15, 2016

Part II of this thesis looks at the setting of linear programs for combinatorial optimization problems. In Chapter 6, we describe the motivations behind extended formulations and give a survey of recent works. Chapter 7 proves the lower bound for approximate extended formulations for the Matching Polytope. This part of the thesis is based on the following work:

- Makrand Sinha. Lower Bounds for Approximating the Matching Polytope. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pages 1585–1604, 2018

2 Information Theory Preliminaries

Most of the proofs in this thesis rely heavily on information theory. In this chapter, we set up the notation and review the basics of information theory that will be applicable to all the chapters. Some information theoretic lemmas that are required for applications will be proved in the chapters where they are used. We will usually omit the proofs for basic facts here and refer the reader to the book by Cover and Thomas [CT06].

2.1 Probability Spaces, Random Variables and Markov Chains

Throughout this thesis, $\log$ (resp. $\ln$) denotes the logarithm taken in base two (resp. base e). We use $[k]$ to denote the set $\{1, 2, \ldots, k\}$ and $[k]^{<n}$ to denote the set of all strings of length less than n over the alphabet $[k]$, including the empty string. The notation $|z|$ denotes the length of the string z.

Random variables are denoted by capital letters (e.g. A) and values they attain are denoted by lower-case letters (e.g. a). Events in a probability space will be denoted by calligraphic letters (e.g. $\mathcal{E}$). Given $a = (a_1, a_2, \ldots, a_n)$, we write $a_{\leq i}$ to denote $a_1, \ldots, a_i$. We define $a_{<i}$ and $a_{>i}$ similarly. We write a_{-i} to denote $a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n$ and a_S to denote the projection of a to the coordinates specified in the set $S \subseteq [n]$.

Given a probability space p and a random variable A in the underlying sample space, we use the notation $p(A)$ to denote the probability distribution of the variable A in the probability space p. We will often consider multiple probability spaces with the same underlying sample space, so for example $p(A)$ and $q(A)$ will denote the distribution of the random variable A under the probability spaces p and q respectively with the underlying sample space of p and q being the same. We write $p(A|b)$ to

denote the distribution of A conditioned on the event $B = b$. We use $p(a)$ and $p(a|b)$, respectively, to denote the probability of the event $A = a$ in the distribution $p(A)$ and $p(A|b)$. Given a distribution $p(A, B, C, D)$, we write $p(A, B, C)$ to denote the marginal distribution on the variables A, B, C. We often write $p(AB)$ instead of $p(A, B)$ for conciseness of notation. Similarly, $p(a, b, c)$ will denote the probability according to the marginal distribution $p(A, B, C)$ and we will often write it as $p(abc)$.

If $\mathcal{W}$ is an event, we write $p(\mathcal{W})$ to denote its probability according to p. Given a probability space p and a random variable A, when we write $A \in \mathcal{W}$ for an event $\mathcal{W}$, we only consider events in the space of values taken by the variable A.

Given a fixed value c, we denote by $\mathbb{E}_{p(b|c)}[g(a, b, c)] := \sum_b p(b|c) \cdot g(a, b, c)$, the expected value of the function $g(a, b, c)$ under the distribution $p(B|c)$. If the probability space p is clear from the context, then we will just write $\mathbb{E}_{b|c}[g(a, b, c)]$ to denote the expectation. For a boolean function $h(a, b)$ and a probability distribution $p(A, B)$, denoting by $\mathbf{1}[h(a, b) = 0]$ the indicator function for the event $h(a, b) = 0$, we write $p(h = 0) := \mathbb{E}_{p(ab)}[\mathbf{1}[h(a, b) = 0]]$ as the probability that h is 0 under inputs drawn from p.

We say that a random variable A determines another random variable B if there is a function h such that $h(A) = B$. We write $A - M - B$ as a shorthand to say that the random variables A, M, and B form a *Markov chain*, or in other words, A and B are independent given M: $p(amb) = p(m) \cdot p(a|m) \cdot p(b|m)$ for every a, b, m. In some cases, we will have Markov chains where M determines B. To emphasize this we will write this Markov chain as $A - M \to B$. For brevity we will write $A|R - M|R - B|R$ to assert that $p(amb|r)$ is a Markov chain for every r.

To get familiar with the notation, it is worthwhile to consider the following example. Let $A \in \{0, 1\}^2$ be a uniformly distributed random variable in a probability space p. Then, $p(A)$ is the uniform distribution on $\{0, 1\}^2$ and if $a = (0, 0)$, $p(a) = 1/4$. Let A_1 and A_2 denote the first and second bits of A: if $B = A_1 + A_2 \bmod 2$, then when $b = 1$, $p(A|b)$ is the uniform distribution on $\{(0, 1), (1, 0)\}$. If $a = (1, 0)$, and $b = 1$, then $p(a|b) = 1/2$, and $p(a, b) = 1/4$. If $\mathcal{E}$ is the event that $A_1 = B$, then $p(\mathcal{E}) = 1/2$. Let $q(A) = p(A|\mathcal{E})$, then $q(A)$ is the uniform distribution on $\{(0, 0), (1, 0)\}$ and $q(A_2)$ is the distribution over the sample space $\{0, 1\}$ which takes the value 0 with probability 1.

2.2 Statistical Distance

Statistical distance is a very useful way to measure how far apart two probability distributions are. For two distributions $p(A)$ and $q(A)$, the *statistical distance* $|p(A) - q(A)|$ between them is defined as

$$|p(A) - q(A)| = \max_{\mathcal{Q}} \left(p(A \in \mathcal{Q}) - q(A \in \mathcal{Q})\right)$$

where $\mathcal{Q}$ is an event.

There are several alternative ways of defining the statistical distance, as given by the following proposition.

Proposition 2.1. $|p(A) - q(A)| = \frac{1}{2}\sum_a |p(a) - q(a)| = \sum_{a:p(a)>q(a)}(p(a) - q(a))$.

We say that distributions $p(A)$ and $q(A)$ are ε-close if $|p(A) - q(A)| \leq \epsilon$ and we write it as $p(A) \overset{\epsilon}{\approx} q(A)$. Intuitively what this means is that the probability of any event under the two distributions can differ at most by an additive factor of ε.

The following basic proposition characterizes the statistical distance of conditional distributions.

Proposition 2.2. *If $p(AB), q(AB)$ are such that $p(A) = q(A)$, then*

$$|p(B) - q(B)| = \mathbb{E}_{p(a)}\left[|p(B|a) - q(B|a)|\right].$$

2.3 Entropy and Mutual Information

For a discrete random variable A in a probability space p, the entropy of A is defined as

$$\mathbf{H}(A) = \mathbb{E}_{p(a)}\left[\log \frac{1}{p(a)}\right].$$

In this thesis, we shall often work with multiple probability spaces over the same sample space. To avoid confusion, we shall explicitly write $\mathbf{H}_p(A)$ to specify the probability space p being used for computing the entropy if it is not clear from the context.

The following proposition gives a useful upper bound on entropy in terms of the size of the support of the random variable:

Proposition 2.3. $\mathbf{H}(A) \leq \log |\mathsf{supp}(A)|$.

For any two random variables A and B in a probability space p, the entropy of A conditioned on B is defined as

$$\mathbf{H}(A|B) = \mathbb{E}_{p(b)}[\mathbf{H}(A|B = b)].$$

The following relation between the conditional entropy $\mathbf{H}(A|B)$ and the entropy $\mathbf{H}(A)$ is easy to see:

Proposition 2.4. $\mathbf{H}(A|B) \leq \mathbf{H}(A)$ *where the equality holds if and only if A and B are independent.*

A subset $\mathcal{S} \subseteq \{0,1\}^*$ is called *prefix-free* if no string in $\mathcal{S}$ is a prefix of another string in $\mathcal{S}$. For any discrete random variable A, a *prefix-free encoding* of A is a function $\ell : \mathsf{supp}(A) \to \{0,1\}^*$ such that the image of ℓ is a prefix-free set.

Proposition 2.5 (Huffman Encoding)**.** *Let A and B be random variables where A is discrete. Then, there exists a prefix-free encoding $\ell : \mathsf{supp}(A) \to \{0,1\}^*$ satisfying*

$$\mathbb{E}_{xy}[|\ell(x)| \mid B = y] \leq \mathbf{H}(A|B) + 1.$$

The mutual information between A and B is defined as

$$\mathbf{I}(A : B) = \mathbf{H}(A) - \mathbf{H}(A|B) = \mathbf{H}(B) - \mathbf{H}(B|A).$$

Similarly, the conditional mutual information $\mathbf{I}(A : B|C)$ is defined to be

$$\mathbf{I}(A : B|C) = \mathbf{H}(A|C) - \mathbf{H}(A|BC).$$

We shall explicitly write $\mathbf{I}_p(A : B|C)$ to specify the probability space p being used for computing the mutual information if it is not clear from the context.

If A and B are independent then $\mathbf{I}(A : B) = 0$. We can always bound the mutual information by the entropy of the variables involved.

Proposition 2.6. *If $A \in \{0,1\}^\ell$, then*

$$0 \leq \mathbf{I}(A : B) \leq \min\{\mathbf{H}(A), \mathbf{H}(B)\} \leq \ell.$$

The binary entropy function[1] is defined to be $\mathsf{h}(x) := -x \log x - (1-x)\log(1-x)$ for $x \in$

[1] We adopt the convention that $x \log x = 0$ at $x = 0$.

$[0,1]$. The function h is concave on the interval $[0,1]$ and is decreasing on the interval $[\frac{1}{2},1]$.

As a corollary of the previous proposition, we get:

Proposition 2.7. *For random variables A and B where $B \in \{0,1\}$, we have*

$$\mathbf{I}(A:B) \leq \mathsf{h}(p(B=1)).$$

The following upper bound for the binary entropy function will be very useful.

Proposition 2.8. $\mathsf{h}(x) \leq 1 - 2\log e\left(x-\frac{1}{2}\right)^2$ *for all* $x \in [0,1]$.

Proof. Define the function $\mathsf{g}(x) = 1 - 2\log e\left(x-\frac{1}{2}\right)^2 - \mathsf{h}(x)$ for $x \in [0,1]$. Then, the first and second derivatives of $\mathsf{g}(x)$ for $x \in [0,1]$ are

$$\mathsf{g}'(x) = -4\log e\left(x-\frac{1}{2}\right) - \log(1-x) - \log x \text{ and } \mathsf{g}''(x) = -4\log e + \frac{\log e}{x(1-x)}.$$

As $x(1-x) \leq \frac{1}{4}$ for $x \in [0,1]$, $\mathsf{g}''(x) \geq 0$ for $x \in [0,1]$. It follows that the function g is convex on the interval $[0,1]$ and hence has a unique minima. Furthermore, the derivative $\mathsf{g}'(\frac{1}{2}) = 0$, so the minimum value of $\mathsf{g}(x)$ is attained at $x = \frac{1}{2}$. Hence, $\mathsf{g}(x) \geq \mathsf{g}(\frac{1}{2}) = 0$. □

2.3.1 Basic Lemmas

A standard fact about entropy and mutual information is the chain rule:

Proposition 2.9 (Chain Rule for Entropy)**.** *If $A = A_1, \ldots, A_n$, then*

$$\mathbf{H}(A) = \sum_{i=1}^{n} \mathbf{H}(A_i|A_{<i}) \leq \sum_{i=1}^{n} \mathbf{H}(A_i).$$

Lemma 2.10 (Chain Rule for Mutual Information)**.** *For random variables $A_1, \ldots, A_n$, B and C,*

$$\mathbf{I}(A_1, \ldots, A_n : B|C) = \sum_{i=1}^{n} \mathbf{I}(A_i : B|A_{<i}C).$$

A clever application of the chain rule gives us the following lemma:

Lemma 2.11 ([BR11]). *Let $X = X_1, \ldots, X_n$ and $Y = Y_1, \ldots, Y_n$ be random variables such that the n-tuples $(X_1, Y_1), \ldots, (X_n, Y_n)$ are mutually independent. Let R be an arbitrary random variable. Then,*

$$\sum_{i=1}^{n} \mathbf{I}\left(X_i : R | X_{<i} Y_{\geq i}\right) \leq \mathbf{I}\left(X : R | Y\right) \text{ and } \sum_{i=1}^{n} \mathbf{I}\left(Y_i : R | X_{\leq i} Y_{>i}\right) \leq \mathbf{I}\left(Y : R | X\right).$$

Proof. Using the chain rule,

$$\begin{aligned}
\sum_{i=1}^{n} \mathbf{I}\left(X_i : R | X_{<i} Y_{\geq i}\right) &\leq \sum_{i=1}^{n} \mathbf{I}\left(X_i Y_{<i} : R | X_{<i} Y_{\geq i}\right) \\
&= \sum_{i=1}^{n} \mathbf{I}\left(X_i : Y_{<i} | X_{<i} Y_{\geq i}\right) + \sum_{i=1}^{n} \mathbf{I}\left(X_i : R | X_{<i} Y\right) \\
&= \sum_{i=1}^{n} \mathbf{I}\left(X_i : R | X_{<i} Y\right) = \mathbf{I}\left(X : R | Y\right).
\end{aligned}$$

where the second last equality follows since $\mathbf{I}\left(X_i : Y_{<i} | X_{<i} Y_{\geq i}\right) = 0$. The second bound follows similarly. □

Next we present some other basic lemmas about mutual information. The next lemma says that the mutual information only increases if we condition on a random variable that is independent of one of the variables in the mutual information term.

Lemma 2.12. *If B and C are independent, $\mathbf{I}\left(A : B\right) \leq \mathbf{I}\left(A : B | C\right)$.*

Proof. We repeatedly use the chain rule:

$$\begin{aligned}
\mathbf{I}\left(A : B\right) &\leq \mathbf{I}\left(A : B\right) + \mathbf{I}\left(C : B | A\right) \\
&= \mathbf{I}\left(AC : B\right) = \mathbf{I}\left(C : B\right) + \mathbf{I}\left(A : B | C\right) = \mathbf{I}\left(A : B | C\right).
\end{aligned}$$

□

It is natural to expect that if we apply a function to a random variable, then the resulting new random variable can only reveal less information, or formally $\mathbf{I}\left(A : f(B)\right) \leq \mathbf{I}\left(A : B\right)$ where f is an arbitrary function of B. The next proposition generalizes this to arbitrary Markov chains.

Proposition 2.13 (Data Processing Inequality). *Let A, M and B be random variables such that $A - M - B$, then $\mathbf{I}\left(A : B\right) \leq \mathbf{I}\left(A : M\right)$.*

A straightforward corollary of the above is the following.

Proposition 2.14. *If B determines C, then* $\mathbf{I}(A:C) \leq \mathbf{I}(A:B)$.

The following propositions are in a similar spirit to the above.

Proposition 2.15. *Let A, B, C and W be random variables such that $AB - C - W$, then*

$$\mathbf{I}(A:B|CW) = \mathbf{I}(A:B|C).$$

Proof. Using the chain rule we expand $\mathbf{I}(AW:B|C)$ in two different ways:

$$\mathbf{I}(W:B|C) + \mathbf{I}(A:B|CW) = \mathbf{I}(AW:B|C) = \mathbf{I}(A:B|C) + \mathbf{I}(W:B|AC).$$

The terms $\mathbf{I}(W:B|C)$ and $\mathbf{I}(W:B|AC)$ are zero since $AB - C - W$. □

Proposition 2.16. *If A and B are independent and $A - M - B$, then* $\mathbf{I}(A:M) = \mathbf{I}(A:M|B)$.

Proof. Since A and B are independent $\mathbf{H}(A|B) = \mathbf{H}(A)$ and since $A - M - B$, $\mathbf{H}(A|MB) = \mathbf{H}(A|M)$. So, we have

$$\mathbf{I}(A:M|B) = \mathbf{H}(A|B) - \mathbf{H}(A|MB) = \mathbf{H}(A) - \mathbf{H}(A|M) = \mathbf{I}(A:M).$$ □

2.4 Divergence

The *divergence*, also sometimes called the *Kullback-Liebler* or *KL* divergence between distributions $p(A)$ and $q(A)$ is defined to be

$$\frac{p(A)}{q(A)} = \sum_a p(a) \log \frac{p(a)}{q(a)}.$$

The standard notation for KL divergence between $p(A)$ and $q(A)$ is $\mathsf{D}(p(A)||q(A))$ but the above notation will be clearer for our purposes.

For three random variables A, B, C and an event $\mathcal{E}$ in a probability space p, we will use the shorthand $\dfrac{A|bc\mathcal{E}}{A|c} = \dfrac{p(A|bc\mathcal{E})}{p(A|c)}$, when p is clear from context.

The *mutual information* between A, B conditioned on C can also be defined in terms of the divergence as:

Proposition 2.17 (Mutual Information and Divergence)**.**

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

2.4.1 Basic Divergence Facts

In the following, p and q are probability spaces (over the same sample space), and A is a random variables over the underlying sample space.

The divergence between two distributions is always non-negative.

Proposition 2.18. $\dfrac{p(A)}{q(A)} \geq 0.$

Similar to the chain rule for mutual information, we also have a chain rule for divergence.

Proposition 2.19 (Chain Rule for Divergence)**.** *If* $A = (A_1, \ldots, A_s)$*, then*

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

Define $\mathsf{d}(a||b) := a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}$ to be the binary divergence. Then, the following holds:

Proposition 2.20. *For any* $0 \leq \varepsilon, \delta < 1/2$*,* $\mathsf{d}\left(1-\varepsilon||\frac{1}{2}+\delta\right)$ *is a decreasing function of both* ε *and* δ*. Furthermore,*

$$\mathsf{d}\left(1-\varepsilon\middle|\middle|\frac{1}{2}+\delta\right) \geq 1 - \varepsilon\log\left(\frac{4}{\varepsilon}\right) - 4\delta.$$

Proof. We can write

$$\begin{aligned}
\mathsf{d}\left(1-\varepsilon \middle|\middle| \frac{1}{2}+\delta\right) &= (1-\varepsilon)\log\left(\frac{2}{1+2\delta}\right) + \varepsilon\log\left(\frac{2}{1-2\delta}\right) - \mathsf{h}(\varepsilon) \\
&= \log 2 - \log(1+2\delta) + \varepsilon\log\left(\frac{1+2\delta}{1-2\delta}\right) - \mathsf{h}(\varepsilon) \\
&\geq 1 - \log(1+2\delta) - \mathsf{h}(\varepsilon) \geq 1 - 4\delta - \mathsf{h}(\varepsilon),
\end{aligned}$$

where we used that $\log(1+2\delta) \leq 2\delta/\ln 2 \leq 4\delta$.

Furthermore, we can upper bound $\mathsf{h}(a) \leq a\log(4/a)$ for $0 \leq a \leq 1/2$. This can be observed by noting that $(1-a)\log\frac{1}{1-a} \leq 2a$ for $0 \leq a \leq 1/2$. Therefore, $\mathsf{h}(a) \leq a\log(1/a) + 2a = a\log(4/a)$. □

If the divergence between two distributions is small, then one can also show that they are close in statistical distance.

Proposition 2.21 (Pinsker's Inequality)**.** $|p(A) - q(A)|^2 \leq \frac{\ln 2}{2} \cdot \frac{p(A)}{q(A)} \leq \frac{p(A)}{q(A)}$.

2.4.2 Divergence Inequalities

The following propositions bound the change in divergence when extra conditioning is involved. Some of these were proved in [BRWY13, GKR16]. For completeness, we include full proofs for all of them. Below, p and q are probability spaces, and A and B are random variables on the underlying sample space.

Proposition 2.22 ([BRWY13])**.** *For an event $\mathcal{W}$ and variables A, B in a probability space p, we have*

$$\mathbb{E}_{b|\mathcal{W}}\left[\frac{A|b\mathcal{W}}{A}\right] \leq \log\frac{1}{p(\mathcal{W})} + \mathbf{I}\left(A : B|\mathcal{W}\right).$$

Proof. By Proposition 2.17, we can write

$$\begin{aligned}\mathbb{E}_{b|\mathcal{W}}\left[\frac{A|b\mathcal{W}}{A}\right] - \mathbf{I}\left(A:B|\mathcal{W}\right) &= \sum_{ab} p(ab|\mathcal{W}) \log \frac{p(a|b\mathcal{W})\cdot p(a|\mathcal{W})}{p(a)\cdot p(a|b\mathcal{W})} \\ &= \sum_{a} p(a|\mathcal{W}) \log \frac{p(a|\mathcal{W})}{p(a)} \\ &= \sum_{a} p(a|\mathcal{W}) \log \frac{p(\mathcal{W}|a)}{p(\mathcal{W})} \leq \log \frac{1}{p(\mathcal{W})}.\end{aligned}$$

□

Proposition 2.23 ([BRWY13])**.** $\mathbb{E}_{p(b)}\left[\frac{p(A|b)}{q(A)}\right] \geq \frac{p(A)}{q(A)}$.

Proof.

$$\mathbb{E}_{p(b)}\left[\frac{p(A|b)}{q(A)}\right] - \frac{p(A)}{q(A)} = \sum_{a,b} p(ab) \log \frac{p(a|b)\cdot q(a)}{q(a)\cdot p(a)} = \mathbb{E}_{p(b)}\left[\frac{p(A|b)}{p(A)}\right] \geq 0,$$

where the last inequality follows from Proposition 2.18. □

Proposition 2.24 ([GKR16])**.** $\mathbb{E}_{p(b)}\left[\frac{p(A|b)}{p(A)}\right] \leq \mathbb{E}_{p(b)}\left[\frac{p(A|b)}{q(A)}\right]$.

Proof.

$$\mathbb{E}_{p(b)}\left[\frac{p(A|b)}{q(A)}\right] - \mathbb{E}_{p(b)}\left[\frac{p(A|b)}{p(A)}\right] = \sum_{a,b} p(ab) \log \frac{p(a|b)\cdot p(a)}{q(a)\cdot p(a|b)} = \frac{p(A)}{q(A)} \geq 0,$$

where the last inequality follows from Proposition 2.18. □

Proposition 2.25. *Let $A \in \{0,1\}$ and let $\gamma = p(a)$ and $\varepsilon = q(a)$ for $a = 1$. Then,*

$$\frac{p(A)}{q(A)} \geq \gamma \log \frac{\gamma}{e\varepsilon}.$$

Proof. We have

$$\begin{aligned}\frac{p(A)}{q(A)} = \gamma \log \frac{\gamma}{\epsilon} + (1-\gamma)\log\frac{1-\gamma}{1-\epsilon} &\geq \gamma \log \frac{\gamma}{\epsilon} + (1-\gamma)\log(1-\gamma) \\ &\geq \gamma \log \frac{\gamma}{\epsilon} - \gamma \log e = \gamma \log \frac{\gamma}{e\epsilon},\end{aligned}$$

where in the last inequality we used the fact that

$$-(1-\gamma)\ln(1-\gamma) = (1-\gamma)\ln\left(1+\frac{\gamma}{1-\gamma}\right) \leq (1-\gamma)\frac{\gamma}{1-\gamma} = \gamma. \qquad \square$$

2.5 Basic Probability Lemmas

For a string $a \in \{0,1\}^n$, the Hamming weight of a is defined to be the number of ones in a. The following proposition is a variant of the standard Chernoff-Hoeffding concentration bound.

Proposition 2.26 (Chernoff Bound)**.** *The number of strings in $\{0,1\}^m$ with Hamming Weight at least $3m/4$ is at most $e^{-m/8} \cdot 2^m$.*

The next lemma is a variant of Markov's inequality:

Lemma 2.27 (Averaging Lemma)**.** *Let A be a bounded random variable such that $\mathbb{E}[A] \geq \alpha$. For any $\beta < \alpha$, let $\mathcal{S} = \{a \mid A(a) \geq \beta\}$. Then, $p(\mathcal{S}) \geq \frac{\alpha-\beta}{m-\beta}$ where $m = \max_a\{A(a)\}$.*

Proof. We have

$$\mathbb{E}[A] = \sum_{a\in\mathcal{S}} p(a)A(a) + \sum_{a\notin\mathcal{S}} p(a)A(a) \leq p(\mathcal{S})m + (1-p(\mathcal{S}))\beta.$$

which gives us that $(m-\beta)p(\mathcal{S}) \geq \mathbb{E}[A] - \beta \geq \alpha - \beta$. $\square$

§ Part I §

Interactive Models

3 Compression in Interactive Models

In a remarkable paper, *A Mathematical Theory of Communication*, Shannon [Sha48] laid the foundations for modeling communication systems mathematically. The fundamental question that Shannon considered is the following: suppose that a sender wants to transmit messages $M \in \{0,1\}^n$ to a receiver. If the distribution of the messages is known to both parties, how many bits does the sender need to transmit, so that the receiver can determine the message M uniquely?

Shannon observed that the statistical properties of the messages that a transmitter is sending to a receiver can be used to transmit the messages much more efficiently. For example, if a message is more probable, then one could encode it using a smaller number of bits and send the corresponding encoding to transmit it uniquely. Shannon defined *entropy* to capture the notion of information contained in a distribution: recall that given a random variable $M \in \{0,1\}^n$ representing the messages, the entropy $\mathbf{H}(M)$ is defined as

$$\mathbf{H}(M) = \sum_{m \in \{0,1\}^n} p(m) \log\left(\frac{1}{p(m)}\right),$$

where $p(M)$ is the distribution on the messages.

If the distribution of M is uniform, then it has n bits of entropy; in this case, we would not expect to send a message shorter than n bits to convey M. On the other hand, if M was a random bit followed by $n-1$ zeros, then its entropy is 1 bit; in this scenario, we could just send the first bit of M to convey its value. Shannon proved that one needs to transmit at least $\mathbf{H}(M)$ bits to convey M uniquely and he also showed that there is a scheme to compress the messages to this limit — one can always transmit M with $\mathbf{H}(M)+1$ bits in expectation. The concepts introduced by Shannon led to the development of the field of information theory and have had wide-reaching impacts in computer science.

Shannon's formulation, although quite natural, is intended only for communication systems where there is no interaction — only the transmitter wants to send messages to the receiver. In the field of theoretical computer science, many systems can be modeled as multiple parties receiving correlated inputs and interacting with each other based on their inputs. In these cases, the interacting parties have information about the distributions of messages of other parties since their inputs are correlated. One could hope to utilize this partial knowledge to transmit the messages more efficiently. For example, if there are two interacting parties and their inputs come from a joint distribution where they are always equal, then the parties need not communicate at all! The notion of entropy does not quite capture this setting, as each message may individually contain a large amount of information; it just happens to be the case that this information is known to the other party, so one could save on the amount of communication by having the second party guess the messages themselves.

As such, to define an analogue of entropy for these interactive models of communication, one needs to be a little bit more careful. In this chapter, we will look at two interactive settings — the model of two-party communication and streaming computation, and see how the compression question can be defined mathematically in these settings. In both of these settings, the compression question also has interesting connections to proving lower bounds and the so-called "direct-sum" question and we shall briefly attempt to survey these connections as well.

3.1 Communication Complexity

In the standard setting of communication complexity, there are two parties who hold inputs $x \in \{0,1\}^n$ and $y \in \{0,1\}^n$ respectively, where the inputs are drawn from some joint distribution. The parties want to compute a function $f(x,y)$ of their inputs and in order to achieve this, they communicate with each other by exchanging messages according to a pre-agreed *communication protocol*. The complexity of the function f is measured by the number of bits communicated to compute $f(x,y)$.

For many interesting functions f, there are clever protocols that allow the parties to compute $f(x,y)$ without exchanging their inputs. However, since there is a distribution on inputs, one may wonder if we can use the statistical properties of the inputs à la Shannon to communicate fewer bits

and still be able to compute the function, possibly with a small error. As there is a correlated distribution on inputs, the players do have some knowledge about each others' inputs and we would like to be able to take advantage of that. One very natural measure of information which captures this correlation was introduced by Barak, Braverman, Chen and Rao [BBCR13]. For any communication protocol, they defined the *internal information cost* of the protocol as the amount of information learned by both parties about each others' inputs. For example, if the joint distribution is such that x was always equal to y, then this measure is zero since the parties do not learn any information that they did not already know. This is what we naturally expect, and similar to Shannon's setting, one can also prove that the internal information cost of a protocol is a lower bound on the communication of any protocol that *simulates* the behavior of the original protocol. A natural question arises — Can we simulate a protocol that has internal information cost I with $O(I)$ bits of communication?

This interactive compression question, although quite natural, arose in the context of the *direct-sum* question in communication complexity. While there could be many other natural measures of information that could be considered as a candidate to generalize Shannon's theory to the setting of two-party communication, internal information cost and similar measures are perhaps the only ones that have turned out to be useful in proving communication lower bounds. As such, the above interactive compression question also has important consequences for proving communication lower bounds and the direct-sum question for communication. For organizational purposes, we first discuss the possibility of interactive compression as a standalone problem, and in a later section, we survey the connection to proving lower bounds and direct-sum statements for communication.

3.1.1 Compressing Communication

Given a communication protocol π, let X and Y denote the inputs to the parties and M denote the messages exchanged by the parties. We will often call M the *transcript* of the protocol. The parties may have access to *public* randomness, denoted by R, which is known to both parties, as well as access to *private* randomness which is only known to the party who owns it. We will use the term *public-coin* (resp. *private-coin*) protocols to refer to protocols that only use public (resp. private) randomness.

For a general communication protocol, it will often be more convenient for us to hide the private randomness and think of the messages sent by each party as being drawn from some distribution that only depends on that party's input, public randomness and previous messages. For example, for the party that holds X, the distribution of the message is determined by X,R and the previous messages.

The *communication complexity* of a protocol π, denoted by $\mathsf{CC}(\pi)$, is the maximum length of any transcript of the protocol. Barak *et al.* [BBCR13] defined the *internal information cost* of a protocol π as the amount of information learned by parties about the inputs to the other party:

$$\mathsf{IC}_p(\pi) = \mathbf{I}_p\left(X : M|YR\right) + \mathbf{I}_p\left(Y : M|XR\right),$$

where $p(X,Y)$ is the input distribution.

Another useful information measure was defined by Chakrabarti, Shi, Wirth and Yao [CSWY01], who were the ones who introduced the notion of information cost of protocols. They defined the *external information cost* of a protocol with respect to the input distribution $p(X,Y)$ as the amount of the information learned about the inputs by an external observer of the messages:

$$\mathsf{IC}_p^{\mathsf{ext}}(\pi) = \mathbf{I}_p\left(XY : M|R\right).$$

Note that the internal and external information costs both depend on the input distribution $p(X,Y)$. We will often omit the input distribution from the notation if it is clear from the context. The internal information cost of a protocol is always at most the external information cost of a protocol, which, in turn, is upper bounded by the communication complexity of the protocol. When $p(X,Y)$ is a *product distribution* on the inputs X and Y, *i.e.* X and Y are independent under $p(X,Y)$, the internal and external information costs are equal, so an external observer of the messages learns the same amount of information as the parties exchanging messages.

Proposition 3.1. *For any protocol π and any distribution $p(X,Y)$,*

$$\mathsf{IC}_p(\pi) \leq \mathsf{IC}_p^{ext}(\pi) \leq \mathsf{CC}(\pi).$$

Moreover, if $p(X,Y)$ is a product distribution on X and Y, then $\mathsf{IC}_p(\pi) = \mathsf{IC}_p^{ext}(\pi)$.

Given an input distribution $p(X, Y)$, we say that a protocol τ *simulates* another protocol π with error ε, if by running the protocol τ, both parties sample the same transcript m with probability at least $1 - \varepsilon$ and conditioned that the same transcript was sampled, the distribution on the sampled transcript is ε-close to the distribution on the transcript induced by the protocol π. The interactive compression question can then be posed as asking if a protocol π with $\mathsf{CC}(\pi) = C$ and $\mathsf{IC}(\pi) = I$ can be simulated with constant error by another protocol with $O(I)$ bits of communication on the same input distribution.

In the last decade or so, rapid progress has been made towards answering the above question in different scenarios, but many fundamental questions still remain open. There are several special cases where we know almost optimal compression of interactive protocols. For bounded-round protocols, Harsha, Jain, McAllester and Radhakrishnan [HJMR10] (see also [BG14]) showed how to give optimal simulations with no error in terms of the external information cost, while Braverman and Rao [BR11] gave near optimal simulations with error in terms of the internal information cost (up to a $1 + o(1)$ factor). For input distributions that are product over the inputs to the parties, Sherstov [She16], building on the work of Kol [Kol16] and Barak *et al.* [BBCR13], showed that protocols with internal information cost I can be simulated with communication $O(I \cdot \log^2 I)$, which is almost optimal.

For unbounded-round protocols and non-product distributions, many of the best known simulations for interactive protocols are far from optimal. We know how to simulate a protocol with internal information I and communication C using a protocol with communication $O(\sqrt{IC} \cdot \log C)$ [BBCR13]. Similarly, we know how to simulate any interactive protocol with external information cost I_{ext} and communication C using a protocol with communication $O(I_{ext} \cdot \log^2 C)$ [BBCR13]. Very recently, Braverman and Kol [BK18] showed that the same compression can be achieved with $O(\mathrm{poly}(I_{ext}) \log\log C)$ communication which is not optimal in terms of I_{ext} but has an exponentially smaller dependence on C. Note that all of these simulations depend on the communication complexity of the original protocol which could be much larger compared to the internal information cost I, or the external information cost I_{ext}. Since the communication of the compressed protocols is decidedly smaller than C, it may be tempting to try a recursive strategy and try to compress the proto-

col again. However, it turns out that using the above known compression methods, the information cost of the compressed protocol blows up, so it is unclear if such a strategy can be made to work. On the other hand, if we do not want a dependence on the communication, then the best known result is due to Braverman [Bra12] who showed how to simulate any protocol with internal information cost I using communication $2^{O(I)}$. This is exponentially larger than what we are asking for.

For unbounded-round protocols and arbitrary input distributions, in some other special cases, we know simulations that improve on the results described in the previous paragraph. Natarajan Ramamoorthy and Rao [NR15] gave better simulations than the ones described above when one party reveals less information than the other. Bauer, Moran and Yehudayoff [BMY15] give better simulations if the protocol has no private randomness. We summarize all the results discussed so far in Table 3.1 for ease of reference.

3.1.2 Information Complexity and Communication Lower Bounds

Given that the best known compression as a function of the internal information cost I uses $2^{O(I)}$ bits of communication, it is reasonable to ask if we can prove lower bounds for interactive compression. It turns out that this question is closely tied to proving communication lower bounds. Information theory based methods have been widely used to prove lower bounds in communication complexity. In fact, the information theoretic lower bound for the set disjointness function (given sets $x \subseteq [n]$ and $y \subseteq [n]$, the set-disjointness function is the indicator function for $x \cap y = \emptyset$) [KS92, Raz92, BYJKS02] can be seen as precursors to some of the compression results.

Let us define the *information complexity* and the *distributional* communication complexity of a boolean function $f(x, y)$ under an input distribution $p(X, Y)$ as follows:

$$\mathsf{IC}_p(f) = \inf_{\pi} \mathsf{IC}_p(\pi) \text{ and } \mathsf{CC}_p(f) = \inf_{\pi} \mathsf{CC}(\pi),$$

where both infimums are taken over all protocols that compute $f(x, y)$ with error at most $1/3$ on the input distribution $p(X, Y)$. We remark that the second infimum in the definition of $\mathsf{CC}_p(f)$ can be replaced with a minimum.

Type of Protocol	Input Distribution	Compressed Protocol		Reference
		Communication	Error	
r-round	Arbitrary	$I_{ext} + r$	zero	[HJMR10] [BG14]
		$I + r$	constant	[BR11]
Arbitrary	Arbitrary	$I_{ext} \cdot \log^2 C$	constant	[BBCR13]
		$\sqrt{IC} \cdot \log C$		[BK18]
		$\mathrm{poly}(I_{ext}) \cdot \log\log C$		[Bra12]
		$2^{O(I)}$		
Asymmetric	Arbitrary	$(I_1^{1/4} C^{3/4} + \sqrt{I_2 C}) \cdot \log C$	constant	[NR15]
		$I_2 \cdot 2^{O(I_1)}$		
Public-coin	Arbitrary	$I^2 \cdot \log\log C$	constant	[BMY15]
Arbitrary	Product	$I^2 \cdot \mathrm{polylog}(I)$	constant	[Kol16]
		$I \cdot \log^2(I)$		[She16]

Table 3.1: Comparison of the various known compression results. Constant factors are suppressed for readability. I, I_{ext} and C respectively denote the internal information cost, the external information cost and the communication complexity of the original protocol. For the asymmetric compression results, $I_1 = \mathbf{I}(X : M|YR)$ and $I_2 = \mathbf{I}(Y : M|XR)$ denote the information revealed by the first and second party separately.

It is straightforward from the definitions to see that Proposition 3.1 also implies that $\mathsf{IC}_p(f) \leq \mathsf{CC}_p(f)$: the information complexity of a boolean function f is a lower bound on the communication complexity. Furthermore, the $2^{O(I)}$ simulation of Braverman [Bra12] implies that $\mathsf{CC}_p(f) \leq 2^{O(\mathsf{IC}_p(f))}$, so, we have $\mathsf{IC}_p(f) \leq \mathsf{CC}_p(f) \leq 2^{O(\mathsf{IC}_p(f))}$. An optimal simulation of protocols with internal information cost I with $O(I)$ bits of communication would, in fact, imply that information complexity is always of the same order as communication complexity, while proving a lower bound on interactive compression for boolean functions would disprove this.

It turns out that many information theoretic lower bound proofs in communication complexity, such as the lower bound proofs for disjointness [Raz92, BYJKS02], also give lower bounds on the information complexity of the corresponding functions. Taking this further, Braverman and Weinstein [BW12] showed that the *discrepancy* method, a standard technique for proving communication lower bounds also gives lower bounds on information complexity. Given a boolean function $f(x, y)$ and a distribution $p(X, Y)$, the discrepancy of f with respect to $p(X, Y)$ is defined as

$$\mathsf{Disc}_p(f) := \max_{\mathcal{R}} |p((X, Y) \in \mathcal{R} \wedge f = 0) - p((X, Y) \in \mathcal{R} \wedge f = 1)|,$$

where the maximum is taken over all rectangles $\mathcal{R} = \mathcal{S} \times \mathcal{T}$ in the input space.

We have the following relation between the distributional communication complexity and discrepancy for boolean functions:

Lemma 3.2.

$$\mathsf{CC}_p(f) = \Omega\left(\log\left(\frac{1}{\mathsf{Disc}_p(f)}\right)\right).$$

Another way to interpret the above relation is the following. In a deterministic communication protocol, the set of inputs that generates a particular transcript is a rectangle in the input space, as each bit sent by one party further partitions their side of the input space into two parts. If the protocol has zero error, then any rectangle $\mathcal{R} = \mathcal{S} \times \mathcal{T}$ that corresponds to a transcript is *monochromatic* under the distribution $p(X, Y)$, *i.e.* the function on the set $\mathsf{supp}(p(X, Y)) \cap \mathcal{R}$ is either all zeros or all ones. Moreover, if the communication complexity of the protocol is c bits, then from a straightforward counting argument, it follows that some rectangle must have at least 2^{-c} fraction of all the inputs. If

the deterministic protocol errs on the input distribution, then the rectangle with density at least 2^{-c} that we obtain, is not monochromatic but nearly monochromatic. If a function f has discrepancy d under an input distribution $p(X,Y)$, then every (nearly) monochromatic rectangle must have size $2^{-O(d)}$ giving us the above lower bound.

Braverman and Weinstein [BW12] showed that if a boolean function $f(x,y)$ has information complexity I with respect to a distribution $p(X,Y)$, then it must have a nearly monochromatic rectangle of density $2^{-O(I)}$, and so large discrepancy. In other words, the discrepancy method, in fact, gives a lower bound on information complexity.

Theorem 3.3 ([BW12])**.**

$$\mathsf{IC}_p(f) = \Omega\left(\log\left(\frac{1}{\mathsf{Disc}_p(f)}\right)\right).$$

The above implies that upper bounds on the size of nearly monochromatic rectangles cannot be used to prove lower bounds on the communication complexity of functions that have small information complexity[1]. Building on [BW12], Kerenedis, Laplante, Lerays, Ronald and Xiao [KLL+12] further proved that, essentially, all communication lower bound methods that were known at the time, in fact prove upper bounds on the size of nearly monochromatic rectangles.

This revealed a significant weakness in our ability to prove new lower bounds in communication complexity, and thus made the question of proving a lower bound for interactive compression also important for the purposes of proving new communication lower bounds. To be able to prove that a function has different information and communication complexities, we need to come up with techniques that will allow us to prove communication lower bounds for functions that have large monochromatic rectangles. Functions which have high communication complexity but no large monochromatic rectangles certainly exist — one can plant large monochromatic rectangles into a random function to obtain such an example with high probability.

In a remarkable sequence of papers, Ganor, Kol and Raz [GKR14, GKR16] showed that there is a boolean function with information complexity I with respect to a distribution that requires $2^{\Omega(I)}$

[1]The information based methods for proving lower bounds on disjointness also prove that disjointness does not have large monochromatic rectangles where all entries are 1.

communication under the same distribution. This proved that Braverman's simulation [Bra12] is tight if we limit the communication cost of the simulation to depend only on the internal information cost of the original protocol. The proof of Ganor *et al.* gives a method to prove communication lower bounds on functions that have many large monochromatic rectangles, potentially leading to fundamentally different methods to prove lower bounds on communication problems. In Chapter 4 of this thesis, we build on the work of Ganor, Kol and Raz [GKR14] and introduce another new technique to prove lower bounds for functions that have large monochromatic rectangles. This new method also allowed us to give a simpler proof for separating information and communication complexity.

3.1.3 Direct Sum for Communication Complexity

As mentioned before, the original motivation for the interactive compression question was the direct-sum question in communication complexity. In general, a direct-sum question in any computational model asks if t independent copies of a task require t times the resources. In many models, such as that of boolean circuits, there are surprising ways of computing t copies of a function much more efficiently than expending t times the resources.

The direct-sum question also has a close connection to proving lower bounds since it gives us a generic way to construct functions that are not efficiently computable, a technique known as *hardness amplification*. In the model of communication complexity, many hard functions on $\{0,1\}^n \times \{0,1\}^n$ are constructed by composing multiple copies of a simple function on two bits. As an example, consider the disjointness function where the parties are given indicator vector for subsets $x \subseteq [n]$ and $y \subseteq [n]$ and their goal is to compute if there is a coordinate i such that $x_i = y_i = 1$. This can be viewed as composing n copies of the AND function on two bits with the n-bit $NAND$ function: $AND(x_i, y_i) = 1$ on some coordinate i if and only if the inputs to the parties are not disjoint. Many such functions for which lower bounds are known, are of a similar form. A general way of proving a lower bound for such functions would be to argue that computing n copies of the composed function requires n times the communication required for computing a single copy of the function, similar to a direct-sum statement. However, the techniques for proving such lower bounds are currently some-

what ad hoc, relying on the structure of these functions. Proving a general direct-sum statement for an arbitrary function f would most likely give us a unified set of techniques to prove communication lower bounds for such composed functions.

To define the direct-sum question for communication complexity, let us consider a boolean function $f(x,y)$ and an input distribution $p(X,Y)$. Suppose that computing f on the input distribution $p(X,Y)$ with error at most $1/3$ requires communicating C bits. For the direct-sum question, the parties receive inputs $x_1,\cdots,x_t$ and $y_1,\cdots,y_t$ where $(x_1,y_1),\cdots,(x_t,y_t)$ are t inputs sampled independently from the input distribution $p(X,Y)$. The direct-sum question then asks — does computing the function $f^t(x_1,\cdots,x_t,y_1,\cdots,y_t) := (f(x_1,y_1),\cdots,f(x_t,y_t))$ with error at most $1/3$ require $\Omega(tC)$ bits of communication? More formally, defining $p^t(X,Y)$ to be the product distribution obtained from taking k independent samples from $p(X,Y)$, the optimal direct-sum conjecture is the following statement.

Conjecture 3.4 (Optimal Direct sum conjecture)**.** *Given an integer $t>1$, a boolean function $f(x,y)$ and an input distribution $p(X,Y)$, the following holds:*

$$\mathsf{CC}_{p^t}(f^t) = \Omega\left(t\cdot \mathsf{CC}_p(f)\right).$$

Barak, Braverman, Chen and Rao [BBCR13] showed that communication complexity of computing t copies of f is closely related to the information complexity of f.

Theorem 3.5 ([BR11])**.** *For any integer $t>1$, the following holds:*

$$\mathsf{IC}_p(f) \leq \frac{\mathsf{CC}_{p^t}(f^t)}{t}.$$

The above implies that if we had a *low communication* protocol for computing t copies of a function, then one could get a *low information* protocol to compute a single copy of a function. This meant that simulating low information protocols with small communication would give us a non-trivial direct-sum statement. As mentioned before, Barak *et al.* [BBCR13] also came up with a way to simulate protocols with communication C and information cost I with communication $O(\sqrt{I\cdot C}\log C)$. This proved that computing t copies of a function requires $\sqrt{t}/\log t$ times more communication.

Theorem 3.6 ([BBCR13]). *For any integer $t > 1$, the following holds*

$$\mathsf{CC}_{p^t}(f^t) = \Omega\left(\frac{\sqrt{t}}{\log t} \cdot \mathsf{CC}_p(f)\right).$$

In a later work, Braverman and Rao [BR11] proved that as t tends to infinity, the communication complexity of computing t copies of f is exactly equal to the information complexity of f. This gives us another operational interpretation of information complexity as the amortized communication complexity of computing f.

Theorem 3.7 ([BR11]). $\mathsf{IC}_p(f) = \lim_{t\to\infty} \frac{\mathsf{CC}_{p^t}(f^t)}{t}$.

Braverman and Rao [BR11] also proved that Conjecture 3.4 is only true if and only if protocols with information cost I can be simulated with communication $O(I)$. In light of this, the recent lower bounds of Ganor, Kol and Raz [GKR14, GKR16] as well the work in Chapter 4 of this thesis gives us a function and a distribution with respect to which Conjecture 3.4 is false.

We remark that if we assume that the communication protocol is bounded-round, then Conjecture 3.4 holds for such a protocol. This is because bounded-round protocols can be optimally compressed [BR11] as mentioned earlier. Furthermore, if the input distribution is product on the inputs to the parties, then Conjecture 3.4 could still be true and known compression results of Sherstov [She16] already imply an almost optimal statement: if p is a product distribution on the inputs, then $\mathsf{CC}_{p^t}(f^t) = \Omega\left(\frac{t}{\log^2 t} \cdot \mathsf{CC}_p(f)\right)$.

3.2 Streaming Computation

Chapter 5 considers the setting of streaming algorithms — where the input to the algorithm is too large to be read into the working memory at once. Such situations are very common in the current age of massive datasets. Companies like Google and Facebook have massive data centers where data is spread across millions of servers. It is prohibitive to even compute simple global statistics of the data because the entire dataset can not be retained in the memory for processing. In such cases, we

can view the input dataset as a stream where we get a chunk of data at a time and study what sort of computations are easy or hard in this model.

Formally, an input to a streaming algorithm is a sequence of n updates $x_1, \ldots, x_n$ arriving sequentially in time, and the algorithm at the end must compute a function $f(x_1, x_2, \ldots, x_n)$. The complexity measure of interest is the memory needed to carry out the computation. Here the memory used at time t is the number of bits stored by the algorithm after reading the inputs $x_1, \ldots, x_t$.

We can simulate the streaming setting by an interactive model in the following manner: there are n parties — let us denote them by numbers in $[n]$ — where party i holds the input x_i. The parties communicate in n rounds, where in round i, party i sends a private message to party $i+1$. At the end of the last round, the last party outputs the value $f(x_1, \cdots, x_n)$. The memory used by the algorithm at time t is then equal to the length of the message received by party t, and hence, the complexity of computing a function can be characterized in terms of the communication. Sometimes this interactive setting is also referred to as *private message-passing number-in-hand* setting.

Although not much is known about the private message-passing number-in-hand setting described above, in many cases, one can use lower bounds in the standard two-party communication model or another interactive model, called the *shared-blackboard number-in-hand* model, to prove lower bounds for many interesting streaming problems. Many of these lower bounds are proved by hardness amplification through direct-sum techniques and again the proof methods are somewhat ad hoc (see [BYJKS02, CKS03, Woo04, EJ08, GH09, Gro09, MWY13] for some of them). Since analogous questions in the setting of two-party communication complexity inspired major progress towards the direct-sum question and led to the theory of two-party interactive compression, a very natural thing to wonder is if we can prove a direct-sum statement for streaming algorithms and if this can lead to a similar theory of compression for streaming algorithms. Viewing streaming algorithms as interactive protocols in the private message-passing number-in-hand setting, this corresponds to proving direct-sum and compression results for such communication protocols. For the rest of this thesis, we will work directly with streaming algorithms and put the view of streaming algorithms as private message-passing number-in-hand protocols aside.

It turns out that some care is needed to even define the correct direct-sum statement in the stream-

ing setting. In Chapter 5, we discuss these issues and introduce a natural direct-sum question for streaming algorithms. We also show that, analogous to two-party communication complexity, this direct-sum question is dual to the question of compressing the memory of a streaming algorithm.

4 An Exponential Separation between Information and Communication

In this chapter, we give an example which shows that the communication complexity of a boolean function can be exponentially larger than its information complexity. This also shows that it is not always possible to simulate a protocol that has information cost I with $O(I)$ bits of communication.

Given a boolean function $f(x, y)$ and an input distribution $p(X, Y)$, recall that the following relation holds between information and communication complexity of the function:

$$\mathsf{IC}_p(f) \leq \mathsf{CC}_p(f) \leq 2^{O(\mathsf{IC}_p(f))},$$

where the upper bound follows from the compression result of Braverman [Bra12].

As discussed previously, a key hurdle in proving that information and communication complexity could be different was to come up with a new lower bound technique — all previously known communication lower bound methods only proved lower bounds for functions that have small monochromatic rectangles (either with all one or all zero entries) and such functions also have large information complexity.

In a remarkable sequence of papers, Ganor, Kol and Raz [GKR14, GKR16] showed that there is a function with internal information cost I with respect to a distribution that requires $2^{\Omega(I)}$ communication under the same distribution. This proved that Braverman's simulation [Bra12] is tight. Their proof introduces the *relative discrepancy* method, which gives a way to prove communication lower bounds on functions that have many large monochromatic rectangles, potentially leading to fundamentally different methods to prove lower bounds on communication problems.

In this chapter, we build on the work of [GKR16] and give a new proof of their main result. Our proofs are shorter and we find them more intuitive. For parameters $k, n \in \mathbb{N}$, we define a boolean function called the *k-ary pointer jumping function* with inputs X, F (given to Alice) and Y, G (given to Bob) and a distribution $q(X, F, Y, G)$ on its inputs. We show that there is a protocol with small internal information cost that computes the k-ary pointer jumping function on the distribution $q(X, F, Y, G)$:

Theorem 4.1. *There is an $O(n \log k)$ communication randomized protocol for the k-ary pointer jumping problem that errs with probability at most $\frac{4}{\log n}$ under the input distribution $q(X, F, Y, G)$ and has internal information cost $O((\log k + \log\log n) \cdot 2^{\frac{2\log n}{k}})$ with respect to the distribution $q(X, F, Y, G)$.*

On the other hand, we show that no protocol with communication complexity significantly smaller than $\min\{k, \log n\}$ can compute the same function on the distribution $q(X, F, Y, G)$:

Theorem 4.2. *For large enough values of k and n, every protocol for the k-ary pointer jumping function that has communication complexity at most $\varepsilon^3 \cdot \min\{k, \log n\}$ errs with probability at least $\frac{1}{2} - 8\varepsilon$ on inputs drawn from the distribution $q(X, F, Y, G)$.*

Note that the statement above applies to all protocols, whether they be deterministic or use public and private randomness. Setting $n = 2^k$ in Theorem 4.1 and Theorem 4.2, we obtain our main result: when the inputs are sampled from the distribution $q(X, F, Y, G)$, the internal information complexity of the k-ary pointer jumping function is $O(\log k)$ but the communication complexity is $\Omega(k)$.

Corollary 4.3. *There is a $2^{O(k)}$ communication randomized protocol for the k-ary pointer jumping function that errs with probability at most $\frac{4}{k}$ under the input distribution $q(X, F, Y, G)$ and has internal information cost $O(\log k)$ with respect to distribution $q(X, F, Y, G)$. Moreover, for large enough k, every protocol for the k-ary pointer jumping function that has communication complexity at most $\varepsilon^3 k$, errs with probability at least $\frac{1}{2} - 8\varepsilon$ on the input distribution $q(X, F, Y, G)$.*

One of the new ideas that simplify our proofs is the use of a new technique: the notion of a *fooling distribution* to prove communication lower bounds. This gives us another method to separate

information and communication apart from the relative discrepancy method introduced in [GKR16]. We describe this technique in Section 4.3.4 and compare it with the relative discrepancy method in Section 4.4.3.

We remark that our function and distribution is a simpler variant of the *bursting noise function* defined in [GKR16] and as such we recover the same parameters in our main theorems as in the work of [GKR16]. For example, [GKR16] proves that any protocol with communication at most 2^t computing the bursting noise function with parameter $t \in \mathbb{N}$ must have error at least $\frac{1}{2} - 2^{-t}$. We recover the same parameters from Corollary 4.3 by setting $k = 512 \cdot 2^{4t}$ and $\varepsilon = \frac{2^{-t}}{8}$. An analogous statement also holds for the information cost upper bound in Corollary 4.3. Moreover, the inputs to the k-ary pointer jumping function are of length N where $\log\log\log N = \Theta(k)$ so the communication and information complexity of this function are really small in terms of the input length (a similar statement holds for the bursting noise function).

Consequences for Compression Even though Corollary 4.3 and the result of Ganor, Kol and Raz [GKR16] implies an exponential gap between information and communication complexity, the consequences for compression of two party protocols are rather weak. Corollary 4.3 gives us a communication problem which has internal information cost $I = \Theta(\log k)$ but requires communication $\mathrm{poly}(k)$. This does rule out the possibility of simulating protocols with internal information cost I and communication C with communication $O(I)$. But note that the protocol that achieves low information cost has communication $C = 2^{O(k)}$, so it is possible that one could compress a protocol with internal information cost I and communication C to $I \cdot \mathrm{polylog}(C)$ bits of communication. For the example given by Corollary 4.3, we have $I \cdot \mathrm{polylog}(C) = \mathrm{poly}(k)$, so such a compression will not violate Corollary 4.3. In particular, the following conjecture is still open:

Conjecture 4.4. *Given an input distribution $p(X, Y)$, if a protocol π has $\mathsf{IC}_p(\pi) = I$ and communication C, then it can be simulated up to a constant error by a protocol with communication $I \cdot \mathrm{polylog}(C)$.*

We remark that Braverman, Ganor, Kol and Raz [BGKR18] present a candidate that might give a counterexample to the above conjecture.

Consequences for Amortized Communication and Direct Sum Corollary 4.3 and the result of Ganor, Kol and Raz [GKR16] also have implications for the amortized communication complexity and the optimal direct-sum conjecture (Conjecture 3.4). The amortized communication complexity of a function $f(x,y)$ under an input distribution $p(X,Y)$ is defined to be the quantity $\lim_{k\to\infty} \frac{\mathsf{CC}_{p^k}(f^k)}{k}$. From the work of Braverman and Rao [BR11] it follows that the amortized communication complexity of a function is exactly equal to its information complexity. So, Corollary 4.3 also gives an example of a boolean function and an input distribution where the amortized communication complexity is exponentially smaller than the communication complexity.

Braverman and Rao [BR11] also showed that the only way to prove Conjecture 3.4 was to give a way to compress protocols that have internal information cost I to $O(I)$ communication. In light of this, Corollary 4.3 also shows that Conjecture 3.4 is false, although the implications are again rather weak.

In particular, Braverman and Rao [BR11] showed that r-round protocols with internal information cost I can be simulated with $O(I+r)$ bits of communication. Note that the low information protocol given by Corollary 4.3 has communication $2^{O(k)}$; in particular it also has $2^{O(k)}$ rounds. If we want to compute t copies of the function in Corollary 4.3 under the input distribution $q^t(X,F,Y,G)$, then using standard arguments one can convert this low information protocol to a $2^{O(k)}$ round protocol that has information cost $O(t\log k)$ under the input distribution $q^t(X,F,Y,G)$. Applying the result of Braverman and Rao [BR11] implies that cost of computing t copies of the function under the distribution $q^t(X,F,Y,G)$ is $O(t\log k)+2^{O(k)}$. If we had computed t copies in the trivial way, then communication would be $t\cdot\mathrm{poly}(k)$. So, when t is large enough so that $t\log k + 2^{O(k)} \ll tk$, we can compute t copies of the function in Corollary 4.3 much faster. But note that the savings start to kick in only when $t = 2^{\Omega(k)}$. For values where $t = 2^{\Theta(k)}$ or alternatively when $k = \Theta(\log t)$, the communication for computing t copies is $\frac{t\log\log t}{\log t}$ times the communication for computing one copy of the function. So far, it is still possible that a direct-sum statement for communication complexity holds with these weaker parameters. In particular, the following conjecture is still open.

Conjecture 4.5. *Given an integer $t>1$, a boolean function $f(x,y)$ and an input distribution $p(X,Y)$,*

it holds that

$$\mathsf{CC}_{p^t}(f^t) = \Omega\left(\frac{t\log\log t}{\log t}\cdot \mathsf{CC}_p(f)\right).$$

Recall that the result of Barak *et al.* [BBCR13] already shows that the right hand side could be taken to be $\sqrt{t}/\log t$ times the communication complexity of a single copy. Moreover, for product distributions $p(X,Y)$, the result of Sherstov [She16] implies that the right hand side can be taken be to $t/\log^2 t$ times the communication complexity of a single copy.

Closely Related Work The results stated in the beginning of this chapter as well as the work of [GKR16] proves that there is a boolean function which has widely different information complexity and communication complexity under a carefully designed input distribution. One can also ask if information complexity and communication complexity are exponentially separated for the worst-case input distribution (the non-distributional setting as defined in [Bra12]). Fontes, Jain, Kerenidis, Laplante, Lauriére and Roland [FJK+15] showed that the *relative discrepancy* technique introduced in [GKR16] cannot be used to separate information and communication complexity for a boolean function in the non-distributional setting.

Ganor, Kol and Raz [GKR15] gave an exponential separation of external information and communication for a *relation* (search problem) that holds in the non-distributional setting. The new proof uses a clever reduction to the randomized communication complexity of set disjointness. However, as the new result of [GKR15] proves a separation only for a search problem, the ideas presented in this chapter still give the most direct proof separating information complexity and communication complexity for boolean functions.

Also, it turns out that one of the technical lemmas that will be used to prove Theorem 4.2 can also be used to show that the randomized communication complexity of the Greater-Than function on n bits is $\Omega(\log n)$. The same lower bound was previously proved by Viola [Vio15] and also by Braverman and Weinstein [BW12] using different techniques. We will present this proof later in this chapter.

Our ideas were also extended by Anshu, Touchette, Yao and Yu [ATYY17] to prove a similar separation between quantum communication and classical information complexity.

Organization and References This chapter is based on the works [NS15] and [RS16]. We start with the preliminaries in Section 5.3. Then, we take a slight detour from the main topic of this chapter and look at the randomized communication complexity of the Greater-Than function. We give a simple lower bound proof for it in Section 4.2. Building upon these proof techniques, we come back to the central question of this chapter, separating information and communication complexity, in Section 4.3 where we define the function and the input distribution and give a high level overview of the proofs. In Section 4.4, we prove the communication lower bound, while Section 4.5 gives the upper bound on the information complexity.

4.1 Preliminaries

4.1.1 Communication Complexity

We briefly describe basic properties of communication protocols that we need. For more details see the textbooks [KN97] or [RY18]. For a deterministic protocol π, let $\pi(x, y)$ denote the messages of the protocol on inputs x, y. For any transcript m of the protocol, define the following events:

$$\mathcal{S}_m = \{x | \exists y \text{ such that } \pi(x, y) = m\}, \qquad \mathcal{T}_m = \{y | \exists x \text{ such that } \pi(x, y) = m\}.$$

We then have:

Proposition 4.6 (Messages Correspond to Rectangles)**.** *If m is a transcript and x, y are inputs to a deterministic protocol π, then, $\pi(x, y) = m \iff x \in \mathcal{S}_m \wedge y \in \mathcal{T}_m$.*

Proposition 4.6 implies:

Proposition 4.7 (Markov Property of Protocols)**.** *Let X and Y be random inputs to a deterministic protocol and let M denote the messages of this protocol. If X and Y are independent then $X - M - Y$.*

Note that the above implies that if x and y are independent inputs sampled from a distribution $p(X, Y)$ and m is a transcript of a deterministic protocol, then $p(xy|m) = p(xy|x \in \mathcal{S}_m, y \in \mathcal{T}_m) = p(x|x \in \mathcal{S}_m)p(y|y \in \mathcal{T}_m)$.

Lemma 4.8 (Errors and Statistical Distance)**.** *Let $h(x,y)$ be a boolean function and $p(X,Y)$ be a distribution such that $p(h=0) = p(h=1) = \frac{1}{2}$. If π is a deterministic protocol with messages M that computes h with error δ on the distribution p, then $|p(M|h=0) - p(M|h=1)| \geq 1 - 2\delta$.*

Proof. Since $|p(M|h=0) - p(M|h=1)| = \max_{\mathcal{Q}}(p(M \in \mathcal{Q}|h=0) - p(M \in \mathcal{Q}|h=1))$ it suffices to exhibit an event $\mathcal{Q}$ such that $p(M \in \mathcal{Q}|h=0) - p(M \in \mathcal{Q}|h=1) = 1 - 2\delta$. Let $\mathcal{M}_0$ denote the event that the protocol outputs a zero. Then, since $p(h=0) = p(h=1) = \frac{1}{2}$, writing the probability of success in terms of $\mathcal{M}_0$, we have

$$\begin{aligned} 1 - \delta &= \frac{p(M \in \mathcal{M}_0|h=0)}{2} + \frac{1 - p(M \in \mathcal{M}_0|h=1)}{2} \\ &= \frac{1}{2} + \frac{p(M \in \mathcal{M}_0|h=0) - p(M \in \mathcal{M}_0|h=1)}{2}. \end{aligned}$$

On rearranging, the above gives us that $p(M \in \mathcal{M}_0|h=0) - p(M \in \mathcal{M}_0|h=1) = 1 - 2\delta$ and hence the statistical distance must be at least $1 - 2\delta$. □

4.2 Communication Complexity of Greater-Than

In the first part of this chapter which is based on [NS15], we give a simple lower bound proof for the communication complexity of the greater-than function. Later we will build upon these techniques to give an exponential separation between communication and information complexities.

For $x \in \{0,1\}^n$, let $\mathsf{bin}(x)$ denote the integer whose binary representation is x[1]. The greater-than function $\mathsf{GT} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ is defined as

$$\mathsf{GT}(x,y) = \begin{cases} 1 \text{ if } \mathsf{bin}(x) \geq \mathsf{bin}(y) \\ 0 \text{ otherwise.} \end{cases}$$

Nisan [Nis94] showed that the *public-coin* randomized communication complexity of the greater-than function is $O(\log n)$ for bit-strings of length n. Using information theoretic techniques Viola

[1] We adopt the convention that the leftmost bit of x is the most-significant bit of $\mathsf{bin}(x)$.

[Vio15] gave a matching lower bound. Braverman and Weinstein [BW12] gave an alternative proof for the lower bound by analyzing the discrepancy of the greater-than function.

We give a very simple proof showing that there is an input distribution under which computing the greater-than function with error $1/10000$ requires communication $\Omega(\log n)$. By the well-known Yao min-max principle (see the textbooks [KN97] or [RY18] for example), this also implies that the *public-coin* randomized communication complexity of the greater-than function is $\Omega(\log n)$. Note that the choice of the error parameter is just for convenience in calculations and could be replaced by any other constant using standard techniques.

Theorem 4.9. *There exists an input distribution $p(X, Y)$ under which computing the function GT with error $1/10000$ requires $\Omega(\log n)$ communication.*

4.2.1 Communication Lower Bound

Consider the following distribution for the lower bound (note that the distribution we use below is a variant of the distribution described in [Vio15] and [BW12]).

Let $J \in [\frac{n}{2}]$ be uniformly random. $X, Y \in \{0,1\}^n$ are sampled uniformly conditioned on the event that $X_{<J} = Y_{<J}$, *i.e.* the most-significant $J-1$ bits of X and Y are always equal.

Figure 4.2.1: Hard Distribution $p(J, X, Y)$ for the greater-than function

The communication lower bound follows from the following two lemmas. The first lemma says that any protocol computing GT with error at most $1/10000$ must reveal a lot of information about the function value $\mathsf{GT}(X, Y)$.

Lemma 4.10. *If M are the messages of a protocol that computes $\mathsf{GT}(X, Y)$ with error at most $1/10000$, then*

$$\mathbf{I}\left(M : \mathsf{GT}(X,Y) | X_{<J} Y_{<J} J\right) \geq 1 - \frac{1}{10} - \frac{1}{2^{n/2-1}}.$$

Next we show that if the length of the transcript is small, then the protocol could not have revealed a lot of information about $\mathsf{GT}(X,Y)$.

Lemma 4.11. *If $M \in \{0,1\}^\ell$ are the messages of a protocol, then*

$$\mathbf{I}\left(M : \mathsf{GT}(X,Y) | X_{<J} Y_{<J} J\right) \leq \frac{2^{\ell+1}}{n} + \mathsf{h}\left(\frac{1}{4} + \frac{1}{2^{n/2+1}}\right).$$

Proof of Theorem 4.9. Let M be the messages of a deterministic protocol with communication ℓ bits that computes the function GT with error less than $1/10000$ on the input distribution $p(X,Y)$. Applying Lemma 4.10 and Lemma 4.11 with $n > 20$, we have that

$$1 - \frac{1}{10} - \frac{1}{2^{n/2-1}} \leq \mathbf{I}\left(M : \mathsf{GT}(X,Y) | X_{<J} Y_{<J} J\right) \leq \frac{2^{\ell+1}}{n} + \mathsf{h}\left(\frac{1}{4} + \frac{1}{2^{n/2+1}}\right).$$

Since for $n > 20$, we have $\mathsf{h}\left(\frac{1}{4} + \frac{1}{2^{n/2+1}}\right) < 0.84$, simple calculations yield that $\ell = \Omega(\log n)$. □

We proceed with the proofs of Lemma 4.10 and Lemma 4.11.

Proof of Lemma 4.10. By Proposition 2.17, we can write

$$\mathbf{I}\left(M : \mathsf{GT}(X,Y) | X_{<J} Y_{<J} J\right) = \mathbb{E}_{mxyj}\left[\frac{\mathsf{GT}(X,Y)|mx_{<j}y_{<j}j}{\mathsf{GT}(X,Y)|x_{<j}y_{<j}j}\right].$$

Note that $p(\mathsf{GT}(X,Y) = 1 | x_{<j}y_{<j}j) = \frac{1}{2} + \frac{1}{2^{n-j+1}} \leq \frac{1}{2} + \frac{1}{2^{n/2+1}}$. Let $\pi(m)$ denote the output of the protocol when the transcript is m and define the event

$$\mathcal{E} = \{m, x_{<j}, y_{<j}, j \mid p(\mathsf{GT}(X,Y) \neq \pi(m) | m x_{<j} y_{<j} j) \geq 1/100\}.$$

Since, the error of the protocol is at most $1/10000$, Markov's inequality implies that $p(\mathcal{E}) \leq 1/100$.

We can now write

$$\begin{aligned}\mathbf{I}\left(M : \mathsf{GT}(X,Y) | X_{<J} Y_{<J} J\right) &\geq p(\overline{\mathcal{E}}) \mathbb{E}_{mxyj|\overline{\mathcal{E}}}\left[\frac{\mathsf{GT}(X,Y)|mx_{<j}y_{<j}j}{\mathsf{GT}(X,Y)|x_{<j}y_{<j}j}\right] \\ &\geq \frac{99}{100} \cdot \mathsf{d}\left(\frac{99}{100} \middle|\middle| \frac{1}{2} + \frac{1}{2^{n/2+1}}\right) \geq 1 - \frac{1}{10} - \frac{1}{2^{n/2-1}},\end{aligned}$$

where the last inequality follows from Proposition 2.20. □

To prove Lemma 4.11, we need the following lemma. The proof of this lemma is based on a subtle application of chain rule and we will use the same ideas later in the proof of Theorem 4.2.

Lemma 4.12. *If $M \in \{0,1\}^\ell$ are the messages of a communication protocol, then*

$$\mathbf{I}\left(M : X_J | X_{<J} Y_{<J} J\right) \leq \frac{2^{\ell+1}}{n}.$$

We first use Lemma 4.12 to prove Lemma 4.11 and then present the proof of Lemma 4.12.

Proof of Lemma 4.11. By the chain rule for mutual information, we have

$$\begin{aligned}
\mathbf{I}\left(M : \mathsf{GT}(X,Y) | X_{<J} Y_{<J} J\right) &\leq \mathbf{I}\left(M : \mathsf{GT}(X,Y) X_J | X_{<J} Y_{<J} J\right) \\
&= \mathbf{I}\left(M : X_J | X_{<J} Y_{<J} J\right) + \mathbf{I}\left(M : \mathsf{GT}(X,Y) | X_{\leq J} Y_{<J} J\right) \\
&\leq 2^{\ell+1}/n + \mathbf{I}\left(M : \mathsf{GT}(X,Y) | X_{\leq J} Y_{<J} J\right),
\end{aligned}$$

where the last inequality follows from Lemma 4.12.

By Proposition 2.7, we have $\mathbf{I}\left(M : \mathsf{GT}(X,Y) | X_{<J} Y_{<J} J, X_J = 0\right) \leq \mathsf{h}\left(\frac{1}{4} + \frac{1}{2^{n/2+1}}\right)$, where the inequality follows since after conditioning on $j, x_{<j}, y_{<j}, x_j = 0$, the event $\mathsf{GT}(X,Y) = 1$ implies that either $Y_J = 0, X_{J+1} = 0$ or $X = Y$, so by the union bound, the probability is at most $1/4 + 2^{-n/2-1}$. Similarly, it holds that $\mathbf{I}\left(M : \mathsf{GT}(X,Y) | X_{<J} Y_{<J} J, X_J = 1\right) \leq \mathsf{h}\left(\frac{3}{4} - \frac{1}{2^{n/2+1}}\right) = \mathsf{h}\left(\frac{1}{4} + \frac{1}{2^{n/2+1}}\right)$. Therefore,

$$\mathbf{I}\left(M : \mathsf{GT}(X,Y) | X_{\leq J} Y_{<J} J\right) \leq \mathsf{h}\left(\frac{1}{4} + \frac{1}{2^{n/2+1}}\right). \qquad \square$$

Proof of Lemma 4.12. Since $X_{<J} = Y_{<J}$, we have

$$\mathbf{I}\left(M : X_J | X_{<J} Y_{<J} J\right) = \mathbf{I}\left(M : X_J | X_{<J} J\right) = \sum_m p(m) \mathbb{E}_{xj|m} \left[\frac{X_j | m x_{<j} j}{X_j | x_{<j} j} \right]. \tag{4.2.1}$$

Recall that any message m induces a rectangle $\mathcal{S}_m \times \mathcal{T}_m$ in the input space as given by Proposition 4.6. Denoting by $\mathcal{S}_m$ (and $\mathcal{T}_m$) the event that $X \in \mathcal{S}_m$ (and $Y \in \mathcal{T}_m$), Proposition 4.6 implies that $M = m$ is equivalent to the events $\mathcal{S}_m \wedge \mathcal{T}_m$. After fixing $x_{<j} j$, X is independent of Y (and hence $\mathcal{T}_m$). So by Proposition 4.7, we have $p(X_j | m x_{<j} j) = p(X_j | \mathcal{S}_m x_{<j} j)$.

Using the above observation, the right hand side in (4.2.1) can be rewritten as

$$\sum_m p(\mathcal{S}_m)p(\mathcal{T}_m|\mathcal{S}_m)\mathbb{E}_{xj|\mathcal{S}_m\mathcal{T}_m}\left[\frac{X_j|\mathcal{S}_m x_{<j}j}{X_j|x_{<j}j}\right] \leq \sum_m p(\mathcal{S}_m)\mathbb{E}_{xj|\mathcal{S}_m}\left[\frac{X_j|\mathcal{S}_m x_{<j}j}{X_j|x_{<j}j}\right], \quad (4.2.2)$$

where the inequality follows from the fact that $\mathbb{E}_a[h(a)] \geq p(\mathcal{W})\mathbb{E}_{a|\mathcal{W}}[h(a)]$, for any non-negative function h.

Since J is independent of X, we have $p(XJ) = p(J)p(X)$ and also $p(XJ|\mathcal{S}_m) = p(J)p(X|\mathcal{S}_m)$. Using that $p(J)$ is uniform over $[\frac{n}{2}]$, we can use the chain rule to write the inner expectation as

$$\mathbb{E}_{xj|\mathcal{S}_m}\left[\frac{X_j|\mathcal{S}_m x_{<j}j}{X_j|x_{<j}j}\right] = \mathbb{E}_j\mathbb{E}_{x|\mathcal{S}_m}\left[\frac{X_j|\mathcal{S}_m x_{<j}}{X_j|x_{<j}}\right] = \frac{2}{n}\frac{X_{\leq\frac{n}{2}}|\mathcal{S}_m}{X_{\leq\frac{n}{2}}}.$$

Now we can bound the right hand side in (4.2.2) by

$$\sum_m p(\mathcal{S}_m)\frac{2}{n}\frac{X_{\leq\frac{n}{2}}|\mathcal{S}_m}{X_{\leq\frac{n}{2}}} \leq \frac{2}{n}\sum_m p(\mathcal{S}_m)\log\frac{1}{p(\mathcal{S}_m)} \leq \frac{2^{\ell+1}}{n},$$

where the first inequality follows from Proposition 2.22 (with $A = X_{\leq n/2}, B = \perp, \mathcal{W} = \mathcal{S}_m$) and the second from the fact that for $0 \leq \gamma \leq 1$, it holds that $\gamma\log(1/\gamma) \leq \frac{\log e}{e} < 1$. □

This finishes the proof of Theorem 4.9.

4.3 Separating Information and Communication

Now we come back to the main topic of this chapter — separating information and communication complexity. We start by defining the k-ary pointer jumping function and the input distribution in Section 4.3.1. We then discuss some basic protocols for the k-ary pointer jumping function in Section 4.3.3. Following this in Section 4.3.4, we give more details on Theorem 4.1 and Theorem 4.2. We prove the communication lower bound in Section 4.4. In Section 4.5, we bound the information complexity of the k-ary pointer jumping problem.

4.3.1 k-ary Pointer Jumping Function

For a parameter $k \in \mathbb{N}, k \geq 2$, we work with the alphabet $[k] = \{1, 2, \ldots, k\}$. Let $X, Y : [k]^{<n} \to [k]$ be functions mapping strings of length less than n to a single character. Let $F, G : [k]^n \to \{0,1\}$ be boolean functions. For $z \in [k]^n$, let $z_{\leq r}$ denote the prefix of z of length r. In the k-ary pointer jumping problem, the first party is given X, F, and the second is given Y, G. The goal of the parties is to compute $F(z) + G(z) \bmod 2$, where $z \in [k]^n$ is the unique string satisfying the n equations

$$X(z_{\leq r}) + Y(z_{\leq r}) = z_{r+1} \bmod k,$$

for every $r \in \{0, 1, \ldots, n-1\}$.

4.3.2 Input and Fooling Distributions

For $z \in [k]^{<n}, J \in \{0, 1, \ldots, n-1\}$ and X, Y as above, we say z is *consistent* with X, Y and J, if $|z| \geq J + 1$, and

$$X(z_{\leq J}) + Y(z_{\leq J}) = z_{J+1} \bmod k.$$

Since X, Y and J will usually be clear from the context, we just write z is consistent and omit X, Y, J.

The distribution on inputs, described in Figure 4.3.1, ensures that $F(z) + G(z) \bmod 2$ is the same for *every* consistent z, so it is enough for the parties to find a consistent z to complete the goal. Also, note that even though the parties do not know the value of J, with probability at least $1 - 1/k$, the value of J is determined by $X(z)$ and $Y(z)$ for any fixed $z \in [k]^n$.

Comparison with the Bursting Noise Function We remark that although our formulation allows us to define the k-ary pointer jumping function and the input distribution much more compactly than the bursting noise function defined in [GKR16], our function is just a simpler variant of their construction. The main difference is that our function can be thought of as being computed over a k-ary tree as opposed to a binary tree (as in the work of [GKR16]) and it is symmetric with respect to both parties since we are taking the sum modulo k of the values computed by both parties (as opposed

Fooling Distribution $p(J, X, F, Y, G)$**:** Index $J \in \{0, \ldots, n-1\}$ is sampled uniformly at random. Functions $X, Y : [k]^{<n} \to [k]$ are sampled uniformly at random subject to the constraint that for any $z \in [k]^{<J}$, $X(z) = Y(z)$. Functions $F, G : [k]^n \to \{0,1\}$ are uniformly random.

Input Distribution $q(J, X, F, Y, G)$**:** Let $\mathcal{E}_0$ denote the event that for all consistent z, $X(z) = Y(z)$ and $F(z) = G(z)$ (when $|z| = n$). Let $\mathcal{E}_1$ denote the event that for all consistent z, $X(z) = Y(z)$ and $F(z) \neq G(z)$ (when $|z| = n$). In the distribution $q_0(J, X, F, Y, G)$, J is sampled uniformly from $\{0, 1, \ldots, n-1\}$, and the rest of the variables are sampled according to the distribution of $p(J, X, F, Y, G|\mathcal{E}_0, j)$. In the distribution $q_1(J, X, F, Y, G)$, J is sampled uniformly, and the rest of the variables are sampled uniformly from the distribution $p(J, X, F, Y, G|\mathcal{E}_1, j)$. The input is sampled by sampling from $q_0(J, X, F, Y, G)$ with probability $\frac{1}{2}$ and from $q_1(J, X, F, Y, G)$ with probability $\frac{1}{2}$.

Figure 4.3.1: Distributions for the k-ary pointer jumping problem. Recall that the inputs to the first and second parties are X, F and Y, G respectively.

to each party going down the tree alternately as in the work of [GKR16]). Our distribution is also very similar to the distribution used by [GKR16].

4.3.3 Simple Protocols

To get a better understanding of the above function, let us present a few simple protocols to compute it. The first protocol below has small error on all distributions but the rest only work for the input distribution $q(X, F, Y, G)$.

Trivial Protocol There is a trivial protocol for this problem that has worst-case communication $O(n \log k)$: in step r, Alice and Bob send each other the values $X(z_{\leq r}), Y(z_{\leq r})$, until they have computed z. The parties can compute $F(z) + G(z) \bmod 2$ with two more bits of communication. Note that this protocol succeeds with zero error under any input distribution.

Binary Search Protocol Under the input distribution $q(J, X, F, Y, G)$, for any $z \in [k]^n$, we have that $X(z_{<J}) = Y(z_{<J})$ with probability 1, while $X(z_{\leq J})$ and $Y(z_{\leq J})$ are different with probability $1 - \frac{1}{k}$. The players can use a version of binary search [FRPU94] to find the first difference as follows: they construct a binary search tree where each node represents an interval $[a, b]$ and its children represent the two sub-intervals $[a, \frac{a+b}{2}]$ and $[\frac{a+b}{2}, b]$. The players repeat the following $O\left(\log\left(\frac{n \log k}{\varepsilon}\right)\right)$ many times: if the players are currently at some node v of this search tree, they exchange $O(1)$ bits of hashes, after applying the hash function to the prefix of length a. If the hashes agree and show that the first difference lies among the children of v, then the players move on to that node. If the hashes disagree, then the players move to the parent of the v. When the players reach a leaf which corresponds to an interval $[a, a]$, they output a, otherwise they output that the strings are equal. It is easy to check using standard concentration bounds that this protocol finds the first difference with error ε if the strings are different.

It follows that with probability at least $1 - \left(\varepsilon + \frac{1}{k}\right)$, the players can compute the index J with $O\left(\log\left(\frac{n \log k}{\varepsilon}\right)\right)$ bits of communication. With an additional $2 \log k$ bits of communication the players can then find a consistent z (satisfying $X(z_{\leq J}) + Y(z_{\leq J}) = z_{J+1} \bmod k$) which suffices

to compute the value of the function on the input distribution $q(J, X, F, Y, G)$. This protocol has communication $O\left(\log\left(\frac{n \log k}{\varepsilon}\right) + \log k\right)$ and the error probability is at most $\varepsilon + \frac{1}{k}$ on the input distribution $q(X, F, Y, G)$.

Sampling Protocol Let $t = \Theta\left(k^2 \log\left(\frac{1}{\varepsilon}\right)\right)$. Using shared randomness, the players draw a subset $\mathcal{S}$ by choosing t strings uniformly at random from $[k]^n$, exchange the values of $F(z)$ and $G(z)$ for each $z \in \mathcal{S}$ and output the majority of the bits $\{F(z) + G(z) \bmod 2\}_{z \in \mathcal{S}}$. Under the input distribution $q(J, X, F, Y, G)$, the probability that a random string is consistent (that it satisfies $X(z_{\leq J}) + Y(z_{\leq J}) = z_{J+1} \bmod k$) is $\frac{1}{k}$, so using standard concentration bounds, this protocol has error at most ε under the distribution $q(J, X, F, Y, G)$. Moreover, the communication is $O\left(k^2 \log\left(\frac{1}{\varepsilon}\right)\right)$.

4.3.4 Information and Communication Complexity

We prove that there is a low information solution for this task, with internal information cost $O\left((\log k + \log\log n) \cdot 2^{\frac{2 \log n}{k}}\right)$ on the distribution $q(X, F, Y, G)$ (Theorem 4.1) but any randomized protocol that errs with a constant probability on the input distribution $q(X, F, Y, G)$ requires communication at least $\Omega(\min\{k, \log n\})$ (Theorem 4.2). The lower bound is tight up to polynomial factors, as the binary search and sampling protocol described previously show that the communication complexity is $O(\min\{k^2, \log n + \log k\})$. Setting $n = 2^k$, we get a correct protocol with information cost $O(\log k)$ even though no protocol with communication $\Omega(k)$ can succeed.

Low Information Protocol The low information protocol for the problem is quite similar to the trivial protocol, so let us first discuss the information cost of the trivial protocol under the distribution $q(X, F, Y, G)$. At the beginning of the protocol, with just their own inputs in hand, the parties do not have any information about J. Using the trivial protocol, both parties learn the value of J with high probability. This happens because J is close to the first point at which their inputs disagree. Since the entropy of J is $\Theta(\log n)$ bits and J is determined with high probability given X and Y, it is not too hard to argue that the parties learn $\Omega(\log n)$ bits of information about each other's inputs. It follows

that the internal information cost of the trivial protocol is $\Omega(\log n)$ bits, much larger than what we are aiming for.

The low information protocol adds some noise to hide the value of J. In step r of the low information protocol, the parties send each other the value $X(z_{\leq r}), Y(z_{\leq r})$ with probability $1 - \frac{1}{\log n}$ and send a uniformly random value otherwise. The parties abort the protocol if they experience $\frac{\log n}{\log \log n}$ rounds where the messages they sent were not the same. The distribution on inputs ensures that they will sample a consistent z with high probability. When the parties sample a consistent z, the messages sent are almost always sampled from a distribution that the receiving party knows, while if they sample a z that is not consistent, the protocol aborts shortly after the inconsistency. These properties can be used to show that under the distribution $q(X, F, Y, G)$, the information cost of the protocol is $O((\log k + \log \log n) \cdot 2^{\frac{2 \log n}{k}})$ and the error probability is at most $\frac{4}{\log n}$.

Intuitively, this protocol does not reveal a lot of information about J because at the end of the protocol the parties see a lot of disagreements, and so they only learn that J belongs to some set of density $\frac{1}{\log n}$. Heuristically, the entropy of J conditioned on the entire exchange is $\approx \log(n/\log n) = \log n - \log \log n$ and the amount of information revealed about J is $O(\log \log n)$.

In Section 4.5, we analyze the aforementioned low information protocol and prove Theorem 4.1. Before moving on, we remark that when $n = 2^k$, using this low information-cost protocol with the simulation result of Braverman [Bra12], one can obtain a deterministic protocol with communication complexity k^{c/ε^2} for a constant $c > 2$, that computes the k-ary pointer jumping function with error probability $O\left(\varepsilon + \frac{1}{k}\right)$ on the distribution $q(X, F, Y, G)$.

Communication Lower Bound Consider any protocol with ℓ bits of communication that solves the k-ary pointer jumping problem. Without loss of generality, we may assume that the protocol is deterministic, since any randomness can always be fixed to obtain a deterministic protocol that succeeds with high probability. Let M denote the messages of the protocol. Our input distribution $q(X, F, Y, G)$ has the property that under the inputs drawn from this distribution the k-ary pointer jumping function is balanced (it takes values 1 and 0 with probability half each). Hence, if the protocol solved the k-ary pointer jumping problem with error δ on the distribution $q(X, F, Y, G)$, then

the statistical distance between $q_0(M)$ (the induced distribution on M when inputs are drawn from $q(X,F,Y,G)$ conditioned on the value of k-ary pointer jumping function being 0) and $q_1(M)$ (the induced distribution on M when inputs are drawn from $q(X,F,Y,G)$ conditioned on the value of k-ary pointer jumping function being 1) would be at least $1-2\delta$ (see Lemma 4.8), since the distributions $q_0(M)$ and $q_1(M)$ have nearly disjoint supports.

To prove the communication lower bound, we define the *fooling distribution* $p(X,F,Y,G)$ on inputs, and using information theoretic techniques show that if the communication complexity ℓ of the protocol is much less than $\min\{k, \log n\}$, then both of the distributions $q_0(M)$ and $q_1(M)$ are close to the fooling distribution $p(M)$, which implies that the statistical distance between $q_0(M)$ and $q_1(M)$ is close to 0. This will give us a contradiction, since we argued in the paragraph above that these distributions must be far apart.

It will be convenient to state our results in terms of the function $\eta : [0,\infty) \to [0,1]$ defined as

$$\eta(\alpha) = \begin{cases} 0 & \text{if } \alpha = 0, \\ \alpha \log(1/\alpha) & \text{if } \alpha \in (0, 1/e), \\ \frac{\log e}{e} & \text{if } \alpha \geq 1/e. \end{cases} \tag{4.3.1}$$

One can check that η is non-decreasing, continuous, and concave, and for $0 \leq \alpha \leq 1$, we have that $\alpha \log(1/\alpha) \leq \eta(\alpha)$. We prove:

Theorem 4.13. *For any deterministic protocol for the k-ary pointer jumping function with communication complexity at most ℓ, we have that*

$$q_0(M) \overset{\gamma}{\approx} p(M) \overset{\gamma}{\approx} q_1(M), \text{ with } \gamma = 4\left(2e\ell/k + 2\ell\sqrt{2^\ell/n} + \eta\left(\sqrt{2^\ell/n}\right)\right)^{1/3}.$$

We stress that the fooling distribution $p(X,F,Y,G)$ is only used in the analysis. The inputs to the protocol come from the distribution $q(X,F,Y,G)$.

Theorem 4.13 implies that $\ell = \Omega(\min\{k, \log n\})$ for any protocol that solves the k-ary pointer jumping problem for the input distribution $q(X,F,Y,G)$. This gives us Theorem 4.2.

Proof of Theorem 4.2. We may assume that $\max\{\frac{1}{k}, \frac{1}{\log n}\} \leq \varepsilon^3 \leq \frac{1}{2}$ since otherwise the statement is trivial. Consider any protocol for the k-ary pointer jumping function that has communication com-

plexity $\ell := \varepsilon^3 \cdot \min\{k, \log n\}$ and errs with probability δ on the input distribution $q(X, F, Y, G)$. We may assume without loss of generality that the protocol is deterministic, since if the protocol used public or private randomness, we can fix the randomness to get a deterministic protocol with the same communication complexity and the same error on the input distribution $q(X, F, Y, G)$. Since, the protocol has error at most δ, the statistical distance between $q_0(M)$ and $q_1(M)$ must be at least $1 - 2\delta$ (see Lemma 4.8). On the other hand, Theorem 4.13 implies that

$$|q_0(M) - q_1(M)| \leq |q_0(M) - p(M)| + |p(M) - q_1(M)| \leq 2\gamma,$$

where $\gamma = 4\left(2e\ell/k + 2\ell\sqrt{2^\ell/n} + \eta\left(\sqrt{2^\ell/n}\right)\right)^{1/3}$. The first term inside the paranthesis is at most $6\varepsilon^3$. The second term can be bounded by

$$\max\{2\varepsilon^3 k\sqrt{2^{-k+\varepsilon^3 k}}, 2\varepsilon^3 \log(n)\sqrt{n^{-1+\varepsilon^3}}\} \leq \max\{2^{-k/4}, n^{-1/4}\} \leq \varepsilon^3$$

for large enough k and n. The last term is at most

$$\max\{\eta\left(\sqrt{2^{-k+\varepsilon^3 k}}\right), \eta\left(\sqrt{n^{-1+\varepsilon^3}}\right)\} \leq \max\{2^{-k/4}, n^{-1/4}\} \leq \varepsilon^3.$$

It follows that $\gamma < 8\varepsilon$ for large enough values of k and n. Then, it must be that $1 - 2\delta \leq 2\gamma < 16\varepsilon$ and hence, the error $\delta > \frac{1}{2} - 8\varepsilon$. □

4.4 Communication Lower Bound for k-ary Pointer Jumping

4.4.1 High-level Proof Sketch for Theorem 4.13

Consider an ℓ-bit deterministic protocol. Our goal is to show that if $\ell \ll \min\{k, \log n\}$, then the induced distribution of the messages M of the protocol is roughly the same under the fooling distribution $p(X, F, Y, G)$ and the distribution $q_0(X, F, Y, G)$ (input distribution $q(X, F, Y, G)$ conditioned on the value of k-ary pointer jumping function being 0) and that a similar statement is true for $p(X, F, Y, G)$ and $q_1(X, F, Y, G)$. As described in the previous section, this proves that communication complexity must be $\Omega(\min\{k, \log n\})$.

How are the distributions $p(J, X, F, Y, G)$ and $q_0(J, X, F, Y, G)$ related? Let S denote the set of consistent strings $z \in [k]^{\leq n}$ and $X_S Y_S F_S G_S$ denote the projection of the random variables $XYFG$ on the strings in the set S and let $X_{\leq J}, Y_{\leq J}$ denote the restriction of X, Y to inputs of length at most J. Let X_J, Y_J denote the restriction of X, Y to inputs of length J.

First note that the set of consistent strings S is determined by $X_J Y_J J$, since given $X_J Y_J J$, one can check whether any string $z \in [k]^{\leq n}$ is consistent or not (whether it satisfies $|z| > J + 1$ and $X(z_{\leq J}) + Y(z_{\leq J}) = z_{J+1} \bmod k$). Since the distribution $p(J, X, F, Y, G)$ and $q_0(J, X, F, Y, G)$ are the same when projected on $X_{\leq J} Y_{\leq J} J$, it follows that the distributions $p(S)$ and $q_0(S)$ are also the same. Fixing the values $X_{\leq J} Y_{\leq J} J$, the distribution $q_0(J, X, F, Y, G)$ is obtained from $p(J, X, F, Y, G)$ by conditioning on the event $X_S F_S = Y_S G_S$. Similarly after fixing $X_{\leq J} Y_{\leq J} J$, the distribution $q_1(J, X, F, Y, G)$ is obtained from $p(J, X, F, Y, G)$ by conditioning on the event $X_S F_S = Y_S \overline{G}_S$, where $\overline{G}$ is the function $1 - G$.

So, after fixing $X_{\leq J} Y_{\leq J} J$, to prove that $p(M) \approx q_0(M)$, we need to show that the distribution $p(M)$ does not change by much, even if we condition on the event $X_S F_S = Y_S G_S$ (and similarly $X_S F_S = Y_S \overline{G}_S$). We prove this in three steps:

- First, we argue using a subtle application of the chain rule, that if $\ell \ll \min\{k, \log n\}$, the protocol does not reveal much information about S under the *fooling distribution* $p(X, F, Y, G)$.
- Next, we show that since the players do not learn much information about S, they also do not learn much information about $X_S F_S$ and $Y_S G_S$ in the *fooling distribution* $p(X, F, Y, G)$. To argue this, we prove a generalization of Shearer's Lemma [CGFS86, Rado3].
- Lastly, using the above facts we show that the distribution $p(M)$ roughly stays the same even after conditioning on the event $X_S F_S$ and $Y_S G_S$.

Note that it is only during the last step that the input distribution $q_0(X, F, Y, G)$ (and $q_1(X, F, Y, G)$) are related to the fooling distribution $p(X, F, Y, G)$. The first two steps analyze the behavior of the protocol only on the fooling distribution $p(X, F, Y, G)$. Next we elaborate on each of the steps.

Information Revealed about Consistent Strings

As discussed before, the set S of consistent strings is determined by $X_{\leq J}Y_{\leq J}J$ in the fooling distribution $p(X, F, Y, G)$. We bound the amount of information revealed about S to each party as follows:

Lemma 4.14. *Let M denote the messages of a deterministic ℓ-bit protocol, then*

$$\mathbf{I}_p\left(M : S|Y_{\leq J}J\right) \leq \mathbf{I}_p\left(M : X_J|Y_{<J}J\right) \leq \frac{2^\ell}{n}$$
$$\mathbf{I}_p\left(M : S|X_{\leq J}J\right) \leq \mathbf{I}_p\left(M : Y_J|X_{<J}J\right) \leq \frac{2^\ell}{n}.$$

Since in the distribution $p(X, F, Y, G)$, we have $X_{<J} = Y_{<J}$, it follows that $\mathbf{I}_p\left(M : X_J|Y_{<J}J\right) = \mathbf{I}_p\left(M : X_J|X_{<J}J\right)$. At first glance, one might think that the expression $\mathbf{I}_p\left(M : X_J|X_{<J}J\right)$ can be upper bounded by ℓ/n using the chain rule (by averaging over all n values of J). However, this is not true: since J is essentially determined by X and Y and the messages M contain information about X and Y, it is not clear if one can directly use the chain rule as simply as above. To understand this point in more detail, it is worthwhile to consider a simple example.

Consider the following variant of the binary search protocol for the k-ary pointer jumping problem from Section 4.3.3. In that section, we discussed the protocol on the input distribution $q(J, X, F, Y, G)$. However the binary search phase of the protocol to find the index J works pretty much the same on the fooling distribution $p(J, X, F, Y, G)$. Using $O(\log n)$ bits the players determine the value of J with a version of binary search and then exchange one bit about the values of X_J and Y_J. So, in this case the $O(\log n)$-bit protocol reveals at least 1 bit of information about X_J and Y_J. So, one could not hope to get a bound like ℓ/n since this $O(\log n)$-bit protocol already reveals 1 bit of information. In fact, this protocol also shows that the above lemma is tight. If we stop the binary search phase after $O(\ell)$ bits, the players will know a set of size $\frac{n}{2^\ell}$ in which J lies. If the players choose a random index K in this set and reveal one bit of information about X_K, then the amount of information revealed about X_J is exactly $\frac{2^\ell}{n}$ as the players reveal one bit about X_J with probability $\frac{2^\ell}{n}$.

Despite the above, we are able to prove Lemma 4.14 by a subtle application of the chain rule

and the Markov property of protocols. Essentially, the proof argues that for each of the 2^ℓ possible transcripts, the protocol, on average, only reveals $1/n$ bit of information.

Information Revealed about $X_S F_S$ and $Y_S G_S$

Next, we want show that the parties do not learn much information about the values of X, F, Y, G restricted to S (denoted X_S, F_S, Y_S, G_S). Note that only $1/k$ fraction of the strings are in S, since for every z, $p(z \in S | J = j, X_{\leq J} = x_{\leq j}) \leq 1/k$. So, recalling the classical Shearer's Lemma [CGFS86, Rado3], one might hope that $\mathbf{I}(M : X_S F_S)$ can be bounded by $\mathbf{I}(M : XF)/k \leq \ell/k$. If S were independent of M, X, F, such a bound would be easy to prove using the chain rule for information (see Lemma 8 in [GKR16]).

In our case, S is not independent of M but almost independent, as the amount of information that M reveals about S is small (after conditioning on $X_{<J}J$). So it is conceivable that $\mathbf{I}(M : X_S F_S)$ can still be bounded by $O(\mathbf{I}(M : XF)/k)$ plus some small error terms. We prove such a generalization of Shearer's Lemma which may be of independent interest. Below $U_{S'}$ denotes the restriction of U to the coordinates in S'. We show:

Lemma 4.15. *Given a probability space p', let $U = (U_1, \ldots, U_t)$ where $U_1, \ldots, U_t$ are mutually independent random variables. Let random variables $C \in \{0,1\}^\ell, S' \subseteq [t]$, and V be such that U is independent of $S'V$, and $U - C - S'V$ and for all $i \in [t]$, $p'(i \in S') \leq 1/k$. Then*

$$\mathbf{I}(C : U_{S'} | V S') \leq \ell \cdot \left(\frac{2e}{k} + 2\sqrt{\mathbf{I}(C : S')} \right) + \eta\left(\sqrt{\mathbf{I}(C : S')} \right).$$

Conditioned on any fixing of $X_{\leq J} Y_{\leq J} J$, we have that XF and YG are independent under p, and since M denotes the messages in a communication protocol, the Markov property (Proposition 4.7) implies that $XF - M - YG$ holds. Thus, Lemma 4.15 in conjunction with Lemma 4.14 and convexity of η can be used to show that the amount of information the messages reveal about $X_S F_S$ (and similarly for $Y_S G_S$) is small:

$$\begin{aligned} \mathbf{I}_p(M : X_S F_S | X_{\leq J} Y_{\leq J} J) &\leq 2e\ell/k + 2\ell\sqrt{2^\ell/n} + \eta\left(\sqrt{2^\ell/n}\right), \\ \mathbf{I}_p(M : Y_S G_S | X_{\leq J} Y_{\leq J} J) &\leq 2e\ell/k + 2\ell\sqrt{2^\ell/n} + \eta\left(\sqrt{2^\ell/n}\right). \end{aligned} \tag{4.4.1}$$

Conditioning on the event $X_S F_S = Y_S G_S$

Lastly, we need to show that the distribution $p(M)$ does not change much after conditioning on the event $X_S F_S = Y_S G_S$ which gives us the distribution $q_0(M)$. Since, both random variables $X_S F_S$ and $Y_S G_S$ are essentially independent of M (the mutual information as in (4.4.1) is small), one can expect that conditioning on the event $X_S F_S = Y_S G_S$ should not change the distribution of M by much.

In fact, we show the following general statement:

Lemma 4.16. *Given a probability space p', if $A, B \in [t]$ are uniform and independent random variables, and $A - C - B$, then*

$$p'(C) \overset{\epsilon}{\approx} p'(C|A = B), \text{ with } \epsilon = 2\mathbf{I}_{p'}(C : A)^{1/3} + 2\mathbf{I}_{p'}(C : B)^{1/3}.$$

Lemma 4.16 together with (4.4.1) and another convexity argument completes the proof of Theorem 4.13.

4.4.2 Proof of Theorem 4.13

We shall prove that $p(M) \overset{\gamma}{\approx} q_0(M)$. The bound for the distribution $q_1(M)$ is analogous. We first give the proof assuming Lemmas 4.14, 4.15, and 4.16 and then we prove the lemmas.

By Lemma 4.14, $\mathbf{I}_p(M : S|X_{\leq J}J) \leq 2^\ell/n$. After fixing $x_{\leq j}, j$, S is determined by Y_J. For any such fixing, we have that $p(z \in S|x_{\leq j}j) \leq 1/k$ holds for any string z, and that XF is independent of YG. Furthermore, by Proposition 4.7 we also have $XF - M - SY_J$ after fixing $x_{\leq j}, j$. Thus we can apply Lemma 4.15 with $U = XF$, $C = M$, $V = Y_J$ and $S' = S$, to conclude that

$$\mathbf{I}_p(M : X_S F_S|SY_J x_{\leq j}j) \leq 2e\ell/k + 2\ell\sqrt{\mathbf{I}_p(M : S|x_{\leq j}j)} + \eta\left(\sqrt{\mathbf{I}_p(M : S|x_{\leq j}j)}\right).$$

Taking the expectation over the choice of $x_{\leq j}j$, and using the concavity of the square-root and η:

$$\begin{aligned}\mathbf{I}_p(M : X_S F_S|Y_{\leq J}X_{\leq J}J) &\leq 2e\ell/k + 2\ell\sqrt{\mathbf{I}_p(M : S|X_{\leq J}J)} + \eta\left(\sqrt{\mathbf{I}_p(M : S|X_{\leq J}J)}\right)\\ &\leq 2e\ell/k + 2\ell\sqrt{2^\ell/n} + \eta\left(\sqrt{2^\ell/n}\right).\end{aligned}$$

The same bound applies to $\mathbf{I}_p\left(M : Y_S G_S | Y_{\leq J} X_{\leq J} J\right)$. For each fixing of $x_{\leq j} y_{\leq j} j$, we have that $X_S F_S - M - Y_S G_S$. Thus ,we can apply Lemma 4.16 to conclude that

$$\begin{aligned}|p(M|x_{\leq j}y_{\leq j}j) - p(M|x_{\leq j}y_{\leq j}j, X_S F_S = Y_S G_s)| \\ \leq 2\sqrt[3]{\mathbf{I}_p\left(M : X_S F_S | x_{\leq j} y_{\leq j} j\right)} + 2\sqrt[3]{\mathbf{I}_p\left(M : Y_S G_S | x_{\leq j} y_{\leq j} j\right)}.\end{aligned}$$

Since $p(X_{\leq J} Y_{\leq J} J) = q_0(X_{\leq J} Y_{\leq J} J)$, we can use Proposition 2.2 to bound

$$\begin{aligned}|p(M) - q_0(M)| &= \mathbb{E}_{p(x_{\leq j} y_{\leq j} j)}\left[|p(M|x_{\leq j}y_{\leq j}j) - p(M|x_{\leq j}y_{\leq j}j, X_S F_S = Y_S G_S)|\right] \\ &\leq \mathbb{E}_{p(x_{\leq j} y_{\leq j} j)}\left[2\sqrt[3]{\mathbf{I}_p\left(M : X_S F_S | x_{\leq j} y_{\leq j} j\right)} + 2\sqrt[3]{\mathbf{I}_p\left(M : Y_S G_S | x_{\leq j} y_{\leq j} j\right)}\right] \\ &\leq 2\sqrt[3]{\mathbf{I}_p\left(M : X_S F_S | X_{\leq J} Y_{\leq J} J\right)} + 2\sqrt[3]{\mathbf{I}_p\left(M : Y_S G_S | X_{\leq J} Y_{\leq J} J\right)} \\ &\leq 4\sqrt[3]{2e\ell/k + 2\ell\sqrt{2^\ell/n} + \eta\left(\sqrt{2^\ell/n}\right)} = \gamma,\end{aligned}$$

where the second to last inequality follows from the concavity of 3^{rd}-root over non-negative reals.

Proof of Lemma 4.14

Once $Y_{\leq J} J$ are fixed, S is determined by X_J, since whether a string z is consistent or not is determined given $X_J Y_J J$. Furthermore, after $Y_{<J} J$ is fixed, XF and YG are independent in the distribution $p(J, X, F, Y, G)$, so the Markov property of protocols (Proposition 4.7) implies that $X_J - M - Y_J$ (conditioned on $X_{<J} J$). Thus, using Propositions 2.14 and 2.16, we get that

$$\mathbf{I}_p\left(M : S | Y_{\leq J} J\right) \leq \mathbf{I}_p\left(M : X_J | Y_{\leq J} J\right) = \mathbf{I}_p\left(M : X_J | Y_{<J} J\right). \tag{4.4.2}$$

Since $X_{<J} = Y_{<J}$, we have that $\mathbf{I}_p\left(M : X_J | Y_{<J} J\right) = \mathbf{I}_p\left(M : X_J | X_{<J} J\right)$, which we shall show is at most $2^\ell/n$. Recall that any message m induces a rectangle $\mathcal{S}_m \times \mathcal{T}_m$ in the input space as given by Proposition 4.6. Denoting by $\mathcal{S}_m$ (and $\mathcal{T}_m$) the event that $X \in \mathcal{S}_m$ (and $Y \in \mathcal{T}_m$), Proposition 4.6 implies that $M = m$ is equivalent to the events $\mathcal{S}_m \wedge \mathcal{T}_m$. Also, since XF and YG are independent given $x_{<j} j$, by Proposition 4.7, we have $p(X|m, x_{<j} j) = p(X|\mathcal{S}_m \wedge \mathcal{T}_m, x_{<j} j) = p(X|\mathcal{S}_m, x_{<j} j)$. So, we can write

$$\mathbf{I}_p\left(M : X_J | X_{<J} J\right) = \mathbb{E}_{yxj|m}\left[\frac{X_j | m x_{<j} j}{X_j | x_{<j} j}\right] = \mathbb{E}_{yxj|m}\left[\frac{X_j | \mathcal{S}_m, x_{<j} j}{X_j | x_{<j} j}\right]. \tag{4.4.3}$$

The right hand side in (4.4.3) can be rewritten as

$$\sum_m p(\mathcal{S}_m)p(\mathcal{T}_m|\mathcal{S}_m)\mathbb{E}_{xj|\mathcal{S}_m,\mathcal{T}_m}\left[\frac{X_j|\mathcal{S}_m, x_{<j}j}{X_j|x_{<j}j}\right] \leq \sum_m p(\mathcal{S}_m)\mathbb{E}_{xj|\mathcal{S}_m}\left[\frac{X_j|\mathcal{S}_m, x_{<j}j}{X_j|x_{<j}j}\right], \quad (4.4.4)$$

where the inequality follows from the fact that $\mathbb{E}_a[h(a)] \geq p(\mathcal{W})\mathbb{E}_{a|\mathcal{W}}[h(a)]$, for any non-negative function h.

Since J is independent of X, we have $p(XJ) = p(J)p(X)$ and also $p(XJ|\mathcal{S}_m) = p(J)p(X|\mathcal{S}_m)$. So, we can write

$$\mathbb{E}_{xj|\mathcal{S}_m}\left[\frac{X_j|\mathcal{S}_m, x_{<j}j}{X_j|x_{<j}j}\right] = \sum_{j=1}^{n} p(j)\mathbb{E}_{x|\mathcal{S}_m}\left[\frac{X_j|\mathcal{S}_m, x_{<j}}{X_j|x_{<j}}\right].$$

Since $p(J)$ is uniform we can use the chain rule to write the right hand side above as

$$\sum_{j=1}^{n} p(j)\mathbb{E}_{x|\mathcal{S}_m}\left[\frac{X_j|\mathcal{S}_m, x_{<j}}{X_j|x_{<j}}\right] = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_{x|\mathcal{S}_m}\left[\frac{X_j|\mathcal{S}_m, x_{<j}}{X_j|x_{<j}}\right] = \frac{1}{n}\cdot\frac{X|\mathcal{S}_m}{X}.$$

Plugging the above into (4.4.4), we get that the left hand side in (4.4.3) can be bounded by

$$\frac{1}{n}\sum_m p(\mathcal{S}_m)\cdot\frac{X|\mathcal{S}_m}{X} \leq \frac{1}{n}\sum_m p(\mathcal{S}_m)\log\frac{1}{p(\mathcal{S}_m)} \leq \frac{2^\ell}{n},$$

where the first inequality follows from Proposition 2.22 (with $B = \perp$ and $\mathcal{W} = \mathcal{S}_m$) and the second from the fact that for $0 \leq \gamma \leq 1$, it holds that $\gamma\log(1/\gamma) \leq \frac{\log e}{e} < 1$.

This proves that $\mathbf{I}_p(M : S|Y_{\leq J}J) \leq \mathbf{I}_p(M : X_J|Y_{<J}J) \leq \frac{2^\ell}{n}$. The bound on $\mathbf{I}_p(M : S|X_{\leq J}J)$ follows analogously.

Proof of Lemma 4.15

We shall first prove:

Claim 4.17. $\mathbf{I}(C : U_{S'}|VS') \leq \sum_{i=1}^{t}\mathbb{E}_{cu}\left[p'(i \in S'|c)\cdot\frac{U_i|cu_{<i}}{U_i|u_{<i}}\right]$.

Call c *bad* if $p'(i \in S'|c) \geq 2e/k + \sqrt{\mathbf{I}(C:S')}$ for some $i \in [t]$ and let $\mathcal{W}$ denote the event that C is bad. Note that when the complement event $\overline{\mathcal{W}}$ occurs, then $p'(i \in S'|c) \leq 2e/k + \sqrt{\mathbf{I}(C:S')}$ for all i. We next show:

Claim 4.18. $p'(\mathcal{W}) \leq \sqrt{\mathbf{I}(C:S')}$.

We can now prove Lemma 4.15, using Claims 4.17 and 4.18:

$$\mathbf{I}\left(C : U_{S'}|VS'\right) \leq p'(\mathcal{W}) \sum_{i=1}^{t} \mathbb{E}_{cu|\mathcal{W}} \left[\frac{U_i|cu_{<i}}{U_i|u_{<i}} \right] + \left(\frac{2e}{k} + \sqrt{\mathbf{I}(C:S')} \right) \cdot \sum_{i=1}^{t} \mathbb{E}_{cu} \left[\frac{U_i|cu_{<i}}{U_i|u_{<i}} \right]. \tag{4.4.5}$$

When c is bad, we have that $p'(U_i|cu_{<i}\mathcal{W}) = p'(U_i|cu_{<i})$ and so, $\dfrac{U_i|cu_{<i}}{U_i|u_{<i}} = \dfrac{U_i|cu_{<i}\mathcal{W}}{U_i|u_{<i}}$. Plugging this into the right hand side in (4.4.5) and using the chain rule gives:

$$\begin{aligned}
&\mathbf{I}\left(C : U_{S'}|VS'\right) \\
&\leq p'(\mathcal{W}) \sum_{i=1}^{t} \mathbb{E}_{cu|\mathcal{W}} \left[\frac{U_i|cu_{<i}\mathcal{W}}{U_i|u_{<i}} \right] + \left(\frac{2e}{k} + \sqrt{\mathbf{I}(C:S')} \right) \sum_{i=1}^{t} \mathbb{E}_{cu} \left[\frac{U_i|cu_{<i}}{U_i|u_{<i}} \right] \\
&= p'(\mathcal{W}) \mathbb{E}_{c|\mathcal{W}} \left[\frac{U|c\mathcal{W}}{U} \right] + \left(\frac{2e}{k} + \sqrt{\mathbf{I}(C:S')} \right) \mathbb{E}_{c} \left[\frac{U|c}{U} \right].
\end{aligned}$$

Using Proposition 2.22 and that $\mathbb{E}_c \left[\dfrac{U|c}{U} \right] = \mathbf{I}(U:C)$, we can bound the above as

$$\begin{aligned}
\mathbf{I}\left(C : U_{S'}|VS'\right) &\leq p'(\mathcal{W}) \left(\log \frac{1}{p'(\mathcal{W})} + \mathbf{I}(U:C|\mathcal{W}) \right) + \left(\frac{2e}{k} + \sqrt{\mathbf{I}(C:S')} \right) \cdot \mathbf{I}(U:C) \\
&\leq \eta\left(p'(\mathcal{W})\right) + p'(\mathcal{W}) \cdot \mathbf{I}(U:C|\mathcal{W}) \\
&\quad + \sqrt{\mathbf{I}(C:S')} \cdot \mathbf{I}(U:C) + \frac{2e\mathbf{I}(U:C)}{k}.
\end{aligned} \tag{4.4.6}$$

The second inequality above follows since $\alpha \log(1/\alpha) \leq \eta(\alpha)$ for $0 \leq \alpha \leq 1$.

Using Claim 4.18, $p'(\mathcal{W}) \leq \sqrt{\mathbf{I}(C:S')}$. Since η (see (4.3.1)) is a non-decreasing function, we can then bound $\eta(p'(\mathcal{W})) \leq \eta\left(\sqrt{\mathbf{I}(C:S')}\right)$. By Proposition 2.6, $\mathbf{I}(U:C|\mathcal{W}), \mathbf{I}(U:C) \leq \ell$, so plugging it into (4.4.6), we get that

$$\mathbf{I}\left(C:U_{S'}|VS'\right) \leq \eta\left(\sqrt{\mathbf{I}(C:S')}\right) + 2\ell\sqrt{\mathbf{I}(C:S')} + \frac{2e\ell}{k}.$$

This finishes the proof of Lemma 4.15. It only remains to prove the claims.

Proof of Claim 4.17. For $i \in [t]$, set $U_i^* = U_i$ if $i \in S'$ and set $U_i^* = \perp$ otherwise. Since we have that $\mathbf{I}(C:U_{S'}|VS') = \mathbf{I}(C:U_1^*U_2^*\ldots U_t^*|VS')$, by the chain rule, we get

$$\begin{aligned}\mathbf{I}\left(C:U_{S'}|VS'\right) &= \mathbb{E}_{cvs'}\left[\frac{U_1^*\ldots U_t^*|cvs'}{U^*|vs'}\right] \\ &= \mathbb{E}_{cuvs'}\left[\sum_{i=1}^{t}\frac{U_i^*|cu_{<i}^*vs'}{U_i^*|u_{<i}^*vs'}\right] = \mathbb{E}_{cuvs'}\left[\sum_{i\in s'}\frac{U_i|cu_{<i}^*vs'}{U_i|u_{<i}^*vs'}\right],\end{aligned}$$

where the last equality holds because when $i \notin S'$, $U_i^* = \perp$ (and so the divergence is 0) and when $i \in S'$, $U_i^* = U_i$.

By assumption, U is independent of VS', $U - C - VS'$ and $p'(u_i|u_{<i}^*) = p'(u_i) = p'(u_i|u_{<i})$, so the right hand side above can be written as

$$\begin{aligned}\mathbb{E}_{cuvs'}\left[\sum_{i\in s'}\frac{U_i|cu_{<i}^*vs'}{U_i|u_{<i}^*vs'}\right] &= \mathbb{E}_{cvs'}\left[\mathbb{E}_{u|c}\left[\sum_{i\in s'}\frac{U_i|cu_{<i}^*}{U_i|u_{<i}^*}\right]\right] \\ &= \mathbb{E}_{cvs'}\left[\mathbb{E}_{u|c}\left[\sum_{i\in s'}\frac{U_i|cu_{<i}^*}{U_i|u_{<i}}\right]\right]. \qquad (4.4.7)\end{aligned}$$

Let $\mathbf{1}[i \in S']$ denote the indicator variable for the event that $i \in S'$. Using linearity of expectation

and Proposition 2.23, we have that

$$(4.4.7) = \mathbb{E}_{cvs'}\left[\sum_{i\in s'}\mathbb{E}_{u|c}\left[\frac{U_i|cu^*_{<i}}{U_i|u_{<i}}\right]\right]$$
$$\leq \mathbb{E}_{cvs'}\left[\sum_{i\in s'}\mathbb{E}_{u|c}\left[\frac{U_i|cu_{<i}}{U_i|u_{<i}}\right]\right] = \mathbb{E}_{cuvs'}\left[\sum_i \mathbf{1}[i\in s']\frac{U_i|cu_{<i}}{U_i|u_{<i}}\right],$$

where the first inequality above follows from Proposition 2.23 and the definition of U_i^* (which depends only on U_i and S').

Finally, using $U - C - VS'$ once more, we get that the right hand side above equals

$$\mathbb{E}_{cu}\left[\sum_i \mathbb{E}_{s'|c}\left[\mathbf{1}[i\in s']\right]\frac{U_i|cu_{<i}}{U_i|u_{<i}}\right] = \sum_{i=1}^{t}\mathbb{E}_{cu}\left[p'(i\in S'|c)\cdot\frac{U_i|cu_{<i}}{U_i|u_{<i}}\right].$$

□

Proof of Claim 4.18. Define $S'_i = 1$ if $i \in S'$ and 0 otherwise. We are going to show that whenever c is bad, we have that $\frac{S'|c}{S'} \geq \sqrt{\mathbf{I}(C:S')}$. Since $\mathbf{I}(C:S') = \mathbb{E}_c\left[\frac{S'|c}{S'}\right]$, Markov's inequality will imply that the probability of c being bad is at most $\sqrt{\mathbf{I}(C:S')}$ and hence $p(\mathcal{W}) \leq \sqrt{\mathbf{I}(C:S')}$.

When c is bad, there is an i^* such that $p'(i^*\in S'|c) \geq 2e/k + \sqrt{\mathbf{I}(C:S')}$. By chain rule and Proposition 2.18, $\frac{S'|c}{S'} \geq \frac{S'_{i^*}|c}{S'_{i^*}}$. Since $p'(i\in S') \leq 1/k$ for any i, by Proposition 2.25,

$$\frac{S'|c}{S'} \geq \frac{S'_{i^*}|c}{S'_{i^*}} \geq p'(i^*\in S'|c)\log\left(\frac{k\cdot p'(i^*\in S'|c)}{e}\right)$$
$$\geq \left(\frac{2e}{k}+\sqrt{\mathbf{I}(C:S')}\right)\log\left(\frac{k}{e}\left(\frac{2e}{k}+\sqrt{\mathbf{I}(C:S')}\right)\right) \geq \sqrt{\mathbf{I}(C:S')}.$$

This finishes the proof of the claim. □

Proof of Lemma 4.16

We assume $\mathbf{I}(C:A), \mathbf{I}(C:B) \leq 1$, since otherwise the lemma is trivially true. For brevity, set

$$\alpha^3 = \mathbf{I}(C:A) = \mathbb{E}_c\left[\frac{A|c}{A}\right] \text{ and } \beta^3 = \mathbf{I}(C:B) = \mathbb{E}_c\left[\frac{B|c}{B}\right].$$

Call c *bad* if $\frac{A|c}{A} \geq \alpha^2$ or $\frac{B|c}{B} \geq \beta^2$, and good otherwise. By Markov's inequality, the probability that C is bad is at most $\alpha+\beta$. To prove Lemma 4.16, we need the following claim proved in [GKR16]. For completeness, we include the short proof after finishing the proof of Lemma 4.16.

Claim 4.19. *Given independent random variables $A^*, B^* \in [t]$ in a probability space p', if A^* is γ_1-close to uniform, and B^* is γ_2-close to uniform, then*

$$p'(A^* = B^*) \geq \frac{1-\gamma_1-\gamma_2}{t}.$$

When c is good, Pinsker's inequality (Proposition 2.21) implies that conditioned on c, A is α-close to uniform and B is β-close to uniform. Then, since $A - C - B$, using Claim 4.19 (with $A^* = A$ and $B^* = B$ in the probability space p' conditioned on c) implies that $p'(A = B|c) \geq \frac{(1-\alpha-\beta)}{t}$. Since $p'(A=B) = \frac{1}{t}$, we have that for a good c,

$$p'(c|A=B) = \frac{p'(c)\cdot p'(A=B|c)}{p'(A=B)} \geq (1-\alpha-\beta)\cdot p'(c). \tag{4.4.8}$$

For any event $\mathcal{Q}$, (4.4.8) implies that

$$\begin{aligned} p'(C\in\mathcal{Q}) - p'(C\in\mathcal{Q}|A=B) &\leq \sum_{c\in\mathcal{Q}, c \text{ bad}} p'(c) + \sum_{c\in\mathcal{Q}, c \text{ good}} (p'(c) - p'(c|A=B)) \\ &\leq p'(C \text{ is bad}) + \sum_c p'(c)(\alpha+\beta) \\ &\leq \alpha+\beta+\sum_c p(c)(\alpha+\beta) \leq 2\alpha+2\beta, \end{aligned}$$

and since $|p'(C) - p'(C|A=B)| = \max_{\mathcal{Q}}(p'(C\in\mathcal{Q}) - p'(C\in\mathcal{Q}|A=B))$ we get the required bound bound on statistical distance.

Proof of Claim 4.19. For each i, let $p'(A^* = i) = \frac{1}{t} + \alpha_i$ and $p'(B^* = i) = \frac{1}{t} + \beta_i$. Then, $\sum_i \alpha_i = \sum_i \beta_i = 0$, and $\alpha_i, \beta_i \geq -\frac{1}{t}$. Using these facts,

$$\begin{aligned} p'(A^* = B^*) &= \sum_i \left(\frac{1}{t} + \alpha_i\right)\left(\frac{1}{t} + \beta_i\right) \\ &= \frac{1}{t} + \frac{\sum_i \alpha_i}{t} + \frac{\sum_i \beta_i}{t} + \sum_i \alpha_i\beta_i = \frac{1}{t} + \sum_i \alpha_i\beta_i. \end{aligned}$$

To lower bound the above, we will only consider the negative terms in the summation:

$$p'(A^* = B^*) \geq \frac{1}{t} + \sum_{i:\alpha_i>0,\beta_i<0} \alpha_i\beta_i + \sum_{i:\alpha_i<0,\beta_i>0} \alpha_i\beta_i \geq \frac{1}{t} - \frac{1}{t}\sum_{i:\alpha_i>0} \alpha_i - \frac{1}{t}\sum_{i:\beta_i>0} \beta_i.$$

From Proposition 2.1, it follows that $\sum_{i:\alpha_i>0} \alpha_i$ is the statistical distance between A^* and the uniform distribution on $[t]$ and likewise for B^*. So, we get,

$$p'(A^* = B^*) \geq \frac{1 - \gamma_1 - \gamma_2}{t}.$$ □

4.4.3 Fooling Distributions and Relative Discrepancy

In this section, we compare our techniques to that of Ganor, Kol and Raz [GKR16]. The key concept introduced in [GKR16] to prove lower bounds is the notion of the *relative discrepancy*. Let $f(x, y)$ be a boolean function and $q(X, Y)$ be a distribution such that $q(f = 0) = q(f = 1) = 1/2$. Then, f has (ϵ, δ) *relative discrepancy* under $q(X, Y)$ if there exists a distribution $u(X, Y)$ such that for every rectangle $S \times T$ in the input space, such that $u(X \in S, Y \in T) \geq \delta$, the following holds

$$\begin{aligned} &q(X \in S, Y \in T | f = 0) \geq (1 - \varepsilon)u(X \in S, Y \in T) \text{ and} \\ &q(X \in S, Y \in T | f = 1) \geq (1 - \varepsilon)u(X \in S, Y \in T). \end{aligned} \tag{4.4.9}$$

Ganor, Kol and Raz [GKR16] proved that if f has $(\epsilon = 1/3, \delta)$ relative discrepancy under $q(X, Y)$, then f has communication complexity $\Omega(\log 1/\delta)$. They then prove a lower bound on the relative discrepancy of the bursting noise function to give a communication lower bound. Note that the relative discrepancy technique individually argues about each rectangle that has large measure under the distribution $u(X, Y)$. Ganor, Kol and Raz [GKR16] prove a lemma (Lemma 12 in

[GKR16]) that is very similar to our Lemma 4.14, however they argue about each rectangle with a large enough measure under $u(X,Y)$.

In contrast, the fooling distribution technique allows us to argue about the distribution *on average* as opposed to arguing about each rectangle individually and this is where our proof becomes more simpler and more intuitive. Lemma 4.14 is an average-case version of the lemma proved in [GKR16]. Together with our generalization of Shearer's Lemma (Lemma 4.15) and Lemma 4.16 we can argue about the messages on average. Note that our fooling distribution $p(X,F,Y,G)$ is analogous to the distribution $u(X,Y)$ in the definition of relative discrepancy.

Next we show some connections between relative discrepancy and fooling distributions. In fact, we show that low relative discrepancy implies the existence of a fooling distribution. The converse does not appear to be true.

Claim 4.20 (Relative Discrepancy implies Fooling Distribution)**.** *If f has relative discrepancy $(\epsilon, 2^{-2\ell})$ then there exists a distribution $u(X,Y)$ such that if $M \in \{0,1\}^\ell$ denotes the messages of a deterministic protocol, then $q(M|f=0) \overset{\gamma}{\approx} u(M) \overset{\gamma}{\approx} q(M|f=1)$ where $\gamma = 2^{-\ell} + \varepsilon$.*

Proof. Let $u(X,Y)$ be the distribution that satisfies (4.4.9) with $\delta = 2^{-2\ell}$. We will show that $u(M) \approx q(M|f=0)$. The proof for $u(M) \approx q(M|f=1)$ is similar. Define $\mathcal{B} = \{m | u(m) < 2^{-2\ell}\}$ and note that $u(M \in \mathcal{B}) < 2^\ell 2^{-2\ell} = 2^{-\ell}$. Also observe that when $m \notin B$, then by (4.4.9) and Proposition 4.6, $u(m) - q(m|f=0) \leq \varepsilon u(m)$. Now for any event $\mathcal{Q}$, we have

$$\begin{aligned} u(M \in \mathcal{Q}) - q(M \in \mathcal{Q}|f=0) &\leq \sum_{m \in \mathcal{Q} \cap \mathcal{B}} u(m) + \sum_{m \in \mathcal{Q} \cap \bar{\mathcal{B}}} (u(m) - q(m|f=0)) \\ &\leq u(\mathcal{B}) + \varepsilon u(\bar{\mathcal{B}}) < 2^{-\ell} + \varepsilon. \end{aligned}$$

Hence $|u(M) - q(M|f=0)| = \max_{\mathcal{Q}}(u(M \in \mathcal{Q}) - q(M \in \mathcal{Q}|f=0)) < 2^{-\ell} + \varepsilon$. □

The above lemma gives another reason as to why our proof is simpler than the one given in [GKR16] — we are proving a weaker statement that still implies a communication lower bound. We mention that in [GKR16], a relaxed notion of relative discrepancy, called the *adaptive* relative discrepancy is also defined. Adaptive relative discrepancy works with a partition of the input space into

rectangles as opposed to working with each rectangle individually, and an upper bound on this measure also suffices to prove a communication lower bound. Using arguments similar to Claim 4.20 one can prove that the existence of a fooling distribution implies that the adaptive relative discrepancy is small. Hence, the lower bounds given by the fooling distribution technique are sandwiched between the lower bounds that can be obtained by the relative discrepancy and the adaptive relative discrepancy methods.

We remark that the results of Fontes *et al.* [FJK+15] do not rule out the possibility that one might be able to separate information and communication complexity in the non-distributional setting (see Section 4) by using either of the two techniques, fooling distributions or adaptive relative discrepancy. In fact, Fontes *et al.* [FJK+15] show that the adaptive relative discrepancy is equal to the worst-case distributional communication complexity of the function up to polynomial factors, but we do not know if the same holds for fooling distributions.

4.5 Information Upper Bound for k-ary Pointer Jumping

Let us recall the trivial protocol and the low information protocol mentioned in Section 4.3.1. In the trivial protocol Alice and Bob send each other $X(z_{\leq i}), Y(z_{\leq i})$ in each step i, until they have computed z. The parties can compute $F(z) + G(z) \bmod 2$ with two more bits of communication. This protocol reveals at least $\Omega(\log n)$ bits of information as both parties learn the value of J with high probability. But note that the z computed in the above manner is consistent and the input distribution $q(X, F, Y, G)$ ensures that $F(z) + G(z) \bmod 2$ is the same for *every* consistent z, so it is enough for the parties to find any consistent z to complete the goal.

For the low information protocol, the parties send each other the values $X(z_{\leq i}), Y(z_{\leq i})$ with probability $1 - \varepsilon$ and send a uniformly random value otherwise. The parties abort the protocol if they see r rounds where the messages they sent were not the same. The distribution on inputs ensures that they will sample a consistent z with probability $1 - O(\varepsilon)$. When the parties sample a consistent z, the messages sent are almost always sampled from a distribution that the receiving party knows, while if they sample a z that is not consistent, the protocol aborts shortly after the inconsistency.

The purpose of the noise is to prevent revealing a lot of information about J to both parties. When the z that the parties sample is consistent they only know that the value of J is among the locations where the values they have exchanged disagree. Since in this case there are εn disagreements in expectation, we intuitively expect that the entropy of J should still be large at the end which means that the information revealed about the value of J is small. In case the z sampled is not consistent (which happens with probability $O(\varepsilon)$), the abort condition prevents the parties from revealing too much information. By choosing the parameter ε carefully we can ensure that the total amount of information revealed is small.

In Figure 4.5.1, we describe the low information protocol which is parameterized by ϵ.

Lemma 4.21. *The protocol in Figure 4.5.1 outputs the correct answer with probability at least* $1 - 4\epsilon$ *on inputs drawn from the distribution* $q(J, X, F, Y, G)$.

Proof. Under the input distribution $q(J, X, F, Y, G)$, whether a sample z with $|z| \geq J + 1$ is consistent or not is determined solely by the $(J+1)^{\text{st}}$ step. So, the probability that the parties sample a z with $|z| \geq J + 1$ and z being inconsistent is at most 2ϵ. Given that this event does not happen, the probability that the parties abort in a particular step is at most $(2\epsilon)^{r-1}$, since to abort one of them must choose to send uniform values in at least $r - 1$ steps where their inputs are equal. By the union bound, the probability that the parties abort in any step is at most $n(2\epsilon)^{r-1}$. If neither of these bad events happens, the protocol computes the correct answer. Thus the probability of making an error is at most $2\epsilon + n(2\epsilon)^{r-1} \leq 4\epsilon$, by the choice of r. □

The following theorem proves that the information cost of the protocol is small. We directly argue about the average information revealed about the messages, as compared to the proof of [GKR16] who use the notion of the *tree divergence cost* to argue about the same.

Theorem 4.22. *If* M *denotes the messages of the protocol in Figure 4.5.1,*

$$\left.\begin{matrix} \mathbf{I}_q\left(M : XF | YG\right) \\ \mathbf{I}_q\left(M : YG | XF\right) \end{matrix}\right\} \leq 2\log(k/\epsilon) \cdot \left(1 + 16\epsilon \cdot (\log n + 3) \cdot 2^{\frac{2\log n}{k\log(1/2\epsilon)}}\right).$$

Input: Alice is given (x, f), Bob is given (y, g). Both know a parameter $\epsilon \in (0, 1)$.

Output: $f(z) + g(z) \bmod 2$ for some consistent z.

Set z to be the empty string;

Set $r = \lceil \frac{\log n}{\log(1/2\epsilon)} + 2 \rceil$;

for $i = 1, 2, \ldots, n$ **do**

Alice sets $a_i = \begin{cases} x(z_{<i}) & \text{with probability } 1 - \epsilon \\ \text{uniform element of } [k] & \text{with probability } \epsilon \end{cases}$,

Bob sets $b_i = \begin{cases} y(z_{<i}) & \text{with probability } 1 - \epsilon \\ \text{uniform element of } [k] & \text{with probability } \epsilon \end{cases}$;

Send $m_i = a_i, b_i$ to each other;

Set $w \in [k]$ so that $w = a_i + b_i \bmod k$, and append w to the string z;

if *$i \geq r$ and $a_{i'} \neq b_{i'}$ for all i' with $i - r + 1 \leq i' \leq i$;* `// if` a_i `and` b_i `disagree in each of the last` r `rounds`

then

Terminate the protocol;

end

end

Send $f(z), g(z)$;

Figure 4.5.1: Protocol π_ϵ

Setting $\epsilon = 1/\log n$ gives Theorem 4.1. To prove Theorem 4.22, we bound $\mathbf{I}_q(M : XF|YG)$. The second term is bounded in the same way.

Let $M_i \in [k] \times [k]$ be the messages exchanged in the i^{th} round for $i \in [n]$ and $M_{n+1} \in \{0,1\} \times \{0,1\}$ be the messages exchanged in the $(n+1)^{\text{th}}$ round. Then, by the chain rule, we can write

$$\mathbf{I}_q(M : XF|YG) = \mathbb{E}_{q(xfyg)}\left[\frac{q(M|xfyg)}{q(M|yg)}\right] = \sum_{i=1}^{|m|} \mathbb{E}_{q(mxfyg)}\left[\frac{q(M_i|m_{<i}xfyg)}{q(M_i|m_{<i}yg)}\right]. \quad (4.5.1)$$

To bound (4.5.1), we apply Proposition 2.24 which allows us to replace the distribution $q(M_i|m_{<i}yg)$ in the divergence expression above with a different distribution at the expense of increasing the divergence. Define distribution $u(MYG) = q(MYG|XF = YG)$. It will be simpler to analyze the divergence with respect to this distribution. Using Proposition 2.24, we then have

$$(4.5.1) \leq \sum_{i=1}^{|m|} \mathbb{E}_{q(mxfyg)}\left[\frac{q(M_i|m_{<i}xfyg)}{u(M_i|m_{<i}yg)}\right].$$

Next, we prove the following claim, which bounds the contribution to the divergence of each possible message:

Claim 4.23.

$$\frac{q(M_i|m_{<i}xfyg)}{u(M_i|m_{<i}yg)} \begin{cases} = 0 & \text{if } i \leq n \text{ and } x(z_{<i}) = y(z_{<i}), \\ \leq \log(k/\epsilon) & \text{if } i \leq n \text{ and } x(z_{<i}) \neq y(z_{<i}), \\ = 1 & \text{if } i = n+1. \end{cases}$$

Before proving Claim 4.23, we show how to use it to bound the information. First note that the above claim intuitively says that information is revealed by M_i only when $X(Z_{<i}) \neq Y(Z_{<i})$. Recall that with probability $1 - O(\varepsilon)$, the parties sample a consistent z and hence, information is revealed only in one step of the protocol (and also one bit at the end when exchanging values of $F(z)$ and $G(z)$). However, in case the parties do not sample a consistent z, there might be a lot of messages M_i such that $X(Z_{<i}) \neq Y(Z_{<i})$.

To bound the information, next we show that on average the number of such messages is small since the parties will abort if they see a lot of disagreements. Towards this end, define $Q_i = 1$ if $|M| \geq i$ and $X(Z_{<i}) \neq Y(Z_{<i})$, and 0 otherwise. Claim 4.23 implies that $\mathbf{I}_q(M : FX|YG) \leq 1 + \log(k/\epsilon) \cdot \mathbb{E}_q\left[\sum_{i=1}^n Q_i\right]$, so it only remains to bound $\mathbb{E}_q\left[\sum_{i=1}^n Q_i\right]$ which is the number of messages where information is revealed.

Claim 4.24.

$$\mathbb{E}_q\left[\sum_{i=1}^n Q_i\right] \leq 1 + \frac{2r\epsilon}{(1-1/k)^r} \leq 1 + 16\epsilon \cdot (\log n + 3) \cdot 2^{\frac{2\log n}{k\log(1/2\epsilon)}}.$$

Claims 4.23 and 4.24 complete the proof of Theorem 4.22. We prove them next.

Proof of Claim 4.23. For any $i \in [n+1]$, let us write $M_i = A_i, B_i$ where A_i denotes Alice's message and B_i denotes Bob's message. Note that, when $i = n+1$, A_i and B_i are bits, and otherwise they are numbers in $[k]$. By the chain rule, for every $i \in [n+1]$,

$$\frac{q(M_i|m_{<i}xfyg)}{u(M_i|m_{<i}yg)} = \frac{q(A_i|m_{<i}xfyg)}{u(A_i|m_{<i}yg)} + \mathbb{E}_{q(a_i|m_{<i}xfyg)}\left[\frac{q(B_i|m_{<i}xfyga_i)}{u(B_i|m_{<i}yga_i)}\right]. \tag{4.5.2}$$

By the definition of the protocol in Figure 4.5.1, for every $i \in [n+1]$, B_i is determined by $YGM_{<i}$. This is because for $i \in [n]$, B_i is either a uniform random value or $Y(z_{<i})$ for some $z_{<i}$ determined by $M_{<i}$ and when $i = n+1$, $B_i = G(z)$ for some z determined by $M_{\leq n}$. It follows that for $i \in [n+1]$, $q(B_i|m_{<i}xfyga_i) = q(B_i|m_{<i}yg)$. Similarly, by the definition of the distribution $u(MYG)$, we have $u(B_i|m_{<i}xfyga_i) = u(B_i|m_{<i}yg) = q(B_i|m_{<i}yg)$ for every $i \in [n+1]$. Hence, the second term in (4.5.2) is zero since the distributions are the same. So, for $i \in [n+1]$ we get that

$$\frac{q(M_i|m_{<i}xfyg)}{u(M_i|m_{<i}yg)} = \frac{q(A_i|m_{<i}xfyg)}{u(A_i|m_{<i}yg)}. \tag{4.5.3}$$

When $i = n+1$, using (4.5.3), a direct calculation shows

$$\frac{q(M_{n+1}|m_{\leq n}xfyg)}{u(M_{n+1}|m_{\leq n}yg)} = 1 \cdot \log(1/2) = 1.$$

Let us consider the case when $i \in [n]$ now. If $x(z_{<i}) = y(z_{<i})$, the definition of the protocol given in Figure 4.5.1, together with (4.5.3) ensures that $u(A_i|m_{<i}yg) = q(A_i|m_{<i}xfyg)$, proving the first bound. On the other hand if $x(z_{<i}) \neq y(z_{<i})$, then $q(A_i = a_i|m_{<i}xfyg) = u(A_i = a_i|m_{<i}yg)$, except when $a_i = x(z_{<i})$ or $a_i = y(z_{<i})$ (since otherwise the probability is ε/k in each case). Thus the divergence can be bounded by the contribution of these two values. We have that $q(A_i = x(z_{<i})|m_{<i}xfyg) = (1-\varepsilon) + (\varepsilon/k)$ where the first and second terms account respectively for the cases when Alice sends the correct value and when Alice sends a random value. On the other hand, we have $u(A_i = x(z_{<i})|m_{<i}yg) = q(A_i = x(z_{<i})|m_{<i}yg, X = y, F = g) = \varepsilon/k$. Similarly, we get that $q(A_i = y(z_{<i})|m_{<i}xfyg) = \varepsilon/k$ and $u(A_i = y(z_{<i})|m_{<i}yg) = 1 - \varepsilon + (\varepsilon/k)$.

Therefore, using (4.5.3) we can bound

$$\begin{aligned}\frac{q(M_i|m_{<i}xfyg)}{u(M_i|m_{<i}yg)} &= (1-\epsilon+(\epsilon/k))\log\frac{1-\epsilon+(\epsilon/k)}{\epsilon/k} + (\epsilon/k)\log\frac{\epsilon/k}{1-\epsilon+\epsilon/k} \\ &\leq \log\frac{1-\epsilon+(\epsilon/k)}{\epsilon/k} \leq \log(k/\epsilon),\end{aligned}$$

as required. □

Proof of Claim 4.24. Let T be such that M_T is the last message sent in the protocol. Let $\mathcal{W}$ be the event that $T > J$ and Z sampled by the protocol is not consistent. Since $q(\mathcal{W}) \leq 2\epsilon$ (if the parties do not send $X(Z_{\leq J})$ and $Y(Z_{\leq J})$ at the $(J+1)^{\text{st}}$ step), and $\mathbb{E}_q\left[\sum_{i=1}^n Q_i \middle| \neg\mathcal{W}\right] \leq 1$ (at most one message where information is revealed when Z is consistent), we have

$$\mathbb{E}_q\left[\sum_{i=1}^n Q_i\right] \leq 2\epsilon \cdot \mathbb{E}_q\left[\sum_{i=1}^n Q_i \middle| \mathcal{W}\right] + 1.$$

If $T > J$, $\sum_{i=1}^n Q_i = \sum_{i=J+1}^T Q_i \leq T - J$, so we get $\mathbb{E}_q\left[\sum_{i=1}^n Q_i \middle| \mathcal{W}\right] \leq \mathbb{E}_q[T - J|\mathcal{W}]$. Note that $T - J$ roughly behaves like a geometric random variable. To bound this expectation, let us bound the contribution when $T - J \leq r$ and when $T - J > r$ separately. The probability that the protocol

does not abort in r rounds conditioned on the event $\mathcal{W}$ is at most $(1-1/k)^r$, so we have

$$\begin{aligned}\mathbb{E}_q\left[T-J|\mathcal{W}\right] &\leq q(T-J\leq r|\mathcal{W})r + q(T-J>r|\mathcal{W})(r+\mathbb{E}_q\left[T-J|\mathcal{W}\right]) \\ &\leq (1-1/k)^r r + (1-(1-1/k)^r)(r+\mathbb{E}_q\left[T-J|\mathcal{W}\right]) \\ \implies \mathbb{E}_q\left[T-J|\mathcal{W}\right] &\leq \frac{r}{(1-1/k)^r},\end{aligned}$$

as required. The second inequality in the statement of the claim follows from the fact that $1/(1-1/k) \leq 2^{2/k}$, for $k \geq 2$, and by the choice of r. □

5 A Direct Sum Theorem for Streaming

In this chapter, we investigate direct-sum questions for streaming algorithms (equivalently, oblivious read-once branching programs). Recall that an input to a streaming algorithm is a sequence of n updates $x_1, \ldots, x_n$ arriving sequentially in time, and the streaming algorithm at the end must compute a function $f(x_1, x_2, \ldots, x_n)$. The complexity measure of interest is the amount of memory that is needed to carry out the computation. Here the memory used by the algorithm at time t is the number of bits the program stores in the memory after reading the inputs $x_1, \ldots, x_t$.

We are interested in how the complexity of a problem changes when the streaming algorithm must process k independent inputs that arrive in parallel. The algorithm now gets k input streams $x^1, x^2, \cdots, x^k$ where $x^i := x^i_1, \ldots, x^i_n$ and the inputs $x^1_t, \ldots, x^k_t$ arrive simultaneously in the t'th time-step. Obviously one can process each of the inputs independently, giving an algorithm that uses k times as much memory. The central question that we investigate in this chapter is: are there interesting functions f for which the best streaming algorithm that computes f on k independent inputs does *not* operate independently on each input? This question is dual to another interesting question: When can we effectively reduce the memory of a streaming algorithm without compromising its accuracy?

These questions also make a lot of sense in the context of the most common applications for streaming algorithms like internet traffic analysis or data from multiple satellites. They also make sense from a theoretical perspective: they help to identify exactly what makes some streaming tasks hard and others easy.

The extensive literature on streaming algorithms is mostly concerned with understanding the maximum number of bits of memory used by the streaming algorithm throughout its run. One can

imagine pathological cases when one can effectively process k inputs at the same cost as processing a single input using this measure of complexity. Suppose that there is a uniformly random block of n/k^3 consecutive updates that contains information in the input, and all other updates are set to 0. Then without loss of generality, the best streaming algorithm uses almost no memory for most of the time, and some memory to process the block of important inputs. When the program processes k parallel inputs, it is very likely that the k informative blocks will not overlap in time, and so the maximum memory usage remains unchanged. Thus, if we are only aiming for a streaming algorithm that succeeds with high probability over this distribution of inputs, one need not increase the memory at all!

However, we see that the *average memory* usage per unit time-step does increase by a factor of k in this last example. The average memory is defined to be the number of bits of memory used on an average time-step. Arguably from the streaming viewpoint, the average memory is what we care about when considering practical applications of streaming algorithms. Another appealing reason to consider average memory as a complexity measure is that some known streaming lower bounds actually yield lower bounds on the average memory. For example, the lower bound proofs for approximating the frequency moments [AMS99, BYJKS02, CKS03, Gro09] and for approximating the length of the longest increasing subsequence [EJ08] can be easily adapted to give matching lower bounds for average memory. In the rest of this work we focus on the average memory used by the streaming algorithm.

5.1 Related Work

The interest in the field of streaming algorithms was renewed by the seminal paper of Alon, Matias and Szegedy [AMS99] who gave algorithms for approximating lower frequency moments and also showed that lower bounds in the *multi-party number-in-hand* communication model implied memory lower bounds for streaming algorithms approximating the higher frequency moments. Since then, lower bounds in communication complexity (and more recently in information complexity) have found applications in proving memory lower bounds in the streaming model (see [AMS99, BYJKS02, CKS03,

Woo04, EJ08, GH09, Gro09, MWY13] for some of them).

As discussed in Chapter 3, questions analogous to the ones we study here have been studied in the setting of two-party communication complexity and information complexity [BBCR13, BR11, Bra12]. It was shown in [GKR16] that there are communication tasks that can be solved much more efficiently in parallel than by naively solving each one independently.

Combining these results about parallelizing communication with known methods for proving lower bounds on streaming algorithms gives several interesting worst-case memory lower bounds for computing natural functions on k parallel streams. To give an example, it is known that computing $(1+\varepsilon)$ approximation of the p^{th} frequency moment for $p \neq 1$ requires worst-case memory $\Omega(1/\varepsilon^2)$ [Woo04, Gro09]. Combining this with the results of [BR11] one can show that computing $(1+\varepsilon)$ approximation of the frequency moment on k streams in parallel requires $\Omega(k/\varepsilon^2)$ memory in the worst-case. We do not give the proof here, since it is relatively straightforward.

A related model is that of *dynamic distributed functional monitoring* introduced by Cormode, Muthukrishnan and Yi [CMY11] where there are multiple sites receiving data streams and communicating with a central coordinator who wants to maintain a function of all the input streams. Recent progress has been made in understanding the communication complexity of various tasks in this model [CMY11, WZ12, WZ14]. Variants of this model have been studied extensively in relation to databases and distributed computing (see [Cor05, CG05, SSK08, SSK10, CMZ06, CMYZ10, ABC09, MSDO05, KCR06, BO03] for some of the applications). Another closely related model is the multi-party *private message passing* model introduced in [DR98]. Any lower bound proved in the message passing model implies a lower bound in the streaming model. Many works have studied this model and its variants (see [GH09, GG10, PVZ12, BEO+13, CRR14, HRVZ15] for some of them). These works do not appear to have any connection to the questions we study here.

5.2 Results

Our results are proved in the setting of average-case complexity: we assume that there is a known distribution on inputs, and consider the performance of algorithms with respect to that distribution. Let

$\mathcal{A}$ be a randomized streaming algorithm which receives an input stream $X = X_1, \ldots, X_n$ sampled from a distribution $p(X_1, \ldots, X_n)$. Throughout this chapter we will only consider the case when $p(X_1, \cdots, X_n)$ is a product distribution except in Section 5.4.1, where we discuss the issues that arise when considering non-product input distributions.

Let $M_1, \ldots, M_n$ denote the contents of the memory of the algorithm at each of the time-steps. Let $|M_t|$ denote the number of bits used to store M_t. The average memory used by the algorithm is $(1/n)\sum_{t=1}^{n} |M_t|$. Let $M(f)$ denote the minimum average memory required to compute a function f with probability $2/3$ when the inputs are sampled according to $p(X)$.

Let $p^k(X)$ denote the product distribution on k independent streams, each identically distributed as $p(X)$, where the resulting streams arrive synchronously in parallel. Thus at time t the input is the t^{th} element of all the k streams. Write f^k to denote the function that computes f on each of the k streams. Then we prove

Theorem 5.1.

$$M(f^k) = \Omega\left(k\left(\frac{M(f)}{n} - 1\right)\right).$$

Theorem 5.1 is proved by a reduction that compresses streaming algorithms with regards to their information complexity. There are several reasonable measures of information complexity for streaming algorithms. Here we define two such information complexity measures. We use Shannon's notion of mutual information, which is defined in the preliminaries (Chapter 2).

The *transitional information content* captures the average amount of information that the algorithm learns about the next input conditioned on its current state.

Definition 5.2 (Transitional Information)**.** $\mathsf{IC}^{tr}(\mathcal{A}) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{I}\left(M_t : X_t | M_{t-1}\right)$.

The *cumulative information content* measures the average amount of information that the streaming algorithm remembers about the inputs seen so far.

Definition 5.3 (Cumulative Information)**.** $\mathsf{IC}^{cum}(\mathcal{A}) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{I}\left(M_t : X_1 \ldots X_t\right)$.

Note that both the transitional and the cumulative information content for an algorithm are bounded by the average memory used by the algorithm. We prove that algorithms with low transitional information can be efficiently simulated:

Theorem 5.4. *Every streaming algorithm with transitional information content I can be simulated with average memory $O(nI + n)$.*

The above theorem is tight as the following example shows. Let the input x be sampled from the uniform distribution on $\{0,1\}^n$ (*i.e.* each update x_i for $i \in [n]$ is a bit). Consider the streaming algorithm $\mathcal{A}$ that remembers all the updates seen so far and outputs $x_1, \ldots, x_n$ at the end. The average memory used by the algorithm is $\Omega(n)$ while the transitional information content of this algorithm is 1. In this case the compression algorithm given by the above theorem would simulate $\mathcal{A}$ with average memory $O(n)$ which is the best one could hope for.

Finally, we show that if algorithms with low cumulative information can be simulated, then one can obtain no savings when parallelizing streaming algorithms:

Theorem 5.5. *If every algorithm with cumulative information I can be simulated using average memory $O(I)$, then $M(f^k) = \Omega\left(k \cdot (M(f) - 1)\right)$.*

In Section 5.5, we discuss more about the possibility of compressing algorithms with low cumulative information content.

5.3 Preliminaries

5.3.1 Common Information and Error-free Sampling

Wyner [Wyn75] defined the quantity *common information* between X and M as

$$\mathbb{C}\left(X : M\right) = \inf_{X - W - M} \mathbf{I}\left(XM : W\right),$$

where the infimum is taken over all jointly distributed W such that, $X - W - M$ and W is supported over a finite set. Wyner showed that the above infimum is always achieved. By the data-processing inequality applied to the Markov chain $X - W - M$ it is easily seen that $\mathbb{C}\left(X : M\right) \geq \mathbf{I}\left(X : M\right)$.

It turns out that the gap between $\mathbb{C}\left(X : M\right)$ and $\mathbf{I}\left(X : M\right)$ can be very large. There are known examples of random variables X and M where $\mathbb{C}\left(X : M\right) = \omega(\mathbf{I}\left(X : M\right))$. We include one simple

example later in this section. Another example is described in the work of Harsha et al. [HJMR10], who also proved a related upper bound. They showed that there always exist C and S, where S is independent of X, $X - CS \to M$ and $\mathbf{H}(C) \approx \mathbf{I}(X : M)$. The random variable S in their work depends on the distribution of M. Braverman and Garg [BG14] showed a similar result that we quote and use in this work:

Lemma 5.6 ([BG14])**.** *Let $p(xm)$ be an arbitrary discrete probability distribution, with finite support. Let S be an infinite list of uniform samples from $\mathsf{supp}(M) \times [0,1]$, independent of XM. Then there exists a random variable C such that $X - CS \to M$ and $\mathbf{H}(C|S) \leq \mathbf{I}(X : M) + \log(\mathbf{I}(X : M) + 1) + O(1)$.*

Separation between common information and mutual information Below we give an explicit example of random variables X and M such that $\mathbf{C}(X : M) = \omega(\mathbf{I}(X : M))$. Let G be a bipartite graph on the vertex set $([n], [n])$ such that the edge density of G is $\approx \frac{1}{2}$ and there are no cliques with more than $3n \log n$ edges in G. As the following lemma shows a random bipartite graph where each edge is picked with probability $1/2$ satisfies these properties with high probability, so such graphs exist.

Lemma 5.7. *With probability $1 - o(1)$, a random bipartite graph on $([n], [n])$ with edge density $1/2$ has no clique $U \times V$ where $U, V \subseteq [n]$ satisfying $\min\{|U|, |V|\} \geq 2 \log n + 2$.*

Proof. Let us set $t := 2 \log n + 2$ for notational convenience. If there is a clique $U \times V$ with $\min\{|U|, |V|\} \geq t$ then there also exists a clique of size $t \times t$. Consequently, to prove the lemma it suffices to upper bound the probability that a $t \times t$ clique exists in the graph. This probability is at most

$$\binom{n}{t}\binom{n}{t}2^{-t^2} \leq n^{2t}2^{-t^2} = 2^{2t\log n - t^2} = 2^{t(2\log n - t)} \leq 2^{-2t} = o(1).$$

□

A corollary of the above lemma is that the maximal clique in a random bipartite graph with edge density $1/2$ has at most $n \cdot 3 \log n$ edges with high probability.

Now we can describe the random variables X and M which will be the end points of a uniformly random edge E in the graph G. It is easily seen that the mutual information $\mathbf{I}(X : M) \approx 1$ since $\mathbf{H}(X) = \log n$ while for any $M = m$, $\mathbf{H}(X|M = m) \approx \log n - 1$. On the other hand, if $X - W - M$, then for any value w attained by W, $\mathsf{supp}(X|W = w)$ and $\mathsf{supp}(M|W = w)$ has to form a clique in the graph G. Since the maximal clique in G has at most $3n \log n$ edges, for any $W = w$, it holds that

$$|\mathsf{supp}(X|W = w)| \cdot |\mathsf{supp}(M|W = w)| \leq 3n \log n.$$

It follows that for any such W we can write

$$\mathbf{H}(XM|W) \leq \log(|\mathsf{supp}(X|W = w)| \cdot |\mathsf{supp}(M|W = w)|) = \log n + O(\log \log n).$$

Hence we have that the mutual information between XM and W is,

$$\begin{aligned}\mathbf{I}(XM : W) &= \mathbf{H}(XM) - \mathbf{H}(XM|W) \\ &\approx (2 \log n - 1) - (\log n + O(\log \log n)) = \log n - O(\log \log n),\end{aligned}$$

for any W satisfying $X - W - M$. It follows that $\mathbb{C}(X : M) = \Omega(\log n)$ while $\mathbf{I}(X : M) \approx 1$.

5.3.2 Streaming Algorithms

Without loss of generality, we associate the values stored by the algorithm with a non-negative integer. Assuming that the inputs to the algorithm come from the domain $\mathcal{X}$, a streaming algorithm defines a function $\mathcal{A} : [n] \times \mathbb{N} \times \mathcal{X} \to \mathbb{N}$. At time $t - 1$, let the memory state of the algorithm be m_{t-1} (we define $m_0 := 1$). On seeing the input x_t at time t, the algorithm computes the t^{th} memory state $m_t := \mathcal{A}(t, m_{t-1}, x_t)$. The output of the algorithm is m_n. Randomized streaming algorithms toss *independent* random coins r_t at each time-step t and sample the memory state at time t as follows: $m_t := \mathcal{A}(t, m_{t-1}, r_t, x_t)$.

The following is obvious from the definition:

Proposition 5.8 (Markov Chain Property)**.** *If $M_1, \ldots, M_n$ denote the memory of a (possibly randomized) streaming algorithm, then for each $t \in [n]$, $X_{\leq n} M_{<t} - X_t M_{t-1} - M_t$.*

The last proposition also implies the following.

Proposition 5.9. *For a randomized streaming algorithm, the following holds,*

$$\mathbf{I}\left(M_{\leq n} : X_{\leq n}\right) = \mathbf{I}\left(M_1 : X_1\right) + \mathbf{I}\left(M_2 : X_2 | M_1\right) + \cdots + \mathbf{I}\left(M_n : X_n | M_{n-1}\right).$$

Proof. Applying the chain-rule, we get

$$\mathbf{I}\left(M_{\leq n} : X_{\leq n}\right) = \sum_{t=1}^{n} \mathbf{I}\left(M_t : X_{\leq n} | M_{<t}\right) \leq \sum_{t=1}^{n} \mathbf{I}\left(M_t : X_t X_{\leq n} M_{<t-1} | M_{t-1}\right).$$

The second inequality follows since $\mathbf{I}\left(M_t : X_t X_{\leq n} M_{<t-1} | M_{t-1}\right) = \mathbf{I}\left(M_t : M_{<t-1} | M_{t-1}\right) + \mathbf{I}\left(M_t : X_{\leq n} | M_{<t}\right) + \mathbf{I}\left(M_t : X_t | M_{<t} X_{\leq n}\right)$ and mutual information is a non-negative quantity.

Applying the chain rule one more time, we have

$$\begin{aligned}\mathbf{I}\left(M_{\leq n} : X_{\leq n}\right) &\leq \sum_{t=1}^{n} \mathbf{I}\left(M_t : X_t X_{\leq n} M_{<t-1} | M_{t-1}\right) \\ &= \sum_{t=1}^{n} \mathbf{I}\left(M_t : X_t | M_{t-1}\right) + \sum_{t=1}^{n} \mathbf{I}\left(M_t : X_{\leq n} M_{<t-1} | X_t M_{t-1}\right).\end{aligned}$$

Proposition 5.8 implies that $X_{\leq n} M_{<t} - X_t M_{t-1} - M_t$ for every $t \in [n]$ and hence the second term on the right hand side is zero.

□

The following proposition states that both the transitional and cumulative information content are upper bounded by the average memory.

Proposition 5.10. *For a randomized streaming algorithm $\mathcal{A}$ with average memory M,*

$$\max\{\mathsf{IC}^{tr}(\mathcal{A}), \mathsf{IC}^{cum}(\mathcal{A})\} \leq M.$$

Definition 5.11 (Simulation)**.** *We say that a streaming algorithm $\mathcal{A}_1$ simulates another algorithm $\mathcal{A}_2$ if for every input $x_1, \ldots, x_n$, the distribution on outputs is exactly the same in both algorithms.*

In general it even makes sense to allow errors during simulation. Our simulations have no error, so we define simulation using the strong definition given above.

5.4 Compression and Direct Sums for Streaming Computation

The following is a natural strategy to prove our direct-sum theorem: given an algorithm that computes f^k correctly with probability $2/3$ on all the streams and uses average memory M, first show that there is some stream "with respect to" which the information content is M/k. Then derive a randomized streaming algorithm that computes f and has information content at most M/k as follows: embed the input stream at the location j about which the memory has small information and simulate the behavior of the algorithm on this stream by generating the other streams randomly, or to say alternately, sample from the distribution $p(M_n|X^{(j)} = x)$. The resulting algorithm would have information content at most M/k but would still use M bits of average memory. The last step would then be to give a way to simulate a streaming algorithm that has information content I with a streaming algorithm that uses average memory approximately I.

For product distributions, we can show that if there exists an algorithm for computing k copies of f with memory M, then there is a randomized algorithm for computing a single copy of f with transitional and cumulative information content at most M/k. To prove our direct-sum result, we are able to show that algorithms with transitional information content I can be simulated with $O(nI + n)$ average memory which as discussed before is best possible. To give an optimal direct-sum result, one could still hope that streaming algorithms with cumulative information content I can be simulated with $O(I)$ average memory. We discuss more about this possibility in Section 5.5.

5.4.1 Non-product Distributions and Correlated Randomness

Before we begin the proof of our compression and direct-sum results, we briefly discuss the difficulty that arises in dealing with non-product distributions. For proving a direct-sum result for non-product distributions using the above strategy, the natural way of using an algorithm that computes k copies of f to compute a single copy of f, is to embed our input stream at position j and generate other streams as randomness so that we can run the algorithm for k copies. The algorithm we get for computing f in this way uses randomness that is correlated across various time-steps if the input stream distribution

is non-product.

Transitional information content is not a useful information measure for compressing such algorithms as the following example shows. We give an example of a function which require $\Omega(1)$ average memory, but can be computed by an algorithm that uses correlated randomness and has transitional information content $1/n$. Let $f(x) = \sum_{t=1}^{n} x_t \mod 2$. Consider the following algorithm that takes as input a random input stream x (each update x_t is a bit) and computes $f(x)$. The algorithm at time t uses randomness r_t where $r_1, \ldots, r_t$ are correlated so that they satisfy $\sum_{t=1}^{n} r_t = 0 \mod 2$. At time t, the algorithm stores in its memory $\sum_{i=1}^{t}(x_t + r_t) \mod 2$ and at time $t = n$ outputs the last value stored in memory. Since $\sum_{t=1}^{n} r_t = 0 \mod 2$, the algorithm outputs $f(x)$. This algorithm has transitional information content $1/n$, but one can not hope to compute the parity of an n bit string without using $\Omega(1)$ bits of average memory.

5.4.2 Compressing Streaming Algorithms

In this section we show how algorithms with small transitional information content can be simulated with small average memory.

Theorem 5.12 (Restated). *Let $\mathcal{A}$ be a randomized streaming algorithm with $\mathsf{IC}^{tr}(\mathcal{A}) = I$. Then there exists a randomized streaming algorithm $\mathcal{A}_{tr}$ with average memory $O(nI + n)$ that simulates $\mathcal{A}$.*

Let $m_1, \ldots, m_n$ denote the memory states of the algorithm $\mathcal{A}$. Recall that Lemma 5.8 implies that for each $t \in [n]$, $X_{\leq n} M_{<t} - X_t M_{t-1} - M_t$. Hence, to prove Theorem 5.12, it suffices to sample from $p(M_t | x_t, m_{t-1})$ if m_{t-1} has been sampled correctly. The compression algorithm will toss random coins to sample an infinite list s_t of samples from $\mathsf{supp}(M_t) \times [0,1]$ and then sample C_t (whose existence is guaranteed by Lemma 5.6) satisfying

$$X_t - C_t S_t | M_{t-1} \to M_t | M_{t-1}, \tag{5.4.1}$$

$$\mathbf{H}(C_t | S_t M_{t-1}) = \mathbf{I}(M_t : X_t | M_{t-1}) + \log(\mathbf{I}(M_t : X_t | M_{t-1}) + 1) + O(1). \tag{5.4.2}$$

The value m_t determined by the sample c_t is distributed according to the distribution $p(M_t | x_t, m_{t-1})$.

The algorithm will store the Huffman encoding (Proposition 2.5) of C_t conditioned on S_t and M_{t-1}. This encoding determines C_t given S_t and M_{t-1}, both of which are known to the algorithm at this time.

Input : Stream x sampled from $p(X)$

Randomness: $s_1, \ldots, s_n$ where s_i is an infinite sequence of uniform samples from $\mathsf{supp}(M_i) \times [0,1]$.

// At time t: the content of the memory are some encodings of $c_{<t}$, where c_i determines m_i given s_i and m_{i-1}.

1. Let m_{t-1} be determined by c_{t-1} and s_{t-1}. On input x_t, sample c_t from the Markov chain in (5.4.1);
2. Append the Huffman encoding of c_t conditioned on s_t and m_{t-1} to the previous memory contents;

Randomized Streaming Algorithm $\mathcal{A}_{tr}$

Note that the algorithm needs to store the encodings of all the previous $c_{\leq t}$ at time t since in order to determine m_t uniquely, the value of m_{t-1} needs to be known which depends on the previous memory contents.

The following proposition is straightforward from (5.4.1).

Proposition 5.13. *The algorithm $\mathcal{A}_{tr}$ simulates $\mathcal{A}$.*

Next we finish the proof of Theorem 5.12 by bounding the total memory used by $\mathcal{A}_{tr}$.

Lemma 5.14. *The average memory used by $\mathcal{A}_{tr}$ is $O(nI + n)$.*

Proof. At time t, the expected number of bits appended to the memory (where the expectation is over the choice of $x_{\leq t}$ and $s_{\leq t}$) is bounded by $\mathbf{H}(C_t|S_t M_{t-1})$. From (5.4.2), it follows that this is at most $\mathbf{2I}\,(M_t : X_t|M_{t-1}) + O(1)$. Hence, the number of bits stored in the memory at a time $t \in [n]$ is at

most

$$\sum_{i=1}^{t}(2\mathbf{I}\left(M_i : X_i | M_{i-1}\right) + O(1)) \leq \sum_{i=1}^{n}(2\mathbf{I}\left(M_i : X_i | M_{i-1}\right) + O(1)) = 2nI + O(n).$$

Since this is true for every time-step t, the average memory is also upper bounded by $2nI + O(n)$. □

5.4.3 Direct Sum for Product Distributions

Recall that we want to prove the following theorem.

Theorem 5.15 (Direct Sum - Restated)**.** *If $p(X)$ is product input distribution, then*

$$M(f^k) = \Omega\left(k\left(\frac{M(f)}{n} - 1\right)\right).$$

To prove the above we first show that if there is a deterministic algorithm for computing k copies of f with average memory M and error probability $1/3$, then there is a randomized algorithm which computes a single copy of f with error at most $1/3$ and has transitional information content at most M/k. Then, we apply Theorem 5.12 to compress this algorithm and get a contradiction if M is smaller than the right hand side in Theorem 5.15.

Computing f with Small Information

Let $\mathcal{A}$ be a *deterministic* streaming algorithm that uses average memory M and computes f^k on inputs sampled from $p^k(X)$ with error at most $1/3$. Let $m_1, \ldots, m_n$ denote the memory states of the algorithm $\mathcal{A}$. Consider the following *randomized* algorithm $\mathcal{A}_{ran}$ that computes f with error at most $1/3$ on inputs sampled from p. The algorithm chooses a random $j \in [k]$, embeds the input stream at position j and at time t, samples and stores the memory state m_t from the distribution $p(M_t | X_t^{(j)} = x_t, m_{t-1})$.

Note that for any fixed value of j, the algorithm $\mathcal{A}_{ran}$ uses independent randomness $x_t^{(-j)}$ in each step as the input distribution $p(X)$ is product. We show that on average over the choice of j, the transitional and cumulative information content of the above algorithm is at most M/k.

Lemma 5.16. $\mathbb{E}_j[\mathsf{IC}^{tr}(\mathcal{A}_{ran} | J = j)] \leq M/k$ *and* $\mathbb{E}_j[\mathsf{IC}^{cum}(\mathcal{A}_{ran} | J = j)] \leq M/k$.

Input : Stream x sampled from $p(X)$

Randomness: j uniformly drawn from $[k]$, streams $x^{(-j)}$

Output : $f(x)$ with error at most $1/3$

1. Set Stream $x^{(j)}$ to be x;
2. At time t, use randomness $x_t^{(-j)}$ to sample m_t from $p(M_t|X_t^{(j)} = x_t, m_{t-1})$;
3. Output the answer of the algorithm on stream j;

Randomized Streaming Algorithm $\mathcal{A}_{ran}$

Proof. Conditioned on any event $J = j$, the transitional information content of $\mathcal{A}_{ran}$ is given by

$$\begin{aligned}
\mathsf{IC}^{tr}(\mathcal{A}_{ran}|J=j) &= \frac{1}{n}\sum_{t=1}^{n} \mathbf{I}\left(M_t : X_t | M_{t-1}, J=j\right) \\
&= \frac{1}{n}\sum_{t=1}^{n} \mathbf{I}\left(M_t : X_t^{(j)} | M_{t-1}, J=j\right) && \text{(with probability 1, } X^{(j)} = X\text{)} \\
&= \frac{1}{n}\sum_{t=1}^{n} \mathbf{I}\left(M_t : X_t^{(j)} | M_{t-1}\right) && (M_t \text{ is independent of the event } J=j).
\end{aligned}$$

Since the input stream comes from a product distribution, $X_t^{(1)}, \ldots, X_t^{(k)}$ are all independent conditioned on M_{t-1}. By Lemma 2.12, the term $\mathbf{I}\left(M_t : X_t^{(j)}|M_{t-1}\right)$ in the above sum is bounded by $\mathbf{I}\left(M_t : X_t^{(j)}|X_t^{(<j)}M_{t-1}\right)$. Taking an expectation over j, we get

$$\begin{aligned}
\mathbb{E}_j[\mathsf{IC}^{tr}(\mathcal{A}_{ran}|J=j)] &\leq \mathbb{E}_j\left(\frac{1}{n}\sum_{t=1}^{n} \mathbf{I}\left(M_t : X_t^{(j)}|X_t^{(<j)}M_{t-1})\right)\right) \\
&= \frac{1}{k}\left(\frac{1}{n}\sum_{t=1}^{n}\sum_{j=1}^{k} \mathbf{I}\left(M_t : X_t^{(j)}|X_t^{(<j)}M_{t-1}\right)\right)
\end{aligned}$$

From the chain rule the right hand side above equals

$$\frac{1}{k}\left(\frac{1}{n}\sum_{t=1}^{n} \mathbf{I}\left(M_t : X_t^{(1)}\ldots X_t^{(k)}|M_{t-1}\right)\right) = \frac{1}{k}\mathsf{IC}^{tr}(\mathcal{A}) \leq \frac{M}{k},$$

where the last inequality follows since the transitional information content is bounded by the average memory (Proposition 5.10).

Analogously, the cumulative information content of $\mathcal{A}_{ran}$ is given by

$$\begin{aligned}
\mathsf{IC}^{cum}(\mathcal{A}_{ran}|J=j) &= \frac{1}{n}\sum_{t=1}^{n}\mathbf{I}\left(M_t : X_{\leq t}|J=j\right) \\
&= \frac{1}{n}\sum_{t=1}^{n}\mathbf{I}\left(M_t : X^{(j)}_{\leq t}|J=j\right) && \text{(with probability 1, } X^{(j)}=X\text{)} \\
&= \frac{1}{n}\sum_{t=1}^{n}\mathbf{I}\left(M_t : X^{(j)}_{\leq t}\right) && \text{(}M_t \text{ is independent of the event } J=j\text{)}.
\end{aligned}$$

As $X^{(1)},\ldots,X^{(k)}$ are all independent, by Lemma 2.12, it follows that $\mathbf{I}\left(M_t : X^{(j)}_{\leq t}\right)$ is at most $\mathbf{I}\left(M_t : X^{(j)}_{\leq t}|X^{(<j)}_{\leq t}\right)$. Taking an expectation over j and using the chain rule, we get

$$\begin{aligned}
\mathbb{E}_j[\mathsf{IC}^{cum}(\mathcal{A}_{ran}|J=j)] &\leq \mathbb{E}_j\left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{I}\left(M_t : X^{(j)}_{\leq t}|X^{(<j)}_{\leq t})\right)\right) \\
&= \frac{1}{k}\left(\frac{1}{n}\sum_{t=1}^{n}\sum_{j=1}^{k}\mathbf{I}\left(M_t : X^{(j)}_{\leq t}|X^{(<j)}_{\leq t}\right)\right) \\
&= \frac{1}{k}\left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{I}\left(M_t : X^{(1)}_{\leq t}\ldots X^{(k)}_{\leq t}\right)\right) = \frac{1}{k}\mathsf{IC}^{cum}(\mathcal{A}) \leq \frac{M}{k}.
\end{aligned}$$

□

Direct-sum Theorem

With the above, we can now apply Theorem 5.12 to get Theorem 5.15.

Proof of Theorem 5.15. Let $\mathcal{A}$ be a streaming algorithm that computes f^k with error at most $1/3$ and average memory M. By Lemma 5.16, there is an algorithm $\mathcal{A}_{ran}$ that uses randomness j and r, computes f with error at most $1/3$ and satisfies $\mathbb{E}_j[\mathsf{IC}^{tr}(\mathcal{A}_{ran})|j] \leq M/k$. Applying Theorem 5.12 to $\mathcal{A}_{ran}$ gives us a randomized algorithm that uses random coins j and r' and computes f using average memory $\mathbb{E}_{j,r'}[\frac{1}{n}\sum_{t=1}^{n}|M_t|] = O(nM/k+n)$.

Since the random coins j and r' are independent of the input, we can fix them to get a deterministic streaming algorithm with average memory $O(nM/k+n)$. Since this must be at least $M(f)$, we have

$$O\left(\frac{nM}{k}+n\right) \geq M(f).$$

Rearranging the above gives us that M is lower bounded by $\Omega\left(k\left(\frac{M(f)}{n}-1\right)\right)$. □

5.5 Towards Optimal Direct Sums

The algorithm $\mathcal{A}_{ran}$ that we gave in the last section also had cumulative information content at most M/k as shown in Lemma 5.16. Analogous to Theorem 5.15, the following result follows. We omit the proof since it is very similar to that of Theorem 5.15.

Theorem 5.17 (Restated)**.** *If every algorithm with cumulative information I can be simulated using average memory $O(I)$, then $M(f^k) = \Omega\left(k \cdot (M(f) - 1)\right)$.*

In this section, we describe a compression algorithm that could possibly simulate an algorithm with cumulative information content I with average memory $O(I + 1)$. However, we are unable to either prove or disprove it.

To give some intuition about the new algorithm, let us recall Algorithm $\mathcal{A}_{tr}$ where the compression algorithm stored Huffman encodings (Proposition 2.5) of C_t satisfying $X_t - C_t S_t | M_{t-1} \to M_t | M_{t-1}$. This necessitated storing the whole history since to determine the sample m_t required knowing encodings of all the previous $c_{<t}$.

The new algorithm that we call $\mathcal{A}_{cum}$, on receiving the input x_t at time t, samples C_t conditioned on the value of x_t and m_{t-1} where C_t satisfies the following properties that follow from Lemma 5.6:

$$X_t - C_t S_t | S_{<t} \to M_t | S_{<t}, \tag{5.5.1}$$

$$\mathbf{H}(C_t | S_{\leq t}) \leq \mathbf{I}\left(M_t : X_t M_{t-1} | S_{<t}\right) + \log(\mathbf{I}\left(M_t : X_t M_{t-1} | S_{<t}\right) + 1) + O(1). \tag{5.5.2}$$

Again the value of m_t determined by the sample c_t is distributed according to the distribution $p(M_t | x_t, m_{t-1})$. Moreover, the algorithm $\mathcal{A}_{cum}$ will store the Huffman encoding of C_t conditioned on $S_{\leq t}$ which avoids the need to store all the previous memory contents since $S_{\leq t}$ is randomness independent of the input and can be fixed in the beginning.

Conjecture 5.18. *Let $\mathcal{A}$ be a randomized streaming algorithm with $\mathsf{IC}^{cum}(\mathcal{A}) = I$. Then, $\mathcal{A}_{cum}$ simulates $\mathcal{A}$ using $O(I + 1)$ average memory.*

The proof that the above compression algorithm gives a correct simulation is straightforward from (5.5.1). We are able to prove the following bounds on the memory used by the above algorithm.

Input : Stream x sampled from $p(X)$

Randomness: $s_1, \ldots, s_n$ where s_i is an infinite sequence of uniform samples from $\mathsf{supp}(M_i) \times [0,1]$.

// At time t: the content of the memory are some encodings of $c_{<t}$, where c_i determines m_i given $s_{\leq i}$.

1. Let m_{t-1} be determined by c_{t-1} and s_{t-1}. On input x_t, sample c_t from the Markov chain in (5.5.1);
2. Store the Huffman encoding of c_t conditioned on $s_{\leq t}$;

Randomized Streaming Algorithm $\mathcal{A}_{cum}$

Lemma 5.19. *In expectation over the choice of $s_{\leq t}$ and $x_{\leq t}$, the memory used by algorithm $\mathcal{A}_{cum}$ at time t is at most $O(\mathbf{I}(M_t : X_{\leq t}|S_{<t}) + 1)$.*

Proof. The memory used by algorithm $\mathcal{A}_{cum}$ at time t is bounded by $\mathbf{H}(C_t|S_{\leq t})$ which as given by (5.5.2) is at most $O(\mathbf{I}(M_t : X_t M_{t-1}|S_{<t}) + 1)$. Moreover, since $M_{t-1}|S_{<t}$ is determined given $C_{t-1}|S_{<t}$,

$$\mathbf{I}(M_t : X_t M_{t-1}|S_{<t}) \leq \mathbf{I}(M_t : X_t C_{t-1}|S_{<t}),$$

by the data processing inequality (Proposition 2.13).

To finish the proof, we show that $\mathbf{I}(M_t : X_t C_{t-1}|S_{<t})$ is upper bounded by $\mathbf{I}(M_t : X_{\leq t}|S_{<t})$. To show this we apply the chain rule as follows,

$$\begin{aligned}
\mathbf{I}(M_t : X_t C_{t-1}|S_{<t}) &\leq \mathbf{I}(M_t : X_{\leq t} C_{t-1}|S_{<t}) \\
&= \mathbf{I}(M_t : X_{<t} C_{t-1}|S_{<t}) + \mathbf{I}(M_t : X_t|S_{<t} X_{<t} C_{t-1}). \\
&= \mathbf{I}(M_t : X_{<t}|S_{<t}) + \mathbf{I}(M_t : C_{t-1}|S_{<t} X_{<t}) \\
&\quad + \mathbf{I}(M_t : X_t|S_{<t} X_{<t} C_{t-1}).
\end{aligned}$$

Note that in the algorithm $\mathcal{A}_{cum}$, $X_{<t}$ and $S_{<t}$ completely determine C_{t-1}. Hence, the second term in the above expression is 0. Moreover, by the same fact $M_t X_t - S_{<t} X_{<t} \to C_{t-1}$ and hence by Proposition 2.15, the last term $\mathbf{I}(M_t : X_t|S_{<t} X_{<t} C_{t-1}) = \mathbf{I}(M_t : X_t|X_{<t} S_{<t})$.

The above discussion yields that

$$\begin{aligned}\mathbf{I}\left(M_t : X_t C_{t-1} | S_{<t}\right) &\leq \mathbf{I}\left(M_t : X_{<t} | S_{<t}\right) + \mathbf{I}\left(M_t : X_t | X_{<t} S_{<t}\right) \\ &= \mathbf{I}\left(M_t : X_{\leq t} | S_{<t}\right),\end{aligned}$$

where the second equality follows by another application of the chain rule. □

Note that since $S_{<t}$ is independent of $X_{\leq t}$, the above quantity $\mathbf{I}\left(M_t : X_{\leq t} | S_{<t}\right)$ is at least as large as $\mathbf{I}\left(M_t : X_{\leq t}\right)$ (recall Lemma 2.12), but it is possible that similar to Lemma 5.6, they could be the same up to some lower order error terms. Towards proving such a statement, we first propose to investigate whether the following stronger version of Lemma 5.6 holds.

Conjecture 5.20. *Let X and M be arbitrary discrete random variables with finite support. Let S be an infinite list of samples from $\mathsf{supp}(M) \times [0,1]$. Then, there exist a random variable C such that*

- $X - CS \to M$.
- $\mathbf{H}(C|S) \leq \mathbf{I}\left(M : X\right) + \log(\mathbf{I}\left(M : X\right) + 1) + O(1)$.
- *For any discrete random variable N such that $X - M - N$, it holds that*

$$\mathbf{I}\left(N : M | S\right) \leq \mathbf{I}\left(N : X\right) + \log(\mathbf{I}\left(N : X\right) + 1) + O(1).$$

We also point out that an inductive use of the above conjecture does not give a non-trivial upper bound on the memory used by the algorithm $\mathcal{A}_{cum}$ because of the error terms in the last statement of the conjecture. But we hope that the techniques used in proving the above conjecture would be helpful in analyzing the memory used by the algorithm $\mathcal{A}_{cum}$. Nonetheless the above conjecture might be interesting in its own right and of potential use somewhere else.

§ Part II §

Linear Programs

6 Extended Formulations and Non-Negative Rank

In the late 1970s, Khachiyan [Kha79] showed that linear programs can be solved in polynomial time. This opened the door to the use of linear programming as a powerful tool in designing polynomial-time algorithms. Many combinatorial optimization problems can be written down naturally as optimizing a linear objective function over a polytope which is the convex hull of all feasible integer solutions. For example, the traveling salesman and maximum matching problems can be expressed as optimizing over the following polytopes respectively:

$$P_{TSP}(n) = \mathsf{conv}\left\{ \mathbb{1}_T \in \mathbb{R}^{\binom{[n]}{2}} \;\middle|\; T \subseteq \binom{[n]}{2} \text{ is a tour in } K_n \right\}$$

$$P_{MAT}(n) = \mathsf{conv}\left\{ \mathbb{1}_M \in \mathbb{R}^{\binom{[n]}{2}} \;\middle|\; M \subseteq \binom{[n]}{2} \text{ is a matching in } K_n \right\}$$

The convex hull of the integer solutions to the above problems are known as the TSP and Matching Polytopes respectively.

For many combinatorial optimization problems, the natural polytopes representing them usually have exponentially many defining inequalities or facets in terms of the size of the instance. This means that if the instance is of size n, one needs to write down exponential in n linear inequalities to even describe the linear program. There are some natural combinatorial optimization problems, such as finding the minimum cost spanning tree or the maximum weight matching, where the natural linear programs have exponential size in n but they can still be solved in time polynomial in n. In these cases, there is a *separation oracle* for these problems, which runs in time polynomial in n, and when

given a point x, answers if x satisfies all the constraints of the linear program, and if it does not, it outputs a violating constraint. One can then use Khachiyan's algorithm to solve the linear program in time polynomial in n. However, for most problems of interest, we do not know how to construct polynomial time separation oracles or whether such oracles even exist.

A beautiful yet simple idea to reduce the size of such linear programs is to use auxiliary variables — such linear programs are usually called *(linear) extended formulations*. As we will see later in this chapter, extended formulations can be exponentially smaller in size than the original linear program in some cases. The best known approximation algorithms for many problems also make use of extended formulations. In this chapter, we will look at what can be said about the power and limitations of such algorithms: Could they solve NP-hard combinatorial optimization problems in polynomial time?

Yannakakis [Yan91] proved that showing limitations on the power of such algorithms boils down to a question about analyzing the *non-negative rank* of certain matrices. In recent years, ideas from communication and information complexity have given us a greater understanding of non-negative rank. This has led to breakthroughs towards answering long-standing questions concerning the limitations of extended formulation based algorithms. We now know that such algorithms cannot solve NP-hard problems like TSP or MAXCUT in polynomial time [FMP+15, CLRS13] and, more surprisingly, they also can not solve the matching problem in polynomial time [Rot17]. In this chapter, we shall chronicle these recent developments.

6.1 Extended Formulations and Extension Complexity

Given a combinatorial optimization problem, let us consider the following way of capturing it with polytopes (or equivalently linear programs). Let $P = P(n)$ be the polytope that is the convex hull of all feasible integer solutions among instances of size n. The optimization problem asks to solve

$$\text{maximize } w^T x$$
$$\text{subject to } x \in P,$$

where w is an objective function that comes from a set of admissible functions.

Let us see some examples to grasp the above. The maximum matching problem asks to find a maximum weighted matching in a given graph. Here, we can take $P = P_{MAT}$ where P_{MAT} is the matching polytope, while the set of admissible functions consists of all non-negative weights w on the edges of the graph. Note that in this case, the graph is encoded only in the objective function w. As another example, consider the maximum weighted independent set problem which asks for the weight of the maximum independent set in a given graph G. In this case, we can take P to be the convex hull of all independent sets in the given graph G, while the set of admissible functions consists of all non-negative weights w on the vertices of the graph G.

Note that the above is not the most general way of capturing a combinatorial optimization problem with linear programs. For example, for the maximum matching problem, an algorithm could encode the graph G in the polytope P, as opposed to optimizing over the matching polytope as in our formulation above. Nevertheless, the formulation above captures natural ways of writing linear programs for many combinatorial optimization problems.

Now that we have formulated how to express an optimization problem as a polytope, we can define the notion of extended formulations. A polytope $K \subseteq \mathbb{R}^{d+r}$ is called a (linear) extended formulation of a given polytope $P \subseteq \mathbb{R}^d$, if there is a linear map $\pi : \mathbb{R}^{d+r} \to \mathbb{R}^d$ such that $\pi(K) = P$. The extended formulation K is also sometimes referred to as a *higher-dimensional lift* of the polytope P. If the polytope P is defined in terms of the variables x, then by performing an appropriate linear transformation, one may assume without loss of generality that the linear map π projects identically on the variables x. In other words, without loss of generality, the polytopes P, K and the linear map π are of the following general form:

$$
\begin{aligned}
&P = \{x \mid Ax \leq b\} \subseteq \mathbb{R}^d, \\
&K = \{(x,y) \mid Ex + Fy = g, y \geq 0\} \subseteq \mathbb{R}^{d+r}, \\
&\pi : \mathbb{R}^{d+r} \to \mathbb{R}^d, \;\; \pi(x,y) = x.
\end{aligned}
$$

Viewed in geometric terms, extended formulations can have fewer facets than the original polytope (see Figure 6.1.1) . This means that extended formulations can be much simpler to write down as

linear programs. If one can find a significantly smaller extended formulation for a polytope of interest, then one can solve a linear optimization problem over the original polytope much faster by optimizing over the extended formulation instead. In light of this, there has been a lot of interest in understanding extended formulations of various natural polytopes. Motivated by this, we define the *extension complexity* of P denoted as $\mathsf{xc}(P)$ to be the minimal number of facets in any extended formulation of P.

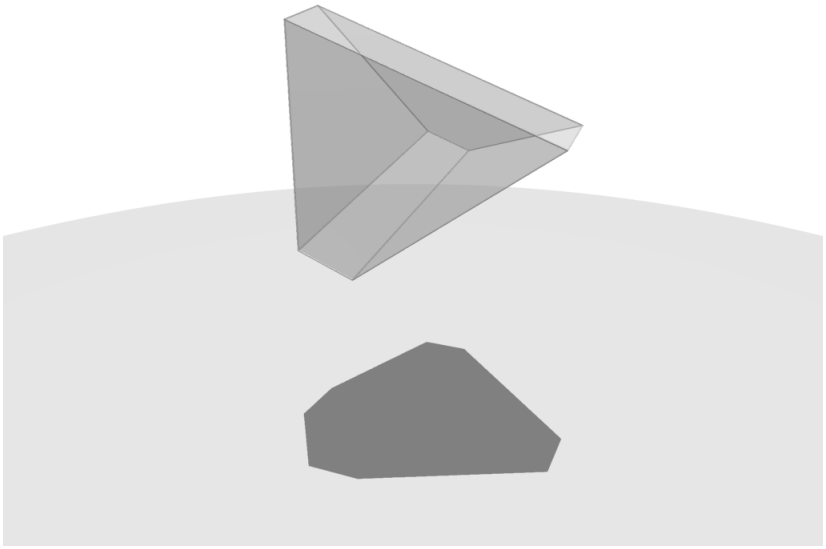


Figure 6.1.1: An example of an extended formulation of a two dimensional polytope. Note that the original two dimensional polytope has eight facets while the extended formulation has six facets.

There are various natural polytopes that have exponentially many facets but only polynomial-sized extended formulations. Below we give examples of a few such polytopes.

Spanning Tree Polytope Consider the spanning tree polytope of the complete graph K_n defined as

$$P_{ST}(n) = \mathsf{conv}\left\{ \mathbb{1}_T \in \mathbb{R}^{\binom{[n]}{2}} \;\middle|\; T \subseteq \binom{[n]}{2} \text{ is a spanning tree of } K_n. \right\}$$

Edmonds [Edm71] proved that the polytope $P_{ST}(n)$ is completely described by the following linear system:

$$\begin{aligned} x &\geq 0; \\ \sum_e x_e &= n-1; \\ \sum_{e \text{ inside } U} x_e &\leq |U|-1 \,, \forall U \subseteq [n]. \end{aligned}$$

Note that $P_{ST}(n)$ has $2^{\Omega(n)}$ facets. However, there is an extended formulation of $P_{ST}(n)$ of size $O(n^3)$ [Mar91].

To describe the extended formulation, we introduce auxiliary variables $z_{(a,b,c)}$ for all ordered triples of vertices (a,b,c) which is meant to encode if $\{a,b\}$ is an edge in the spanning tree and whether removing the edge $\{a,b\}$ leaves c and b in the same connected component. Formally consider the following polytope $K(n)$ defined in the original variables $x \in \mathbb{R}^{\binom{[n]}{2}}$ and auxiliary variables $z \in \mathbb{R}^{[n]\times[n-1]\times[n-2]}$ and described by the following linear constraints:

$$\begin{aligned} \sum_e x_e &= n-1; \\ x_{\{a,b\}} - z_{(a,b,c)} - z_{(b,a,c)} &= 0, \ \forall \text{pairwise distinct } a,b,c; \\ x_{\{a,b\}} + \sum_{c \in [n]\setminus\{b,c\}} z_{(a,c,b)} &= 1, \ \forall \text{distinct } a,b; \\ x \geq 0; z &\geq 0. \end{aligned}$$

Note that in the above formulation, the x-variables are defined with respect to unordered pairs while the z-variables are defined with respect to ordered triples. The polytope $K(n)$ defined above has $O(n^3)$ facets and Martin [Mar91] showed that the projection of $K(n)$ on the x-variables is the polytope $P_{ST}(n)$.

Permutahedron Consider the polytope known as the *permutahedron*:

$$P_{PERM}(n) = \text{conv}\left\{ x \in \mathbb{R}^n \;\middle|\; x \text{ is a permutation of } [n]. \right\}$$

Rado [Rad52] showed that the facets of the permutahedron are completely described by the $2^n - 2$ inequalities given below:

$$\begin{aligned} x &\geq 0; \\ \sum_i x_i &= \binom{n+1}{2}; \\ \sum_{i \in S} x_i &\geq \binom{|S|+1}{2} \ \forall S, \emptyset \neq S \subsetneq [n]. \end{aligned}$$

On the other hand, it turns out that there is an extended formulation of $P_{PERM}(n)$ [Goe15] that is of size $O(n \log n)$.

There are many other examples where the size of extended formulations is significantly smaller than that of the original polytope. For a reference, see the survey by Conforti, Cornuéjols and Zambelli [CCZ10].

6.1.1 Approximation using Extended Formulations

The definition of extended formulations given above only captures linear programs that solve the original optimization problem exactly. This does not capture many approximation algorithms which use linear programs with auxiliary variables and only yield approximate solutions. To include such algorithms, we need the notion of approximate extended formulations. Informally in these cases, the higher dimensional lift will not project to the original polytope but will project to some other polytope that approximates the original one.

Let us consider the case of maximization problems and let OPT and LP respectively denote the optimal solution value and the maximum value of the approximating linear program. We say that the approximating linear program has approximation ratio α if $OPT \leq LP \leq \alpha \cdot OPT$. To capture this property with extended formulations, let $P \subseteq \mathbb{R}^d$ denote the polytope corresponding to the convex hull of all feasible integer solutions for the problem. Then, we say that a polytope $K = \{(x, y) \in$

$\mathbb{R}^{d+r} \mid Ex + Fy = g, y \geq 0\}$ is an *α-approximate extended formulation* of P if

$$\max\{w^T x \mid (x,y) \in K\} \geq \max\{w^T x \mid x \in P\} \ \forall w, \text{ and}$$
$$\max\{w^T x \mid (x,y) \in K\} \leq \alpha \max\{w^T x \mid x \in P\} \ \forall \text{ admissible } w.$$

We can also give an equivalent definition in terms of a projecting linear map. Let $\pi : \mathbb{R}^{d+r} \to \mathbb{R}^d$ denote the linear map $\pi(x,y) = x$ and let Q denote the following convex set:

$$Q = \{x \in \mathbb{R}^d \mid w^T x \leq \max\{w^T z \mid z \in P\} \ \forall \text{ admissible } w\}.$$

Then, saying that K is an α-approximate extended formulation of P is equivalent to saying that $P \subseteq \pi(K) \subseteq \alpha Q$. Note that in general, the set Q might not always be bounded or even polyhedral, but in many interesting cases, it is a polytope and for the rest of this thesis we shall only be concerned with such cases. When the polytope P is *monotone down-closed* (*i.e.* $y \in P$ implies $x \in P$ if $x \leq y$) the set Q can in fact be taken to be the original polytope itself. For example, the matching polytope is monotone down-closed, so when P is the matching polytope, an α-approximate extended formulation K for it satisfies $P \subseteq \pi(K) \subseteq \alpha P$.

We will omit it here, but one can define the notion of approximate extended formulations for minimization problems in an analogous way.

6.2 Slack Matrices and Non-negative Rank

It turns out that the extension complexity of a polytope is exactly characterized by the *non-negative rank* of a certain *slack matrix* associated with the polytope. This connection was discovered by Yannakakis [Yan91] who proved the first lower bounds for extended formulations that satisfy certain symmetry properties.

The *non-negative rank* of a non-negative matrix $S \in \mathbb{R}^{s \times t}$ is defined as

$$\mathsf{rk}_+(S) = \min\{r | \exists U \in \mathbb{R}^{s \times r}_{\geq 0}, V \in \mathbb{R}^{r \times t}_{\geq 0} : S = UV\}.$$

Alternatively, saying that $\mathsf{rk}_+(S) = r$ is equivalent to saying that S can be written as a sum of r non-negative rank one matrices, *i.e.* there are vectors $u_1, \cdots, u_r \in \mathbb{R}^s_{\geq 0}$ and $v_1, \cdots, v_r \in \mathbb{R}^t_{\geq 0}$ such

that $S = \sum_{i=1}^{r} u_i v_i^T$. Note that if we drop the non-negativity constraint from the above definition, we get the usual definition of rank of a matrix.

To state the connection between extension complexity and non-negative rank, let $P \subseteq \mathbb{R}^d$ be a polytope. Denote by $\{x_1, \ldots, x_s\}$ the vertices of the polytope P and let $\langle a_i, x \rangle \leq b_i$, $\forall i \in [t]$ denote the facets of the polytope P. Then, the slack matrix S^P with respect to the polytope P is a $s \times t$ matrix with non-negative entries defined by $S^P_{ij} = b_i - \langle a_i, x_j \rangle$. The rows of this matrix are indexed by the vertices of P, the columns are indexed by facets of P, while the entry corresponding to a vertex and a facet pair is the distance of the vertex from the facet.

Yannakakis proved that the extension complexity of a polytope is equal to the non-negative rank of the associated slack matrix up to an additive value of one.

Theorem 6.1 ([Yan91]). *For any polytope P, we have $\mathsf{rk}_+(S^P) - 1 \leq \mathsf{xc}(P) \leq \mathsf{rk}_+(S^P)$. Moreover, if P has dimension at least one, then $\mathsf{xc}(P) = \mathsf{rk}_+(S^P)$.*

The above theorem is quite powerful as it reduces the question of proving upper and lower bounds on the size of extended formulations to analyzing non-negative rank of slack matrices.

For analyzing approximate extended formulations, we need a generalization of the above theorem. Recall that an approximate extended formulation K is associated with polytopes P and Q such that $P \subseteq Q$ and there is a linear map π so that $P \subseteq \pi(K) \subseteq Q$. Let $P = \mathsf{conv}\{x_1, \ldots, x_s\} \subseteq \mathbb{R}^d$ and $Q = \{x \in \mathbb{R}^d \mid \langle a_i, x \rangle \leq b_i, \ \forall i \in [t]\} \subseteq \mathbb{R}^d$ be such that $P \subseteq Q$. The *Slack Matrix* $S^{P,Q} \in \mathbb{R}^{s \times t}$ with respect to an inner polytope P and an outer polytope Q is a non-negative matrix defined by $S^{P,Q}_{ij} = b_i - \langle a_i, x_j \rangle$. Note that rows of this matrix are indexed by the vertices of the inner polytope P while the columns are indexed by the facets of Q.

Pashkovich [Pas12] and independently, Braun, Fiorini, Pokutta and Steurer [BFPS15b] showed that to lower bound the size of extended formulations of a polytope sandwiched between an inner polytope P and an outer polytope Q, it suffices to lower bound the non-negative rank of the corresponding slack matrix $S^{P,Q}$. They proved the following generalization of Yannakakis's theorem.

Theorem 6.2 ([Pas12, BFPS15b]). *For any polytopes P and Q such that $P \subseteq Q$, we have that*

$$\mathsf{rk}_+(S^{P,Q}) - 1 \leq \min_{P \subseteq K \subseteq Q} \mathsf{xc}(K) \leq \mathsf{rk}_+(S^{P,Q}).$$

Moreover, if P and Q have dimension at least one, then $\min_{P \subseteq K \subseteq Q} \mathsf{xc}(K) = \mathsf{rk}_+(S^{P,Q})$.

Note that the above theorem also says that to prove a lower bound on the extension complexity of a polytope P, we can choose any polytope Q that contains it and prove bounds on the non-negative rank of the slack matrix associated with P and Q. The choice of the outer polytope gives us some freedom in choosing an appropriate slack matrix so that the non-negative rank is easier to analyze. This will be very useful for us in some cases.

6.3 Lower Bounds on Extension Complexity and Size of Linear Programs

The first lower bounds for extended formulations were proved in the seminal paper by Yannakakis [Yan91] which related the extension complexity to the non-negative rank of slack matrices. He showed that any *symmetric* extended formulation (one whose formulation is invariant under permutation of variables) that projects to the TSP and Matching Polytopes in $\mathbb{R}^{\binom{[n]}{2}}$ must have size $2^{\Omega(n)}$. This refuted a purported $\mathsf{P} = \mathsf{NP}$ proof which encoded the Traveling Salesman problem as a polynomial size linear program.

The question of proving lower bounds on the size of extended formulations without such symmetry properties remained open for a long time as there were no techniques to analyze the non-negative rank. In a breakthrough work Fiorini, Massar, Pokutta, Tiwary and de Wolf [FMP+15] used ideas from the lower bound proofs for communication complexity of disjointness to prove lower bounds on non-negative rank. Using this connection they managed to prove exponential lower bounds on the extension complexity of the TSP and Matching Polytopes. Since then there has been a lot of progress in proving strong lower bounds for extension complexity of various natural polytopes. In this section, we briefly survey known lower bounds. For a summary and comparison of all the results in this section, see Table 6.1.

Polytope or Optimization Problem	Notes	Dimension/ Instance Size	Extension Complexity	Reference
TSP & Matching Polytopes	Symmetric Lifts	$d = \binom{n}{2}$	$2^{\Theta(n)}$	[Yan91]
Random Polytope		d	$2^{\Theta(d)}$	[Rot13]
TSP Polytope		$d = \binom{n}{2}$	$2^{\Omega(\sqrt{n})}$	[FMP+15]
Correlation Polytope		$d = n^2$	$2^{\Theta(n)}$	[FMP+15]
Cut Polytope		$d = \binom{n}{2}$	$2^{\Theta(n)}$	[FMP+15]
SAT, Knapsack, Independent Set Polytopes		$d = n$	$2^{\Omega(\sqrt{d})}$	[AT13, PV13]
Correlation Polytope	α approx	$d = n^2$	$2^{\Theta(n/\alpha)}$	[FMP+15]
TSP & Matching Polytopes		$d = \binom{n}{2}$	$2^{\Theta(n)}$	[Rot17]
Matching Polytope	$1 + \frac{1}{n}$ approx	$d = \binom{n}{2}$	$2^{\Theta(n)}$	[BP15]
SAT, Knapsack, Independent Set Polytopes		$d = n$	$2^{\Omega\left(\frac{n}{\log n}\right)}$	[GJW16a]
MAXCUT	$2 + \varepsilon$ approx	n	$2^{\Omega\left(\frac{\log^2 n}{\log\log n}\right)}$ 2^{n^γ}	[CLRS13] [KMR17]

Table 6.1: Comparison of the extension complexity lower bounds for various polytopes. Constant factors are suppressed for readability. In the above, d denotes the dimension of the polytope and n denotes the instance size. In the last row, $\gamma = \gamma(\varepsilon)$ is a constant that depends on ε.

6.3.1 Correlation Polytope and Unique Disjointness

A polytope that is central to proving strong exponential lower bounds for extension complexity is the so called *correlation* polytope. The correlation polytope $P_{CORR} \subseteq \mathbb{R}^{n \times n}$ is defined as the convex hull of all binary $n \times n$ rank one matrices:

$$P_{CORR} = \mathsf{conv}\left\{ bb^T \;\middle|\; b \in \{0,1\}^n \right\}.$$

Fiorini, Massar, Pokutta, Tiwary and de Wolf [FMP+15] proved that the extension complexity of the correlation polytope is $2^{\Omega(n)}$. To prove this, they implicitly used a version of Theorem 6.2. The slack matrix for the polytope P_{CORR} is difficult to analyze directly, but by choosing a suitable outer polytope Q, we get a much nicer slack matrix. We shall take a detailed look at this matrix since this will be very relevant to Chapter 7.

Consider the following outer polytope

$$Q = Q(n) = \{x \in \mathbb{R}^{n \times n} \mid \langle 2\mathsf{diag}(a) - aa^T, x \rangle \leq 1, a \in \{0,1\}^n\},$$

where $\mathsf{diag}(a) \in \mathbb{R}^{n \times n}$ is the diagonal matrix which has the vector a on the diagonal and is zero otherwise.

We now proceed to show that $P_{CORR} \subseteq Q$. For any vector $b \in \{0,1\}^n$, consider the vertex bb^T of P_{CORR}. Then, we have that

$$\begin{aligned} 1 - \langle 2\mathsf{diag}(a) - aa^T, bb^T \rangle &= 1 + \langle a, b \rangle^2 - 2 \left\langle \mathsf{diag}(a), bb^T \right\rangle \\ &= 1 + \langle a, b \rangle^2 - 2 \langle a, b \rangle = (1 - \langle a, b \rangle)^2 \geq 0, \end{aligned}$$

where the second equality follows since $(bb^T)_{ii} = b_i^2 = b_i$ since b is a $0/1$ vector.

The above means that for all vertices x of P_{CORR} the inequality $\langle 2\mathsf{diag}(a) - aa^T, x \rangle \leq 1$ is valid. By convexity, it also follows that the inequality is valid for all points $x \in P_{CORR}$. It follows that $P_{CORR} \subseteq Q$.

Furthermore, the slack matrix S of the correlation polytope with respect to the inner polytope P_{CORR} and the outer polytope Q is indexed by vectors $a, b \in \{0,1\}^n$ and the entries are given by

$$S_{ab} = 1 - \left\langle 2\mathsf{diag}(a) - aa^T, bb^T \right\rangle = (1 - \langle a, b \rangle)^2.$$

As we have seen above, this is a non-negative matrix. Recalling Theorem 6.2, to prove a lower bound on the extension complexity of P_{CORR}, it suffices to prove a lower on the non-negative rank of the slack matrix S defined above.

Note that the matrix S has the following property: if we consider a and b as indicator vectors for subsets of $[n]$, then the entry $S_{ab} = 0$ when the sets intersect uniquely while the entry $S_{ab} = 1$ when the sets are disjoint. Let us call a non-negative matrix A indexed by vectors $a, b \in \{0,1\}^n$ and satisfying

$$A_{ab} = \begin{cases} 1, \text{ if } |a \cap b| = 0 \text{ and} \\ \leq 1 - \rho, \text{ if } |a \cap b| = 1. \end{cases} \tag{6.3.1}$$

as a *unique disjointness* matrix with parameter ρ. Note that we do not have any constraints on the entries that do not correspond to disjoint or uniquely intersecting sets — they could be arbitrary non-negative real numbers. The slack matrix S defined above is a unique disjointness matrix with parameter $\rho = 0$.

Fiorini *et al.* [FMP+15] used techniques used in proving communication lower bounds for disjointness to prove that the non-negative rank of any unique disjointness matrix with parameter $\rho = 0$ is $2^{\Omega(n)}$ proving that $\mathsf{xc}(P_{CORR}) = 2^{\Omega(n)}$. A sequence of works by Braun *et al.* [BFPS15b], Braverman and Moitra [BM13] and Braun and Pokutta [BP16] culminated in proving that for an arbitrary $0 \leq \rho \leq 1$, the non-negative rank of any unique disjointness matrix with parameter ρ is $2^{\Omega(\rho n)}$. In Chapter 7, we will prove a tight lower bound for a lopsided version of the unique disjointness matrix (where sets a and b are of different sizes).

The lower bound on the non-negative rank of a unique disjointness matrix with parameter ρ has implications for approximate extended formulations for the correlation polytope which we now describe. Firstly, it is not hard to see that an α-approximate extended formulation K for P_{CORR} satisfies $P_{CORR} \subseteq \pi(K) \subseteq \alpha Q$ for some linear map π. Easy calculations show that the slack matrix with respect to the inner polytope P and the outer polytope αQ is a matrix indexed by $a, b \in \{0,1\}^n$ and the entry corresponding to disjoint sets a, b is α and the entry corresponding to uniquely intersecting sets a, b is $\alpha - 1$. Scaling it by a factor of $1/\alpha$, it follows that this slack matrix is a unique disjointness

matrix with parameter $\rho = 1/\alpha$. Hence, using Theorem 6.2, a non-negative rank lower bound of $2^{\Omega(n/\alpha)}$ implies that the extension complexity of any polytope K that satisfies $P_{CORR} \subseteq K \subseteq \alpha Q$ is $2^{\Omega(n/\alpha)}$.

An implication of the above result concerns the Maximum Clique problem which can be naturally written as a linear optimization problem over the correlation polytope (see [BM13]). Braun, Fiorini, Pokutta and Steurer [BFPS15b] observed that if we had an α-approximate extended formulation K satisfying $P_{CORR} \subseteq K \subseteq \alpha Q$ then this would give us a linear program which gives an α approximation for the Maximum Clique problem. The lower bound of $2^{\Omega(n/\alpha)}$ for the size of any α-approximate extended formulation of such a polytope K then implies that any linear program of size $2^{O(n^\varepsilon)}$ that is a higher dimensional lift of K can not give a better than $n^{1-\varepsilon}$ approximation for the Maximum Clique problem. This was first established by Braverman and Moitra [BM13] and a different proof was given later by Braun and Pokutta [BP16]. These results match the parameters for known NP-hardness inapproximability result for the Maximum Clique problem from the celebrated work of Håstad [Hås96].

6.3.2 Linear Embeddings of Polytopes

In the last section, we described a way to prove a lower bound on the extension complexity of the correlation polytope. We now show how this can be used to obtain extension complexity lower bounds for other important polytopes that arise in combinatorial optimization.

The underlying idea is quite general and based on the observation that, in many cases, we can give a linear transformation mapping a polytope P into another polytope Q or we can embed the polytope P into a face of the polytope Q using a linear transformation. Such operations are very useful for us since if we have a bound on the extension complexity of P then we also get a bound on the extension complexity of Q.

Proposition 6.3. *Let P, Q and F be polytopes. Then, the following holds:*

(i) If F is an extended formulation of P, then $\mathsf{xc}(F) \geq \mathsf{xc}(P)$.

(ii) If F is a face of Q, then $\mathsf{xc}(Q) \geq \mathsf{xc}(F)$.

The Correlation polytope turns out to be fundamental in proving strong lower bounds for many other polytopes since it can be linearly transformed or embedded into them. Let us look at some examples.

Cut Polytope The Cut polytope on n vertices is defined to be the convex hull of all cuts in the complete graph. For a subset of vertices $T \subseteq [n]$, define $\delta(T)$ to be the set of edges with exactly one endpoint in T. Then, the Cut polytope $P_{CUT}(n) \subseteq \mathbb{R}^{\binom{[n]}{2}}$ is defined to be

$$P_{CUT}(n) = \mathsf{conv}\left\{ \mathbb{1}_{\delta(T)} \in \mathbb{R}^{\binom{[n]}{2}} \;\middle|\; T \subseteq [n] \right\}.$$

It is well known that there is a linear bijection between $P_{CUT}(n+1)$ and $P_{CORR}(n)$ (see [FMP+15] for example). Hence, from the known lower bounds on extension complexity of P_{CORR}, it follows that the extension complexity of $P_{CUT}(n)$ is also $2^{\Omega(n)}$. Note that this lower bound is tight up to constant factors in the exponent since the cut polytope has $2^{O(n)}$ facets.

TSP Polytope Fiorini *et al.* [FMP+15] proved that the $P_{CORR}(\Theta(\sqrt{n}))$ is a projection of a face of $P_{TSP}(n)$. This implied that the extension complexity of the polytope $P_{TSP}(n)$ is $2^{\Omega(\sqrt{n})}$, solving the long-standing open problem following the work of Yannakakis [Yan91]. Later work by Rothvoß [Rot17] improved the extension complexity lower bound to $2^{\Omega(n)}$.

SAT, Stable Set, Subset Sum and Knapsack Polytopes The work by Fiorini *et al.* [FMP+15], Avis and Tiwary [AT13], and Pokutta and Van Vyve [PV13] further showed that there are families of SAT, Stable Set, Subset Sum and Knapsack polytopes in d dimensions so that the Correlation polytope on $\Theta(\sqrt{d})$ dimensions can be embedded as a face of these polytopes. This gave a lower bound of $2^{\Omega(\sqrt{d})}$ on the extension complexity of these polytopes. Later, Göös, Jain and Watson [GJW16b] gave a different construction based on Karchmer-Wigderson games and improved the lower bound to $2^{\Omega(d/\log d)}$.

Note that Rothvoß [Rot13] showed that a random $0/1$ polytope in d dimensions has extension complexity $2^{\Omega(d)}$ with high probability. So, the lower bounds given by Göös, Jain and Watson

[GJW16b] almost approach the optimal bounds one could prove for extension complexity of polytopes. We remark that the extension complexity of the Matching, Cut and Correlation polytopes in $d = \Theta(n^2)$ dimensions are upper bounded by $2^{O(n)} = 2^{O(\sqrt{d})}$, so one could not hope to prove a $2^{\Omega(d)}$ lower bound for their extension complexities.

6.3.3 Matching Polytope

The Matching Polytope is another example whose extension complexity was left open by the work of Yannakakis [Yan91]. It was proven by Edmonds [Edm65] that the facets of the Matching Polytope $P_{MAT}(n)$ are given by the following inequalities:

$$\begin{aligned} x &\geq 0; \\ \sum_{\substack{e \text{ incident} \\ \text{on } i}} x_e &\leq 1, \forall i \in [n]; \\ \sum_{\substack{e \text{ inside} \\ U}} x_e &\leq \frac{|U|-1}{2}, \forall U \subseteq [n], |U| \text{ odd}. \end{aligned}$$

The above shows that the Matching Polytope $P_{MAT}(n)$ has $2^{\Omega(n)}$ facets. Yannakakis [Yan91] showed that any symmetric extended formulation for $P_{MAT}(n)$ must also have $2^{\Omega(n)}$ facets and it remained a long-standing open problem if the same was true for arbitrary extended formulations.

The Matching Polytope is not known to be linearly reducible (as in Proposition 6.3) to the Correlation polytope. In fact, we should not expect it to be the case since optimizing over the Correlation polytope is NP-hard while one can optimize over the Matching Polytope in polynomial time. Even though the Matching Polytope $P_{MAT}(n)$ has $2^{\Omega(n)}$ facets, to optimize over it, one need not write down the linear program explicitly as there is a polynomial time *separation oracle* for the polytope (see [PR82]). What this means is that given an input graph G, in polynomial time, one can check if all the constraints of the Matching Polytope are satisfied and if not then one can find a violated constraint. With such a separation oracle, one can use the Ellipsoid method [Kha79] and optimize over the Matching Polytope in polynomial time. We emphasize that such an algorithm is not captured in the framework of extended formulations.

With the above in mind, the Matching Polytope presented an interesting example, since it seemed likely that its extension complexity was exponential even though one could optimize over it in polynomial time. If the Matching Polytope had truly exponential complexity, this would identify a major weakness of (linear) extended formulation based algorithms. In a breakthrough work, Rothvoß [Rot17] showed that this was indeed true by proving that the extension complexity of the Matching Polytope is $2^{\Omega(n)}$. Later, Braun and Pokutta [BP15] extended the ideas in [Rot17] and proved the same lower bound for $\left(1+\frac{1}{n}\right)$-approximate extended formulations for the Matching Polytope. In Chapter 7, we will generalize these results and prove a tight lower bound for α-approximate extended formulations for the Matching Polytope for all parameters α.

6.3.4 Lower Bounds for Maximum Constraint Satisfaction Problems

In this section, we look at linear programs for Maximum Constraint Satisfaction Problems. For illustrative purposes, we will choose the MAXCUT problem as our running example. The MAXCUT problem can be naturally defined as a linear optimization problem over the Cut Polytope and in Section 6.3.2, we derived a lower bound on the extension complexity of the Cut polytope by reducing it from the Correlation Polytope using Proposition 6.3. However, it turns out that for approximate extended formulations, such reductions do not work and so we will start from first principles.

Given a graph $G = ([n], E)$, let $x_1, \cdots, x_n \in \{\pm 1\}$ be variables where $x_u = 1$ iff the vertex $u \in [n]$ was selected in the cut. Let us denote the fraction of edges crossing a cut x in a graph $G = ([n], E)$ by $\mathsf{cut}_G(x) := \frac{1}{2|E|}\sum_{(u,v)\in E}(1 - x_u x_v)$. Then, the MAXCUT problem is defined as

$$\begin{aligned} &\text{maximize } \mathsf{cut}_G(x) \\ &\text{subject to } x \in \{\pm 1\}^n. \end{aligned}$$

Note that traditionally the Cut polytope is defined over the edges of the graph, but the formulation above is over the vertices of the graph. This is just chosen for notational convenience and all the results hold even if we consider variables corresponding to the edges. Moreover, note that scaling by the $1/2|E|$ factor ensures that the MAXCUT value for any graph is at most one.

Recall that we want to capture approximating linear programs of the form

$$\begin{aligned} &\text{maximize } \langle w_G, x \rangle \\ &\text{subject to } Ax \leq b, \end{aligned} \tag{6.3.2}$$

where the polytope $K = \{x \mid Ax \leq b\} \subseteq \mathbb{R}^n$ contains $\mathsf{conv}(\{\pm 1\}^n)$, the graph G is encoded only in the objective function and the value of the linear program is an α-approximation for the MAXCUT value. Such linear programs can be seen to be slightly more general than the framework of approximate extended formulations which deals with higher dimensional lifts of fixed polytopes. However, even though the polytopes K above can be arbitrary, it turns out that proving lower bounds for such linear programs still boils down to lower bounding the non-negative rank of certain slack matrices that can be described explicitly.

For the MAXCUT problem, Chan, Lee, Raghavendra and Steurer [CLRS13] showed that if a linear program of the form (6.3.2) achieves an approximation ratio of $2 + \varepsilon$ for a constant $\varepsilon > 0$, then it must have super-polynomially many inequalities. Later, Kothari, Meka and Raghavendra [KMR17] improved this lower bound to 2^{n^γ} where $\gamma = \gamma(\varepsilon) > 0$ is another constant that depends on ε. We remark that this technique is very general and yields strong lower bounds for other Maximum Constraint Satisfaction problems but we shall focus only on the case of MAXCUT here. We also note that by reductions given by [BFPS15a], these lower bounds also translate to similar lower bounds for approximating linear programs for the Minimum Vertex Cover and Maximum Independent Set problem.

The approach of [CLRS13] consists of showing that linear programs of the form (6.3.2) are not more powerful than the linear programs that come from the Sherali-Adams hierarchy. Let us briefly recall the definition of the Sherali-Adams hierarchy. Let ℓ be a linear map from the space of all multilinear degree d polynomials to the Reals. Note that the value of ℓ is specified for all polynomials by specifying the values over $\{\Pi_{i \in S} x_i \mid |S| \leq d\}$, the set of monomials of degree at most d.

A degree d Sherali-Adams relaxation for the MAXCUT problem is the following optimization

problem:

$$\begin{aligned} &\text{maximize } \ell(\mathsf{cut}_G(x)) \\ &\text{subject to } \sum_p \ell(p) = 1 \\ &\qquad \ell(p) \geq 0, \ \forall \text{non-negative polynomials } p \text{ that depend on at most } d \text{ variables,} \end{aligned}$$

where $\mathsf{cut}_G(x)$ is the degree two polynomial representing the MAXCUT objective function. Note that the above is a linear program where the variables are the values taken by the linear map ℓ on degree d monomials. It is not too difficult to see that the above optimization problem can be written as a linear program with $n^{O(d)}$ inequalities. By taking larger values of d, we can get linear programs that give better approximations. It is well-known that Sherali-Adams relaxations with polynomial degree cannot achieve better than $2 + \varepsilon$ approximation for MAXCUT for any constant $\varepsilon > 0$.

Theorem 6.4 ([CMM09]). *For any $\varepsilon > 0$, there is a constant $\gamma = \gamma(\varepsilon)$ and a graph G such that the MAXCUT value on G is at most $\frac{1}{2} + \varepsilon$ while the optimum value of any degree n^γ Sherali-Adams relaxation is at least $1 - \varepsilon$.*

Using the graph G given by the above theorem in a black-box manner, Kothari *et al.* [KMR17] construct a family of graphs where the MAXCUT value is at most $\frac{1}{2} + \varepsilon$ but the optimum value of any linear program of the form (6.3.2) is at least $1 - O(\varepsilon)$ if it has fewer than 2^{n^γ} inequalities for a constant $\gamma = \gamma(\varepsilon)$. The family of graphs is constructed by embedding the graph on n^c vertices for $0 < c < 1$ given by Theorem 6.4 into an n vertex graph according to a certain gadget. We shall refer the reader to the paper [KMR17] for more details since they are not directly relevant to the results of this thesis. Finally, we remark that the lower bounds of [CLRS13, KMR17] in fact follow from lower bounding the non-negative rank of a special class of slack matrices. We discuss this in Section 6.4.4.

6.3.5 Lower Bounds for Semidefinite Programs

Semidefinite extended formulations provide a way generalizing the definition of linear extended formulations to capture semidefinite programs (SDPs). We shall not define them here as they are not

directly relevant to this thesis, but a semidefinite extended formulation, also sometimes known as a *SDP lift*, gives a way to describe the original polytope or optimization problem as a suitable projection of the feasible region of a semidefinite program.

In Section 6.3.4, we saw that the works [CLRS13, KMR17] imply that small linear programs for Maximum Constraint Satisfaction problems are not much more powerful than the linear programs that come from the Sherali-Adams hierarchy. Lee, Raghavendra and Steurer [LRS15] proved that a similar situation arises for SDP relaxations: when it comes to Maximum Constraint Satisfaction problems, SDPs of size polynomial size cannot perform much better than SDPs that come from the *sums of squares* hierarchy of constant degree. The sums of squares hierarchy can be thought of as a generalization of the Sherali-Adams hierarchy to the setting of SDPs and strong lower bounds are known for SDPs that come from this hierarchy. Using the result of [LRS15], these lower bounds also imply lower bounds for the size of SDPs for Maximum Constraint Satisfaction problems. For instance, it proves that for MAX-3-SAT, no SDP of polynomial size can beat the trivial approximation factor of $7/8$.

Using the same framework, Lee *et al.* [LRS15] also prove that the TSP, Cut and Correlation polytopes also require semidefinite extended formulations of subexponential size. It is a major open problem to determine if the same holds for the Matching Polytope. At this point however, it is unclear, if the techniques used in [LRS15] could be used to prove a non-trivial lower bound for the semidefinite extension complexity of the Matching Polytope.

6.4 Techniques to Lower Bound Non-Negative Rank

In what follows, let $A \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ be a non-negative matrix with $\mathsf{rk}_+(A) = r$ and let $u_1, \cdots, u_r \in \mathbb{R}^{\mathcal{X}}_{\geq 0}$ and $v_1, \cdots, v_r \in \mathbb{R}^{\mathcal{Y}}_{\geq 0}$ be its non-negative rank decomposition. In other words, we can write

$$A_{xy} = \sum_{i=1}^{r} u_i(x)v_i(y).$$

Below, we look at various techniques to prove lower bounds on non-negative rank.

6.4.1 Rectangle Covering Bound

Since $A = \sum_{i=1} u_i v_i^T$ where u_i's and v_i's are non-negative, it follows that if an entry $u_i(x)v_i(y)$ is non-zero for some $i \in [r]$, then the corresponding entry A_{xy} must also be non-zero. In particular, for a non-negative matrix M, denoting by $\mathsf{supp}(M)$ the set of non-zero entries of M, we have that

$$\mathsf{supp}(A) \subseteq \cup_{i=1}^{r} \mathsf{supp}(u_i v_i^T).$$

Note that since $u_i v_i^T$ is a rank one matrix, $\mathsf{supp}(u_i v_i^T) = \mathsf{supp}(u_i) \times \mathsf{supp}(v_i^T)$ where $\mathsf{supp}(w)$ for a vector w denotes the set of non-zero coordinates in w. Hence, the support of each rank one factor is a rectangle. If follows that a non-negative rank decomposition gives us a set of rectangles whose union equals $\mathsf{supp}(A)$. Such a set of rectangles is called a *rectangle covering* of A and the size of the smallest rectangle covering is called as the *rectangle covering number* of A and denoted by $\mathsf{rc}(A)$. From the above, it follows that

$$\mathsf{rk}_+(A) \geq \mathsf{rc}(A).$$

The rectangle covering number is closely related to communication complexity. In particular, if A is a boolean matrix, then logarithm of the rectangle covering number is exactly the non-deterministic communication complexity of the boolean function to which A corresponds to. If A is not a boolean matrix, define the boolean matrix $\mathsf{supp}A \in \{0,1\}^{\mathcal{X}\times\mathcal{Y}}$ as

$$\mathsf{supp}A_{xy} = 1 \text{ iff } (x,y) \in \mathsf{supp}(A).$$

Then a lower bound on the non-deterministic communication complexity of $\mathsf{supp}A$ implies a lower bound on $\log(\mathsf{rc}(A))$ and hence, translates into a lower bound on $\mathsf{rk}_+(A)$.

The above connection was exploited by Fiorini *et al.* [FMP+15] to prove exponential lower bounds for the non-negative rank of any unique disjointness matrix with parameter $\rho = 0$. In particular, they observed that the lower bound on the non-deterministic communication complexity of disjointness given by de Wolf [dW03] already implied that the rectangle covering number of any unique disjointness matrix with parameter $\rho = 0$ is $2^{\Omega(n)}$, giving us the same lower bound on its non-negative rank.

6.4.2 Common Information

Given a non-negative matrix $A \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$, we can view it as a probability distribution $p(XY)$ supported over $\mathcal{X} \times \mathcal{Y}$ as follows:

$$p(xy) = \frac{A_{xy}}{\sum_{x',y'} A_{x'y'}}.$$

Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ denote random variables with distribution $p(XY)$.

In the distribution $p(XY)$, the random variables X and Y are correlated. A non-negative rank decomposition of A allows us to break the correlation between X and Y. Recalling that the notation $X - R - Y$ denotes that X, R and Y form a Markov chain, we have the the following proposition.

Proposition 6.5 ([BM13, BP16])**.** *There is a random variable R such that*

$$X - R - Y \text{ and } |\mathsf{supp}(R)| \leq \mathsf{rk}_+(A).$$

Proof. Recall that $A = \sum_{i=1}^{r} u_i v_i^T$ where $u_i \in \mathbb{R}^{\mathcal{X}}_{\geq 0}$ and $v_i \in \mathbb{R}^{\mathcal{Y}}_{\geq 0}$ are non-negative (column) vectors and r is the non-negative rank of A. Then, $p(XY)$ can be expressed as a convex combination of r product distributions by setting

$$p(xy|r) = \frac{u_r(x)v_r(y)}{\sum_{x',y'} u_r(x')v_r(y')} \text{ and } p(r) = \frac{\sum_{x',y'} u_r(x')v_r(y')}{\sum_{x',y'} A_{x'y'}}.$$ □

Hence, to prove a lower bound on $\mathsf{rk}_+(A)$, it suffices to prove that for any random variable R such that $X - R - Y$, the size of the support of R must be large. This is closely related to the *common information* between X and Y. Recall that the common information between X and Y is defined to be $\mathbb{C}(X : Y) = \inf_R \mathbf{I}(XY : R)$ where the infimum is taken over all finite random variables R satisfying $X - R - Y$. In particular, common information between X and Y gives a lower bound on the size of support of any R that breaks the correlation between X and Y and hence on the non-negative rank of A.

Braun and Pokutta [BP16] extended the above definition of common information to allow conditioning on other random variables and events in the probability space. Let Z be another random

variable and $\mathcal{E}$ be an event. The common information between X and Y conditioned on $Z, \mathcal{E}$ is defined as

$$\mathbb{C}\left(X : Y | Z\mathcal{E}\right) = \inf_{R} \mathbf{I}\left(XY : R | Z\mathcal{E}\right),$$

where the infimum is taken over all finite random variables R in all extensions of the probability space satisfying the following

1. $X - R - Y$
2. Given XY, we have that $Z\mathcal{E}$ and R are independent. In other words, for all x, y, r, z, it holds that $p(r, z, \mathcal{E}|xy) = p(r|xy)p(z, \mathcal{E}|xy)$.

It is easily seen that if R satisfies $X - R - Y$, then $|\mathsf{supp}(R)| \geq \mathbb{C}\left(X : Y | Z\mathcal{E}\right)$ for any Z and $\mathcal{E}$ satisfying the above condition, and so, we also get a lower bound on the non-negative rank of A. Having the flexibility to condition on arbitrary random variable Z and an arbitrary event $\mathcal{E}$ satisfying the above conditions is often helpful in analysis as it allows us to focus on only a part of the matrix.

Proposition 6.5 was first used by Braverman and Moitra [BM13] to prove a tight lower bound of $2^{\Omega(\rho n)}$ for unique disjointness matrices with arbitrary parameter $\rho \in [0, 1]$. Later, Braun and Pokutta [BP16] showed that the above definition of conditional common information allows a general framework to prove lower bounds on non-negative rank and they used it reprove tight lower bounds for unique disjointness matrices with arbitrary parameters ρ. The same framework was used by Braun and Pokutta [BP15] to prove a lower bound on the slack matrix that comes from approximate extended formulations for the Matching Polytope. We will also use Proposition 6.5 in Chapter 7 to prove a tight lower bound on approximate extended formulations for the Matching Polytope.

6.4.3 Hyperplane Separation Bound

Given the non-negative matrix $A \in \mathbb{R}_{\geq 0}^{\mathcal{X} \times \mathcal{Y}}$ whose non-negative rank we want to bound, let $W \in \mathbb{R}^{\mathcal{X} \times \mathcal{Y}}$ be an arbitrary matrix of weights. For a matrix M, let $\|M\|_\infty = \max_{xy} |M_{xy}|$ denote the infinity norm of M and let $\langle W, M \rangle = \sum_{xy} W_{xy} M_{xy}$ denote the Frobenius inner product between the matrices W and M. Then, the hyperplane separation bound, due to Fiorini, is captured by the following proposition.

Proposition 6.6. *Given* $W \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$, *let* $\alpha = \max_R\{\langle W, R\rangle \mid R \in \{0,1\}^{\mathcal{X}\times\mathcal{Y}}, \mathsf{rk}(R) \leq 1\}$ *and* $\beta = \max_{(x,y)\in\mathsf{supp}(W)} |A_{xy}|$. *Then, we have*

$$\mathsf{rk}_+(A) \geq \frac{\langle W, A\rangle}{\alpha\beta} \geq \frac{\langle W, A\rangle}{\alpha\|A\|_\infty}.$$

Proof. We know that when maximizing a linear function over a polytope, the maximum is always attained at a vertex of the polytope. Using this fact and standard linear optimization arguments (see Lemma 5 in [Rot17]) one can prove that

$$\alpha = \max_R\{\langle W, R\rangle \mid R \in \{0,1\}^{\mathcal{X}\times\mathcal{Y}}, \mathsf{rk}(R) \leq 1\} = \max_R\{\langle W, R\rangle \mid R \in [0,1]^{\mathcal{X}\times\mathcal{Y}}, \mathsf{rk}(R) \leq 1\}.$$

The above implies that for any *fractional rank* 1 matrix $R \in [0,1]^{\mathcal{X}\times\mathcal{Y}}$, we have $\langle W, R\rangle \leq \alpha$.

Write $A = \sum_{i=1}^r R_i$ where R_i are rank 1 non-negative matrices for $i \in [r]$. Let $\widetilde{A} \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ denote the restriction of the matrix A to $\mathsf{supp}(W)$. In other words, $\widetilde{A}_{xy} = A_{xy}$ when $xy \in \mathsf{supp}(W)$ and $\widetilde{A}_{xy} = 0$ otherwise. Similarly, define $\widetilde{R}_i$ for $i \in [r]$ as the restriction of R_i to $\mathsf{supp}(W)$. Note that by definition, $\langle W, A\rangle = \left\langle W, \widetilde{A}\right\rangle$ and $\langle W, R_i\rangle = \left\langle W, \widetilde{R}_i\right\rangle$. Also, notice that for each $i \in [r]$, the matrices $\frac{\widetilde{R}_i}{\|\widetilde{R}_i\|_\infty}$ have entries in $[0,1]$.

Now, we can write

$$\begin{aligned}\langle W, A\rangle = \left\langle W, \widetilde{A}\right\rangle &= \sum_{i=1}^r \langle W, R_i\rangle = \sum_{i=1}^r \left\langle W, \widetilde{R}_i\right\rangle \\ &= \sum_{i=1}^r \|\widetilde{R}_i\|_\infty \left\langle W, \frac{\widetilde{R}_i}{\|\widetilde{R}_i\|_\infty}\right\rangle \leq \alpha \sum_{i=1}^r \|\widetilde{R}_i\|_\infty \leq \alpha \cdot r \cdot \beta.\end{aligned}$$

The last inequality above follows since $\|\widetilde{R}_i\|_\infty \leq \beta$ for each $i \in [r]$, while the second inequality in the statement of the lemma follows simply because $\beta \leq \|A\|_\infty$. □

There is nothing special about the infinity norm in the above proposition and one may replace it with other norms (see [FP16]). The above proposition was used by Rothvoß [Rot17] to prove a lower bound on the non-negative rank of the slack matrix S for the Matching Polytope. Rothvoß came up with a weight matrix W such that the inner product $\langle W, S\rangle = 1$ while the inner product of the weight matrix with every rank one rectangle is at most $2^{-O(n)}$. This proved a $2^{\Omega(n)}$ lower bound on

the non-negative rank of S and hence on the extension complexity of the Matching Polytope. Later, Braun and Pokutta [BP15] proved the same lower bound using the common information method mentioned above.

In all the examples we know of so far, the hyperplane separation bound and the common information method yield bounds that are polynomially apart. Based on this, Pokutta [Pok15] asked if one can prove that the hyperplane separation bound and the common information methods are polynomially related. In Chapter 7, we give a candidate that might give an exponential separation between these two techniques.

6.4.4 Approximation by Conical Juntas

A different technique was introduced by Chan *et al.* [CLRS13] and further developed by Kothari *et al.* [KMR17] for lower bounding the non-negative rank of a special class of slack matrices. Let $f : \{\pm 1\}^n \to \mathbb{R}_{\geq 0}$ be a non-negative real function and let $\mathcal{X} = \mathcal{Y} = \left(\{0,1\}^{n^{100}}\right)^n$. Then, the slack matrices $M = M_f \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ used in [KMR17] are of the following form:

$$M_{xy} = f\left(\langle x_1, y_1\rangle, \cdots, \langle x_n, y_n\rangle\right), \tag{6.4.1}$$

where $x_i, y_i \in \{0,1\}^{n^{100}}$ for $i \in [n]$ and $\langle x, y\rangle = (-1)^{\sum_i x_i y_i \bmod 2}$ is a variant of the inner product function.

Such slack matrices arise in the context of proving lower bounds on the size of linear programs that approximate Maximum Constraint Satisfaction problems. Towards proving lower bounds on the non-negative rank of such matrices, let us call a non-negative real function $h(z_1, \cdots, z_n)$ a *conical d-junta* if h can be written as a conic (non-negative) combination of functions that depend only on d variables. Let $\deg_+(f)$ denote the minimum d such that f is a d-conical junta. For matrices of the form (6.4.1), it is easy to see that if $\deg_+(f) = d$, then $\mathsf{rk}_+(M_f) = n^{O(d)}$. Kothari *et al.* [KMR17] showed that in fact, $\mathsf{rk}_+(M_f) = n^{\Omega(\deg_+(f+\frac{1}{n}))}$. Note that $\deg_+(f+\eta) \leq \deg_+(f)$ for all $\eta > 0$ so the above is not necessarily a converse, but it is often sufficient for applications.

For the MAXCUT problem discussed before, to prove that linear programs of the form (6.3.2) cannot achieve better than $2+\varepsilon$ approximation for any constant $\varepsilon > 0$, the slack matrix M_f corre-

sponds to an f that has $\deg_+(f + \frac{1}{n})$ at least n^γ for some constant $\gamma = \gamma(\varepsilon)$. Hence, it translates into a 2^{n^η} lower bound on the size of linear programs for some constant $\eta = \eta(\varepsilon)$. In fact, the above statement follows from a more general observation of Chan *et al.* [CLRS13] who showed that in a generic way, known Sherali-Adams lower bound imply lower bounds for $\deg_+(f + \frac{1}{n})$ for functions f that corresponds to Maximum Constraint Satisfaction problems. This also yields lower bounds for other constraint satisfaction problems like MAX-3-SAT.

7 Inapproximability of the Matching Polytope by Small Linear Programs

In this chapter, we will prove tight lower bounds for approximate extended formulations for the Matching Polytope. Recall that the Matching Polytope $P_{MAT}(n) \subseteq \mathbb{R}^{\binom{[n]}{2}}$ is defined to be the convex hull of all matchings in the complete graph on the vertex set $[n]$:

$$P_{MAT}(n) = \text{conv}\left\{ \mathbb{1}_M \in \mathbb{R}^{\binom{[n]}{2}} \;\middle|\; M \subseteq \binom{[n]}{2} \text{ is a matching in } K_n \right\}$$

We will just write P_{MAT} when n is clear from the context. Also recall that by a result of Edmonds [Edm65], the facets of P_{MAT} are given by the following degree constraints and *odd set* inequalities:

$$\begin{aligned} x &\geq 0; \\ \sum_{\substack{e \text{ incident} \\ \text{on } i}} x_e &\leq 1, \forall i \in [n]; \\ \sum_{\substack{e \text{ inside} \\ U}} x_e &\leq \frac{|U|-1}{2}, \forall U \subseteq [n], |U| \text{ odd}. \end{aligned}$$

Even though the description of the Matching Polytope has $2^{\Omega(n)}$ facet-defining inequalities, one can still optimize any linear function in polynomial time over the Matching Polytope using the algorithm of Edmonds [Edm65] or the polynomial time separation oracle given by Padberg and Rao [PR82]. The question of whether one could optimize over the Matching Polytope with a small extended formulation remained open until recently when, in a breakthrough result, Rothvoß [Rot17] proved that the extension complexity of P_{MAT} is indeed $2^{\Omega(n)}$.

The central question that we want to answer here concerns whether the Matching Polytope can be approximated by a small extended formulation. Let us say that polytope K is a $(1+\varepsilon)$ *approximation of a monotone down-closed* polytope P if $P \subseteq K \subseteq (1+\varepsilon)P$. As discussed in Section 6.1.1, for monotone down-closed polytopes P, the above is equivalent to requiring that for any non-negative vector w, the following holds

$$\max\{w^T x \mid x \in P\} \leq \max\{w^T x \mid x \in K\} \leq (1+\varepsilon)\max\{w^T x \mid x \in P\}.$$

Let us recall that a $(1+\varepsilon)$-approximate extended formulation for a monotone down-closed polytope P is a higher dimensional lift of a polytope K such that $P \subseteq K \subseteq (1+\varepsilon)P$.

The Matching Polytope P_{MAT} is monotone down-closed and it is well-known (see Section 7.3.1 for a proof) that for any $\frac{2}{n} \leq \varepsilon \leq 1$, the following polytope Q_ε with at most $\binom{n}{1+1/\varepsilon} + O(n^2)$ facets approximates the Matching Polytope up to a factor of $1+\varepsilon$ (note that for $\varepsilon \leq \frac{1}{n-1}$, Q_ε is the Matching Polytope itself):

$$Q_\varepsilon(n) = \left\{ x \in \mathbb{R}^{\binom{[n]}{2}} \;\middle|\; \sum_{\substack{e \text{ incident} \\ \text{on } i}} x_e \leq 1 \;\forall i \in [n];\; x \geq 0; \right. \\ \left. \sum_{\substack{e \text{ inside} \\ U}} x_e \leq \frac{|U|-1}{2} \;\forall U \subseteq [n], |U| \text{ odd},\; |U| \leq \frac{1+\varepsilon}{\varepsilon} \right\} \tag{7.0.1}$$

This gives us a *polynomial type approximation scheme* (PTAS) style linear program (size at most n^c for a constant $c = c(\varepsilon)$ depending on ε) that approximates the Matching Polytope. Building on the ideas of Rothvoß [Rot17], Braun and Pokutta [BP15] showed that any extended formulation that approximates the Matching Polytope up to a factor of $1 + O\left(\frac{1}{n}\right)$ has size $2^{\Omega(n)}$ ruling out a *fully polynomial type approximation scheme* (FPTAS) type extended formulation (size polynomial in both n and $\frac{1}{\varepsilon}$) for matching. Rothvoß [Rot17] observed that this already implies that for any $\frac{2}{n} \leq \varepsilon \leq 1$, any $(1+\varepsilon)$ approximating extended formulation must have $2^{\Omega(1/\varepsilon)}$ size.

Theorem 7.1 ([Rot17, BP15]). *For any $\frac{2}{n} \leq \varepsilon \leq 1$ and any polytope $P_{MAT} \subseteq K \subseteq (1+\varepsilon)P_{MAT}$,*

$$\mathsf{xc}(K) \geq 2^{\Omega(1/\varepsilon)}.$$

Note that the above implies a tight lower bound of $2^{\Omega(n)}$ for $\varepsilon \leq \frac{2}{n}$ but leaves a gap between the upper and lower bounds. Moreover, the gap gets larger as ε increases and in particular when $\varepsilon = \Omega(1)$ we do not even have non-trivial lower bounds.

The above theorem is proven using the connection between extension complexity and non-negative rank of slack matrices that was established in the work of Yannakakis [Yan91] and subsequently extended by Pashkovich [Pas12] and Braun, Fiorini, Pokutta and Steurer [BFPS15b] to handle approximations of polytopes (recall Section 6.2). The lower bound on extension complexity above then follows from a lower bound on the non-negative rank of the slack matrix associated with the Matching Polytope.

A matrix that is closely related to the matching slack matrix is the unique disjointness matrix. Denoting by $\mathcal{Y}$ the collection of all subsets of $[n]$ and we say that a non-negative matrix $A \in \mathbb{R}^{\mathcal{Y}\times\mathcal{Y}}$ is a unique disjointness matrix with parameter $\rho \in [0,1]$ if

$$A_{xy} = \begin{cases} 1, \text{ if } |x \cap y| = 0 \text{ and} \\ \leq 1-\rho, \text{ if } |x \cap y| = 1. \end{cases} \tag{7.0.2}$$

As previously discussed in Section 6.3, many recent lower bounds in extended formulations follow from a lower bound on the non-negative rank of such matrices. Starting with the breakthrough work of Fiorini, Massar, Pokutta, Tiwary and de Wolf [FMP+15], a sequence of works [BFPS15b, BM13, BP16] proved that the non-negative rank of any unique disjointness matrix is $2^{\Omega(\rho n)}$.

For this work, the matrix that will be of relevance is the lopsided version of the unique disjointness matrix where the rows are indexed by k-subsets of $[n]$ where $k \leq \frac{n}{2}$, while the columns are indexed by all subsets of $[n]$. This is useful for us since it turns out that to prove an extension complexity lower bound for any $(1+\varepsilon)$-approximation for the Matching Polytope it suffices to prove a lower bound on similar lopsided slack matrices. The rows of any such slack matrix (an example is exhibited by the slack matrix for the polytope Q_ε) are indexed by odd sets of size at most $O(\frac{1}{\varepsilon})$ and the columns are indexed by all possible matchings of which there are exponentially many in n.

From the known lower bounds for the unique disjointness matrix discussed above, one could infer that the non-negative rank of the lopsided unique disjointness matrix must be $2^{\Omega(\rho k)}$. One could

however, hope to improve this bound to $\approx \binom{n}{k}$ (which is much larger than $2^{\Omega(k)}$ when k is small) at least for the case $\rho = 1$ by making use of the lopsided structure. For the case $\rho = 1$, Lee, Raghavendra and Steurer [LRS15] proved that the non-negative rank of such lopsided unique disjointness matrices is $\left(\frac{n}{k^3 \log n}\right)^{\Omega(k)}$. However, it is unclear whether their proof strategy can be extended to handle the case of the Matching Polytope. On the other hand, lopsided versions of disjointness have previously been studied in the setting of communication complexity and tight lower bounds [MNSW98, AIP06, Pat11, NR15] were proven for them by exploiting the lopsided structure.

Our Results

In this chapter, we show that the simple upper bound exhibited by the polytope Q_ε defined above is tight in the sense that any extended formulation of a polytope that $(1+\varepsilon)$ approximates P_{MAT} must (roughly) have as many facets as Q_ε.

Theorem 7.2. *For any $\frac{2}{n} \leq \varepsilon \leq 1$ and any polytope $P_{MAT} \subseteq K \subseteq (1+\varepsilon)P_{MAT}$, it holds that*

$$\mathsf{xc}(K) \geq \binom{n}{\alpha/\varepsilon},$$

where $0 < \alpha < 1$ is an absolute constant.

Note that the above theorem also covers the case for which tight bounds were previously known: when $\varepsilon \leq \frac{2}{n}$, we get an asymptotically tight lower bound of $\binom{n}{\alpha n/2} = 2^{\Omega(n)}$.

We also prove a tight lower bound on the non-negative rank of the lopsided unique disjointness matrix. Let $k \leq \frac{n}{2}$ and $\mathcal{X} = \binom{[n]}{k}$ be the collection of k-subsets of $[n]$ and let $\mathcal{Y}$ be the collection of all subsets of $[n]$. For a parameter $\rho \in [0,1]$, let $A \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ be any non-negative matrix satisfying

$$A_{xy} = \begin{cases} 1, \text{ if } |x \cap y| = 0 \text{ and} \\ \leq 1-\rho, \text{ if } |x \cap y| = 1. \end{cases} \tag{7.0.3}$$

Then denoting by $\mathsf{rk}_+(A)$ the non-negative rank of A, we prove that:

Theorem 7.3. *For any non-negative matrix satisfying (7.0.3) and any $\frac{3\cdot 1000^8}{\log(n/k)} \leq \rho^8 \leq 1$, it holds that $\mathsf{rk}_+(A) \geq \binom{n}{\alpha\rho^8 k}$ with $0 < \alpha < 1$ an absolute constant.*

Upto a polynomial dependence in ρ, the above theorem proves a tight bound for all parameters $\rho \in [0,1]$ improving upon the result of [LRS15] who only studied the case of $\rho = 1$ and proved a slightly weaker bound of $\left(\frac{n}{k^3 \log n}\right)^{\Omega(k)}$ for this case.

From known results discussed in Section 6.3.1 (also see Section 7.2), proving a lower bound on the non-negative rank of any lopsided unique disjointness matrix also gives a lower bound on the extension complexity of a lopsided version of the correlation polytope. Formally, we have the following result for the polytope $P_{LCORR} \in \mathbb{R}^{n \times n}$ defined as

$$P_{LCORR} = \mathsf{conv}\left\{ bb^T \;\middle|\; b \in \{0,1\}^n, \sum_{i=1}^{n} b_i \leq k \right\}.$$

Corollary 7.4. *Let* $1 \leq \sigma^8 \leq \frac{1}{3 \cdot 1000^8} \log\left(\frac{n}{k}\right)$. *For any polytope* $P_{LCORR}(n) \subseteq K \subseteq \sigma P_{LCORR}(n)$, *it holds that* $\mathsf{xc}(K) \geq \binom{n}{\alpha k/\sigma^8}$ *with* $0 < \alpha < 1$ *an absolute constant.*

Organization

In Section 7.1 we give the high-level intuition for our results before we delve into technical details. Section 7.2 discusses the slack matrix for the Matching and Lopsided Correlation polytopes. Section 7.3 contains some preliminaries. In Section 7.4 we prove the lower bounds on the non-negative rank of the lopsided unique disjointness matrix and the matching slack matrix. Section 7.5 proves a limitation of our technique.

7.1 Overview of Techniques

Our proof is based on the *common information* approach described in Section 6.4.2. In this approach, we view a non-negative matrix L indexed by x and y as a probability distribution $p(x,y)$ by normalizing by the total weight of the matrix. Let X and Y be sampled from $p(X,Y)$. If $\mathsf{rk}_+(L) = r$, then the corresponding distribution $p(X,Y)$ is a convex combination of r product distributions given by the non-negative rank one factors. Viewed this way, a non-negative factorization of L gives us a random variable R with support of size $\log \mathsf{rk}_+(L)$ that then breaks the dependency between X and Y.

In other words, X and Y are independent if we know the value of R (see Section 7.3) or formally, X, R, Y form a Markov chain (denoted by $X - R - Y$).

7.1.1 Lopsided Unique Disjointness

To give intuition behind the proof of Theorem 7.3, let us take a look at what the corresponding distribution looks like when the matrix we start with is the lopsided unique disjointness matrix A given by (7.0.3). Let X' and Y' be the random variables sampled from the distribution obtained by normalizing A. Then X' is a random subset of $[n]$ of size at most k and Y' is an arbitrary random subset of $[n]$. It turns out that by using a direct-sum argument (dividing the universe $[n]$ into blocks of size $[\frac{n}{k}]$) and appropriate conditioning we can reduce our question to a problem about random subsets X and Y of the universe $[\frac{n}{k}]$.

In the reduced problem, the set $X \subseteq [\frac{n}{k}]$ is a random subset of size 1 and $Y \subseteq [\frac{n}{k}]$ is an arbitrary random set while the probability that $X \cap Y = \varnothing$ is at least $\frac{1}{2} + \Omega(\rho)$. We will argue that if the non-negative rank was in fact small, then the probability that $X \cap Y = \varnothing$ cannot be much larger than $\frac{1}{2}$, deriving a contradiction.

In the reduced problem, the entropies of X and Y are given by $\mathbf{H}(X|X \cap Y = \varnothing) = \log\left(\frac{n}{k}\right) - O(1)$ and $\mathbf{H}(Y|X \cap Y = \varnothing) = \frac{n}{k} - O(1)$. If $\log \mathsf{rk}_+(A) \ll \gamma^8 k \log\left(\frac{n}{k}\right)$ for a small constant γ, then it turns out that we get a random variable R such that $X - R - Y$ and even after conditioning on R the entropies remain large:

$$\mathbf{H}(X|R, X \cap Y = \varnothing) \geq \log\left(\frac{n}{k}\right) - \gamma^8 \log\left(\frac{n}{k}\right) - O(1), \text{ and },$$
$$\mathbf{H}(Y|R, X \cap Y = \varnothing) \geq \frac{n}{k} - \gamma^8 \log\left(\frac{n}{k}\right) - O(1).$$

A random set of size 1 from $[\frac{n}{k}]$ has entropy $\log(\frac{n}{k})$ and a uniformly random set from $[\frac{n}{k}]$ has entropy $\frac{n}{k}$. So, the above expressions characterize how far are X and Y from the corresponding uniform distributions, in terms of entropy. We remark that the conditioning on $X \cap Y = \varnothing$ is needed to carry out the direct-sum argument and is quite essential.

To prove a non-negative rank lower bound for lopsided unique disjointness, we exploit the lopsided structure to prove the following key technical lemma which intuitively says that for most values

of R, the probability that the event $X \cap Y = \varnothing$ happens conditioned on R is smaller than $\frac{1}{2} + \gamma$. In the following lemma, we view the set X as an element of $\binom{[n]}{k}$ while we view $Y \in \{0,1\}^{\frac{n}{k}}$ as an indicator vector for a subset of $\binom{[n]}{k}$, so $X \cap Y = \varnothing$ is equivalent to the event that $Y_X = 0$.

Lemma 7.5. *Let m be a large enough integer, $X \in [m], Y \in \{0,1\}^m$ and R be random variables with distribution $p(XYR)$ such that $X - R - Y$. For any γ satisfying $\frac{3}{\log m} \leq \gamma^8 \leq \frac{1}{2^{64}}$ define $\mathcal{B} = \{(x,r) | p(Y_x = 0|r) \geq \frac{1+\gamma}{2}\}$. If*

$$\mathbf{H}(X|R, Y_X = 0) \geq (1-\gamma^8)\log m - 3 \text{ and } \mathbf{H}(Y|R, Y_X = 0) \geq m - \gamma^8 \log m - 3,$$

then, $p((X,R) \in \mathcal{B}) \leq 64\gamma$.

With a little work, one can conclude from the above lemma that if the non-negative rank of A is small, then the probability of the event $X \cap Y = \varnothing$ is at most $\frac{1}{2} + O(\gamma)$. Choosing γ to be sufficiently small, we can ensure that this probability is much smaller than $\frac{1}{2} + \Omega(\rho)$, and so we derive a contradiction to our initial assumption, that the non-negative rank of the lopsided unique disjointness matrix is small.

To understand Lemma 7.5 in more detail, it is worthwhile to consider some simple examples. If it was the case that $\mathbf{H}(X|R, Y_X = 0) \geq \log m - \alpha$ and $\mathbf{H}(Y|R, Y_X = 0) \geq m - \alpha$ where $\alpha \ll 1$, then using Pinsker's inequality, conditioned on the event $Y_X = 0$ and most values of R, X and Y would be close to uniform in statistical distance. Let us ignore the conditioning on the event $Y_X = 0$ for now. Then, closeness to the uniform distribution in statistical distance implies that the probability that x, r is such that $p(Y_x = 0|r)$ is significantly larger than $\frac{1}{2}$ is small. It turns out that conditioning on the event $Y_X = 0$ biases the probability space in a way so that we get the same statement. Lower bounds on the non-negative rank of the standard unique disjointness matrix (rows and columns indexed by all possible subsets of $[n]$) essentially follow from a variation of this argument since in those cases the entropy loss is small enough so that we can say that the distributions of X and Y (conditioned on R and the event $Y_X = 0$) are close to uniform in statistical distance.

In our case, however, the entropy loss is large enough that the distributions are quite far from uniform in statistical distance. Let us try to construct an example where the entropy loss is larger. Let

S be a random subset of $[m]$ of size $m^{1-\alpha}$ and $T \subseteq S$ be a random subset of size $\alpha \log m$. X will be a uniform index chosen from the subset S while the string Y is chosen uniformly conditioned on the event that the $Y_i = 0$ for every $i \in T$. In this case, $R = ST$ is a random variable that breaks the dependency between X and Y. Also, we have that $\mathbf{H}(X|R, Y_X = 0) = (1-\alpha)\log m$ and $\mathbf{H}(Y|R, Y_X = 0) = m - \alpha \log m$, so the entropy loss is of the same order as in the assumptions of Lemma 7.5. Here, the event that x, r is such that $p(Y_x = 0|r)$ is significantly larger than $\frac{1}{2}$ occurs only when $x \in T$, and hence, the total probability measure of such x, r is at most $|T|/|S| \leq \frac{\alpha \log m}{m^{1-\alpha}}$.

Generalizing the intuition gained from the example given above, the proof of Lemma 7.5 proceeds by showing that if the measure of x, r such that $p(Y_x = 0|r) \geq \frac{1+\gamma}{2}$ is large, and the entropy $\mathbf{H}(X|R, Y_X = 0)$ is large, then for most values of R, there is a large set of coordinates of Y that must be very biased given R, and hence the entropy $\mathbf{H}(Y|R, Y_X = 0)$ must be small. This intuition is borrowed from the lower bounds on lopsided disjointness in communication complexity [Pat11, NR15], but since the setting of non-negative rank is different, the technical details involved for converting this intuition into proof are more challenging here.

Before moving on, we stress two key points about Lemma 7.5. Firstly, given the assumptions on entropy here, one might hope to say that the probability that x, r is such that $p(Y_x = 0|R = r) \leq \frac{1-\gamma}{2}$ must also be small. However, this is not true — the lemma is one-sided, as the following example shows: Let R be a random subset of $[m]$ of size $m^{1-\alpha}$. X will be a uniform index chosen from the subset R while the string Y is chosen as follows: with probability γ, the string Y is a uniform n-bit string conditioned on the event that the $Y_i = 1$ for every $i \in R$, and with probability $1-\gamma$, the string Y is a uniform n-bit string. In this case, R is a random variable that breaks the dependency between X and Y. Also, we have that $\mathbf{H}(X|R, Y_X = 0) = (1-\alpha)\log m$ and $\mathbf{H}(Y|R, Y_X = 0) = m - O(1)$ (the conditioning on $Y_X = 0$ is crucial here), so the entropies satisfy the assumptions of Lemma 7.5. Letting $\mathcal{E}$ be the event that x, r is such that $p(Y_x = 0|r) \leq \frac{1-\gamma}{2}$, it is easy to see that $p(\mathcal{E}) = 1$. The above example also shows the importance of conditioning on the event $Y_X = 0$: without conditioning on this event, the entropy $H(Y|R)$ is, in fact, quite small.

Secondly, the lopsided structure is crucial in Lemma 7.5. Such a lemma is not true if one considers the case where Y is also a random subset of $[m]$ of size 1 satisfying $\mathbf{H}(Y|R, X \cap Y = \varnothing) \geq (1 -$

$\gamma^8) \log m - O(1)$ for a constant $\gamma > 0$. So, even though one could hope that the non-negative rank of the small set unique disjointness matrix (where we restrict both rows and columns to be indexed by sets of size at most k) is $\approx \binom{n}{k}$, a common information based approach, as used here, will be unable to prove this (see Section 7.5 for more details).

7.1.2 Matching Slack Matrix

Recalling the connection between extension complexity and non-negative rank of slack matrices, it turns out that to prove Theorem 7.2 it is sufficient to prove a lower bound on the non-negative rank of an appropriate slack matrix associated with the Matching Polytope (see Section 7.2). For the sake of providing intuition, we will work with a slightly simpler slack matrix S in this section. The rows of this slack matrix are indexed by odd cuts (subsets) of $[2n+6]$ of size at most $1/\varepsilon$ and the columns are indexed by perfect matchings in the complete graph on the vertex set $[2n+6]$. The entry corresponding to cut u and perfect matching m is $S_{um} = |\delta(u) \cap m| - 1$ where $\delta(u)$ is the set of edges of the complete graph on $[2n+6]$ crossing u (edges with exactly one end point in u). The true slack matrix whose non-negative rank we need to bound to prove Theorem 7.2 is a noisy version of this slack matrix where a small constant $\beta > 0$ is added to every entry.

Using a direct-sum argument with appropriate conditioning (dividing vertices into $1/\varepsilon$ chunks of size $O(\varepsilon n)$ each), we can reduce our problem to a question about a random odd cut U and a random perfect matching M in a graph with $2t+6$ vertices where $t := \varepsilon n$. Furthermore, the non-negative rank decomposition gives a random variable R such that $U - R - M$ where the size of support of R is $\log \mathsf{rk}_+(S)$. Denoting by $q(U, M, R)$ the distribution of U, M, R, it turns out that the probability $q(U = u, M = m)$ is proportional to $|\delta(u) \cap m| - 1$. In particular, if only one edge of the matching m crosses the cut u, then it has probability zero under the distribution $q(U, M, R)$.

To describe the high-level idea of the proof we need some notation. We first fix an arbitrary perfect matching $\mathcal{A}$ in the graph and an arbitrary cut $\mathcal{Z}$ that cuts all edges of $\mathcal{A}$. We pick a uniformly random partition $B = (B_0, B_1, \ldots, B_t)$ of the set $\mathcal{Z}$ such that $|B_0| = 3$ and $|B_i| = 2$ for each $i \in [t]$. We call a cut U *consistent* if $U = B_0 \cup B_j$ for some $j \in [t]$ and note that the size of any such cut is always 5.

We call a perfect matching M *consistent* if M always includes the edges of $\mathcal{A}$ touching B_0 and inside the other blocks B_j for $j \in [t]$, either M includes the edges of $\mathcal{A}$ or M matches the block B_j to itself and the neighbors of B_j under $\mathcal{A}$ to itself (see Figure 7.1.1).

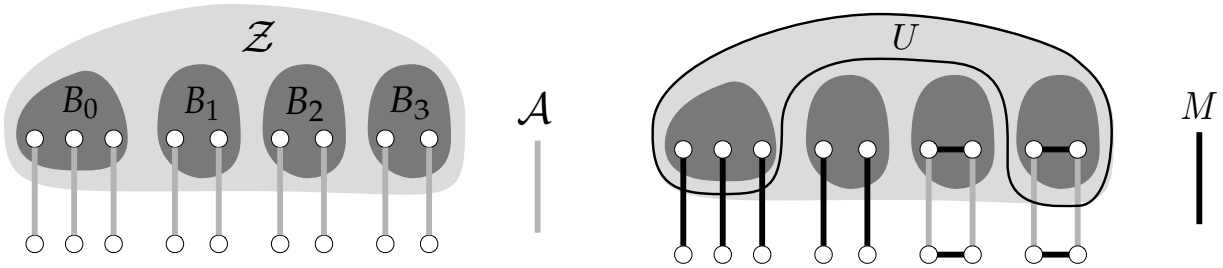


Figure 7.1.1: An example of $\mathcal{A}, \mathcal{Z}, B$ with $t = 3$ (left) and U, M conditioned on $\mathcal{D}$ (right). Note that U and M are both consistent.

Let $\mathcal{E}$ be the event that U and M are both consistent. Note that given the partition B, one can check whether the cut U is consistent without knowing what the matching M is and vice-versa. Hence, as $U - R - M$, it follows that even conditioned on the event $\mathcal{E}$, U and M are still independent given R and B. Furthermore, when the event $\mathcal{E}$ occurs then either $|\delta(U) \cap M| = 3$ (when U does not cut the edges of M apart from those incident on B_0) or $|\delta(U) \cap M| = 5$ otherwise.

Let $\mathcal{D} \subseteq \mathcal{E}$ denote the event that U and M are consistent and U does not cut the edges of M apart from those incident on B_0. Comparing this setup to the case of lopsided unique disjointness, one can see that apart from choosing B_0, U corresponds to picking a set of size one among the t blocks $(B_1, \ldots, B_t)$, and apart from the edges touching B_0, M corresponds to picking a subset of the same t blocks by considering the blocks where M crosses B_j to be in the set. The event $\mathcal{D}$ then exactly corresponds to the event that these sets are disjoint. In the "disjoint" case, there are exactly 3 edges of M crossing the cut U whereas in the "intersecting" case the number of edges of M crossing U is exactly 5.

As the probability under $q(U, M, R)$ is proportional to $|\delta(U) \cap M| - 1$, it is not too hard to see that the probability of the event $q(\mathcal{D}|\mathcal{E}) = \frac{(3-1)}{(3-1)+(5-1)} = \frac{1}{3}$. Furthermore, the entropies $\mathbf{H}(U|B, \mathcal{D}) = \log t - O(1)$ and $\mathbf{H}(M|B, \mathcal{D}) = t - O(1)$. If $\log \mathsf{rk}_+(S) \ll \gamma^8 \cdot \frac{n}{t} \log t$ for a

small constant γ, then even after conditioning on R the entropies remain large: $\mathbf{H}(U|RB,\mathcal{D}) \geq \log t - \gamma^8 \log t - O(1)$ and $\mathbf{H}(M|RB,\mathcal{D}) \geq t - \gamma^8 \log t - O(1)$.

We want to proceed similarly to the case of lopsided unique disjointness: We want to conclude that the assumptions on entropy imply that $q(\mathcal{D}|\mathcal{E})$ must be much smaller than $\frac{1}{3}$. In the case of lopsided disjointness, it was enough for us to use Lemma 7.5 and bound the contribution to $p(\mathcal{D}|\mathcal{E}, R = r, B = b)$ by $\frac{1}{2} + \gamma$ for most r, b. But now as we want to prove that the probability is smaller than $\frac{1}{3}$, we need to exploit the combinatorial structure of the Matching Polytope.

This is where the random partition B, which is the key new idea introduced by Rothvoß [Rot17], is useful. We are going to argue that only certain kinds of partitions can contribute to the probability of the event $\mathcal{D}$; otherwise, there is a non-zero probability of sampling a cut U and a matching M that satisfies $|\delta(U) \cap M| = 1$ and any such pair has probability zero in the distribution $q(U, M, R)$, since the corresponding slack matrix entry in S is zero. Then, averaging over all the choices of the random partition B, we can show the total contribution of these random partitions to $q(\mathcal{D}|\mathcal{E})$ is indeed less than $\frac{1}{3}$.

Let us make some simplifying assumptions first. Define M_j to be the edges of M corresponding to block B_j. Let us assume that for all values r, b the probability that $M_j = \mathcal{A}_j$ (M crosses B_j) is roughly $\frac{1}{2}$ conditioned on $\mathcal{E}, R = r, B = b$ for each $j \in [t]$. Since conditioned on $\mathcal{E}$ either $M_j = \mathcal{A}_j$ or $M_j \neq \mathcal{A}_j$, this implies that

$$q(M_j \neq \mathcal{A}_j|\mathcal{E}, R = r, B = b) \approx q(M_j = \mathcal{A}_j|\mathcal{E}, R = r, B = b) \approx \frac{1}{2}.$$

Also assume that U is almost uniform among the m possible cuts. Then, as U and M are independent conditioned on $\mathcal{E}, R = r, B = b$, we get that

$$q(U = B_0 \cup B_j, M_j \neq \mathcal{A}_j|\mathcal{E}, R = r, B = b) \approx q(U = B_0 \cup B_j, M_j = \mathcal{A}_j|\mathcal{E}, R = r, B = b).$$

The event on the right hand side above fixes a cut of size 5 in the graph and a matching that crosses on all the edges. Fix the blocks outside this cut arbitrarily. By symmetry all of the $\binom{5}{3}$ ways of splitting this cut into B_0 and B_j are equally likely (see Figure 7.1.2). It turns out that the right hand side above can be non-zero only when the support of B_0 forms a family where every two sets intersect in at least

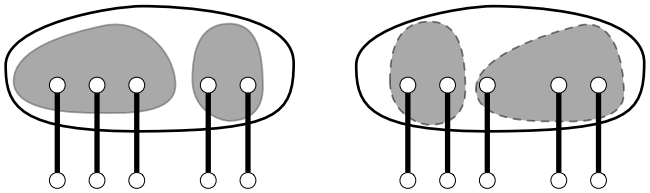


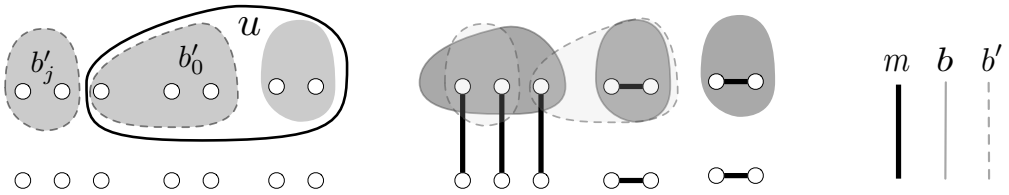
Figure 7.1.2: An example of a cut and matching corresponding to the event $U = B_0 \cup B_j, M_j = \mathcal{A}_j$ and an example of two different splits of the cut into $B_0 \cup B_j$. By symmetry any such partition is equi-probable.

two vertices (note that this determines B_j as the parts outside are already fixed) and hence averaging over b and r, we get that $q(\mathcal{D}|\mathcal{E}) \approx \frac{\Gamma}{10} q(\overline{\mathcal{D}}|\mathcal{E})$ where Γ is a bound on the size of any such family. It turns out that Γ is small enough so that we can conclude $q(\mathcal{D}|\mathcal{E}) < \frac{1}{3}$ and derive a contradiction.

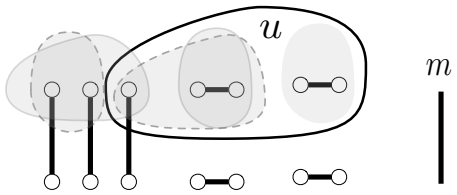
Why must the probability be zero when there are two partitions b and b' (which are same everywhere outside this cut) such that $|b_0 \cap b'_0| = 1$? This is because we can choose an appropriate cut u that is consistent with b' (See Figure 7.1.3(a)) and has non-zero probability $q(U = u|\mathcal{E}, R = r, B = b') > 0$ and an appropriate matching m that is consistent with b (see Figure 7.1.3(a)) and has non-zero probability $q(M = m|\mathcal{E}, R = r, B = b) > 0$. Note that then it must also hold that the cut and the matching has non-zero probability even without conditioning on $\mathcal{E}$: $q(U = u|R = r) > 0$ and $q(M = m|R = r) > 0$. But since U and M are independent given R then such a pair would have non-zero probability even though $|\delta(u) \cap m| = 1$ and any such pair must have zero probability under $q(U, M)$.

If it was the case that $\mathbf{H}(U|B, \mathcal{D}) \geq \log t - \alpha$ and $\mathbf{H}(M|B, \mathcal{D}) \geq t - \alpha$ where $\alpha \ll 1$, then one could work with statistical distance as is done in [Rot17, BP15], but since in our case the entropy loss is much larger, we use Lemma 7.5 in conjunction with the random partition idea.

As mentioned before, to prove Theorem 7.2 we have to work with a noisy version of the slack matrix used above, where the probability of a pair U and M satisfying $|\delta(U) \cap M| = 1$ is not zero, but a small constant. This along with the limitations of Lemma 7.5 and the lopsided structure makes translating this intuition into a formal proof considerably more involved.



(a) An example of a cut u consistent with b' (left) and a matching m consistent with b (right). Note that b and b' agree on all the blocks except b_0 and b_j.



(b) The cut u and matching m satisfy $|\delta(u) \cap m| = 1$.

7.2 Slack Matrices and Non-Negative Rank

Let us take a look at the slack matrix for the Matching and Lopsided Correlation polytopes.

Slack Matrix for the Matching Polytope

Let $0 < \beta < 1$ be a constant which will be determined by the proof. In proving Theorem 7.2, we may assume without loss of generality that $\frac{c'}{n} \leq \varepsilon \leq 1$ for a sufficiently large constant c'. We call a subset of vertices a *cut* and for a cut U, we define $\delta(U)$ to be the set of edges with exactly one endpoint in U and $E(U)$ to be the set of edges inside U.

Consider the following outer polytope $Q'_\varepsilon(n) \subseteq \mathbb{R}^{\binom{[n]}{2}}$ for matching:

$$Q'_\varepsilon(n) = \left\{ x \in \mathbb{R}^{\binom{[n]}{2}} \;\middle|\; \sum_{e \in \delta(\{i\})} x_e \leq 1 \,, \forall i \in [n];\; x \geq 0; \right.$$
$$\left. \sum_{e \in E(U)} x_e \leq \frac{|U| + \beta - 1}{2} \,, \forall U \subseteq [n],\; |U| \text{ odd} \,,\; |U| \leq 1 + \frac{\beta}{\varepsilon}. \right\}$$

For any cut U of size at most $1 + \frac{\beta}{\varepsilon}$, we have $(1+\varepsilon)(|U|-1) \leq |U| + \beta - 1$. Since every $x \in (1+\varepsilon)P_{MAT}(n)$ satisfies

$$x(E(U)) \leq (1+\varepsilon)\frac{|U|-1}{2} \quad \forall U \subseteq [n], |U| \text{ odd},$$

the following proposition holds.

Proposition 7.6. $(1+\varepsilon)P_{MAT}(n) \subseteq Q'_\varepsilon(n)$.

We remind the reader that the polytope $Q_\varepsilon(n)$ defined in (7.0.1) is also contained inside $Q'_\varepsilon(n)$ since $Q_\varepsilon(n) \subseteq (1+\varepsilon)P_{MAT}(n)$.

Using Proposition 7.6, it suffices to lower bound the non-negative rank of the slack matrix $S^{P_{MAT}, Q'_\varepsilon}$ to lower bound the extension complexity of polytopes that approximate the Matching Polytope. Since there are only $O(n)$ degree constraints, let us focus on the submatrix S of $S^{P_{MAT}, Q'_\varepsilon}$ where the entries are given by odd cuts and perfect matchings. Then, the entry corresponding to cut u and perfect matching m is

$$S_{um} = \frac{|u|+\beta-1}{2} - |E(u) \cap m| = \frac{|\delta(u) \cap m| + \beta - 1}{2}, \tag{7.2.1}$$

since for any perfect matching m and odd cut u, $|E(u) \cap m| = \frac{|u| - |\delta(u) \cap m|}{2}$. Note that this matrix has $\binom{n}{\Theta(1/\varepsilon)}$ rows. We prove that

Theorem 7.7. *For the matrix S defined in (7.2.1), it holds that*

$$\mathsf{rk}_+(S) \geq \binom{n}{\alpha/\varepsilon}$$

where $\frac{c'}{n} \leq \varepsilon \leq 1$ for a large enough constant $c' = c'(\beta)$ and $0 < \alpha < 1$ is an absolute constant.

Theorem 6.2 and Theorem 7.7 then give us Theorem 7.2.

Proof of Theorem 7.2. Note that Proposition 7.6 implies that for any polytope K such that $P_{MAT} \subseteq K \subseteq (1+\varepsilon)P_{MAT}$, it holds that $P_{MAT} \subseteq K \subseteq Q'_\varepsilon(n)$. Using Theorem 6.2 and Theorem 7.7, it follows that

$$\mathsf{xc}(K) \geq \mathsf{rk}_+(S) - 1 \geq \binom{n}{\alpha/\varepsilon},$$

for an absolute constant $0 < \alpha < 1$. □

Slack Matrix for the Lopsided Correlation Polytope

Consider the polytope $Q = Q(n) = \{x \in \mathbb{R}^{n\times n} \mid \langle 2\mathsf{diag}(a) - aa^T, x\rangle \leq 1, a \in \{0,1\}^n\}$ where $\mathsf{diag}(a) \in \mathbb{R}^{n\times n}$ denotes the diagonal matrix which has the vector a on the diagonal and is zero otherwise. Recall from Section 6.3.1 that $P_{CORR} \subseteq Q$, and hence, it also follows that $P_{LCORR} \subseteq Q$ since P_{LCORR} is contained inside P_{CORR}. The slack matrix corresponding to the inner polytope P_{LCORR} and outer polytope σQ is the non-negative matrix σA where A is a lopsided unique disjointness matrix satisfying the conditions in (7.0.3) with parameter $\rho = \frac{1}{\sigma}$. Corollary 7.4 then directly follows from Theorem 7.3 and Theorem 6.2.

7.3 Preliminaries

7.3.1 Approximating the Matching Polytope

Below we give a proof of the folklore result which says that the polytope $Q_\varepsilon(n)$ defined in (7.0.1) is a $(1+\varepsilon)$ approximation for the Matching Polytope. Recall that for a cut U, we define $\delta(U)$ to be the set of edges with exactly one end point in U and $E(U)$ to be the set of edges inside U.

Claim 7.8. $P_{MAT}(n) \subseteq Q_\varepsilon(n) \subseteq (1+\varepsilon)P_{MAT}(n)$.

Proof. First inclusion is trivial since any vector $x \in P_{MAT}(n)$ is also in $Q_\varepsilon(n)$. To see the second inclusion, let $x \in Q_\varepsilon(n)$ be an arbitrary vector. We only need to show that for any odd set $U \subseteq [n]$, $|U| > \frac{1+\varepsilon}{\varepsilon}$, $x(E(U)) \leq (1+\varepsilon)\frac{|U|-1}{2}$ since the other inequalities are already satisfied. Note that $x(E(U)) \leq \frac{1}{2}\sum_{v\in U} x(\delta(\{v\})) \leq |U|/2 \leq (1+\varepsilon)\frac{|U|-1}{2}$ where the last inequality holds when $|U| \leq \frac{1+\varepsilon}{\varepsilon}$. □

7.3.2 Information Theory Lemmas

Lemma 7.9. *Let X be a random variable such that $\mathbf{H}(X) \geq \log \ell - a$ where $\ell = |\mathsf{supp}(X)|$ and $a \geq 0$. Define $\mathcal{S} = \{x | p(x) \leq \frac{2^{(a+1)/\gamma}}{\ell}\}$. Then, $p(\mathcal{S}) \geq 1-\gamma$.*

Proof. Set $b = 2^{(a+1)/\gamma}$. Denoting by $\overline{\mathcal{S}}$ the complement of $\mathcal{S}$, we can write

$$\begin{aligned}\mathbf{H}(X) &= \sum_{x\in\mathcal{S}} p(x)\log\left(\frac{1}{p(x)}\right) + \sum_{x\notin\mathcal{S}} p(x)\log\left(\frac{1}{p(x)}\right)\\ &\le \sum_{x\in\mathcal{S}} p(x)\log\left(\frac{1}{p(x)}\right) + p(\overline{\mathcal{S}})\log\left(\frac{\ell}{b}\right)\\ &= p(\mathcal{S})\sum_{x\in\mathcal{S}}\frac{p(x)}{p(\mathcal{S})}\log\left(\frac{1}{p(x)}\right) + p(\overline{\mathcal{S}})\log\left(\frac{\ell}{b}\right)\\ &\le p(\mathcal{S})\log\left(\sum_{x\in\mathcal{S}}\frac{1}{p(\mathcal{S})}\right) + p(\overline{\mathcal{S}})\log\left(\frac{\ell}{b}\right),\end{aligned}$$

where the first inequality follows from the definition of $\mathcal{S}$ and the second inequality follows from concavity of the log function. We can further upper bound

$$\begin{aligned}\mathbf{H}(X) &\le p(\mathcal{S})\log\left(\frac{1}{p(\mathcal{S})}\right) + p(\mathcal{S})\log|\mathcal{S}| + p(\overline{\mathcal{S}})\log\left(\frac{\ell}{b}\right)\\ &\le 1 + p(\mathcal{S})\log\ell + p(\overline{\mathcal{S}})\log\left(\frac{\ell}{b}\right),\end{aligned}$$

where we used that $x\log(1/x) \le 1$ for $0 \le x \le 1$. Since $\mathbf{H}(X) \ge \log\ell - a$, we get that

$$\log\ell - a \le 1 + p(\mathcal{S})\log\ell + (1-p(\mathcal{S}))\log\left(\frac{\ell}{b}\right),$$

which gives that $p(\mathcal{S}) \ge 1 - \frac{a+1}{\log b} = 1-\gamma$. □

Lemma 7.10. *Let $Y \in \{0,1\}^\ell$ and define $\mathsf{bias}_i(Y) := p(Y_i = 0) - p(Y_i = 1)$. If there is a set $\mathcal{S} \subseteq [\ell]$ such that $\mathbb{E}_{i\in\mathcal{S}}[\mathsf{bias}_i(Y)] \ge 2\gamma$ where $\gamma > 0$, then $\mathbf{H}(Y) \le \ell - \gamma^2|\mathcal{S}|$.*

Proof. We may write

$$\mathbb{E}_{i\in\mathcal{S}}[\mathsf{bias}_i(Y)] = \mathbb{E}_{i\in\mathcal{S}}[p(Y_i=0) - (1-p(Y_i=0))] = 2\mathbb{E}_{i\in\mathcal{S}}[p(Y_i=0)] - 1,$$

which by the assumption implies that $\mathbb{E}_{i\in\mathcal{S}}[p(Y_i=0)] \ge \frac{1}{2} + \gamma$.

We can upper bound

$$\mathbb{E}_{i\in\mathcal{S}}[\mathbf{H}(Y_i)] = \mathbb{E}_{i\in\mathcal{S}}[\mathsf{h}(p(Y_i=0))] \le \mathsf{h}(\mathbb{E}_{i\in\mathcal{S}}[p(Y_i=0)]),$$

where the last inequality follows from the concavity of the binary entropy function h. Recall that h is a decreasing function on $\left[\frac{1}{2}, 1\right]$ and Proposition 2.8 implies that $\mathsf{h}(\frac{1}{2}+x) \leq 1 - 2\log e \cdot x^2 \leq 1 - x^2$. It follows that

$$\mathbb{E}_{i\in\mathcal{S}}[\mathbf{H}(Y_i)] \leq \mathsf{h}\left(\frac{1}{2}+\gamma\right) \leq 1-\gamma^2.$$

Denoting by $\overline{\mathcal{S}}$ the complement of $\mathcal{S}$ and applying the chain rule we get:

$$\begin{aligned}\mathbf{H}(Y) &\leq \mathbf{H}(Y_{\overline{\mathcal{S}}}) + \mathbf{H}(Y_{\mathcal{S}}) \leq \mathbf{H}(Y_{\overline{\mathcal{S}}}) + \sum_{i\in\mathcal{S}} \mathbf{H}(Y_i)\\ &\leq (\ell - |\mathcal{S}|) + |\mathcal{S}|(1-\gamma^2) = \ell - \gamma^2|\mathcal{S}|.\end{aligned}$$

□

7.3.3 Intersecting Families

The following lemma will be crucial for the analysis. It is a special case of the Erdős-Ko-Rado Theorem (see [Wil84]) which says that when $n \geq 6$, then the size of any family of $\binom{[n]}{3}$ that intersects in two elements is at most $n-2$ (Lemma 7.11 follows from the case $n = 6$). We give a self-contained proof here.

Lemma 7.11. *Let $\mathfrak{F} \subseteq \binom{[5]}{3}$ be a family of subsets such that any two sets in $\mathfrak{F}$ intersect in two elements. Then, $|\mathfrak{F}| \leq 4$.*

Proof. Take any two sets $\mathcal{F}_1$ and $\mathcal{F}_2$ in the family. Since, $\mathcal{F}_1$ and $\mathcal{F}_2$ intersect in two elements, $|\mathcal{F}_1 \cup \mathcal{F}_2| = 4$. There are two possibilities: if there is another set $\mathcal{F}_3$ in the family that contains the unique element $x \in [5]$ that is not in $\mathcal{F}_1 \cup \mathcal{F}_2$, then $\mathcal{F}_3 = (\mathcal{F}_1 \cap \mathcal{F}_2) \cup \{x\}$ and there can be no other sets in the family which have intersection of size 2 with $\mathcal{F}_1, \mathcal{F}_2$ and $\mathcal{F}_3$ simultaneously; if there is no such set $\mathcal{F}_3$, then $|\cup_{\mathcal{F}\in\mathfrak{F}} \mathcal{F}| \leq 4$, and the size of the family $\mathfrak{F}$ is bounded by $\binom{4}{3} = 4$. □

7.4 Lower Bounds on Non-negative Rank

Let us recall the main technical lemma which we use to derive a lower bound on the non-negative rank of the lopsided unique disjointness matrix as well as the matching slack matrix.

Lemma 7.5. *Let m be a large enough integer, $X \in [m], Y \in \{0,1\}^m$ and R be random variables with distribution $p(XYR)$ such that $X - R - Y$. For any γ satisfying $\frac{3}{\log m} \leq \gamma^8 \leq \frac{1}{2^{64}}$ define $\mathcal{B} = \{(x,r) | p(Y_x = 0|r) \geq \frac{1+\gamma}{2}\}$. If*

$$\mathbf{H}(X|R, Y_X = 0) \geq (1-\gamma^8)\log m - 3 \text{ and } \mathbf{H}(Y|R, Y_X = 0) \geq m - \gamma^8 \log m - 3,$$

then, $p((X,R) \in \mathcal{B}) \leq 64\gamma$.

We first prove the above lemma in Section 7.4.1. In the following sections, Section 7.4.2 and Section 7.4.3, we use Lemma 7.5 to derive non-negative rank lower bounds.

7.4.1 Proof of Main Technical Lemma

In this section, we prove Lemma 7.5. For the sake of contradiction, we assume that $p((X,R) \in \mathcal{B}) \geq 64\gamma$. The proof will proceed by first fixing a value of r, such that the entropies $\mathbf{H}_p(X|r, Y_X = 0)$ and $\mathbf{H}_p(Y|r, Y_X = 0)$ will remain large but the distribution of X and Y conditioned on r will be quite biased (conditioned on r, the probability of the event $Y_X = 0$ will be significantly larger than $\frac{1}{2}$). Then, we will argue that if this was the case, then in fact, the entropy $\mathbf{H}_p(Y|r, Y_X = 0)$ must be much smaller than our assumption.

Let $\mathcal{G}$ denote the set of r such that

(a) $\mathbf{H}_p(X|r, Y_X = 0) \geq (1-\gamma^4)\log m$ and $\mathbf{H}_p(Y|r, Y_X = 0) \geq m - \gamma^4 \log m$.

(b) $p(X \in \mathcal{S}_r | r) \geq 32\gamma$ where $\mathcal{S}_r = \{x | (x,r) \in \mathcal{B}\}$.

We will be able to argue that

Claim 7.12. $p(R \in \mathcal{G} | Y_X = 0) \geq 256\gamma^2$.

In particular, this means that the set $\mathcal{G}$ is not empty. For the rest of the proof, we will fix some $r \in \mathcal{G}$ and work with the distribution $q(XY) = p(XY|r)$ which is product. Consider the random variable $\mathsf{Bias} = \mathbb{1}_{Y_X=0} - \mathbb{1}_{Y_X=1}$. Note that when $x \in \mathcal{S}_r$,

$$\mathbb{E}_{q(xy|X=x)}[\mathsf{Bias}] = q(Y_x = 0) - q(Y_x = 1) \geq \gamma.$$

We will prove that there exists a *rectangle* with the following properties.

Claim 7.13. *There exists events $\mathcal{S} \subseteq \mathsf{supp}(X)$, $\mathcal{T} \subseteq \mathsf{supp}(Y)$ such that*

(a) $\mathbb{E}_{q(xy|(X,Y)\in\mathcal{S}\times\mathcal{T})}[\mathsf{Bias}] \geq \gamma/2$.

(b) $q(x|X \in \mathcal{S}) \leq \frac{2^{2/\gamma^2}}{m^{1-\gamma^2}}$ *for every* x.

(c) $\mathbf{H}_q(Y|Y \in \mathcal{T}) \geq m - \gamma^2 \log m - \frac{2}{\gamma^2}$.

Lets finish the proof of Lemma 7.5 first. For this, we define

$$\beta_x := q(Y_x = 0|Y \in \mathcal{T}) - q(Y_x = 1|Y \in \mathcal{T}).$$

Since $q(XY|(X,Y) \in \mathcal{S} \times \mathcal{T})$ is a product distribution, Claim 7.13(b) implies that

$$\mathbb{E}_{q(xy|(X,Y)\in\mathcal{S}\times\mathcal{T})}[\mathsf{Bias}] = \sum_x q(x|X \in \mathcal{S})\beta_x \leq \frac{2^{2/\gamma^2}}{m^{1-\gamma^2}} \sum_{x\in\mathcal{S}} \beta_x.$$

So, using Claim 7.13(a), we can say

$$\sum_{x\in\mathcal{S}} \beta_x \geq \frac{\gamma}{2^{2/\gamma^2+1}} m^{1-\gamma^2}, \text{ or equivalently, } \mathbb{E}_{x\in\mathcal{S}}[\beta_x] \geq \frac{\gamma}{2^{2/\gamma^2+1}} \frac{m^{1-\gamma^2}}{|\mathcal{S}|},$$

From Lemma 7.10, it then follows that

$$\mathbf{H}_q(Y|Y \in \mathcal{T}) \leq m - \left(\frac{\gamma}{2^{2/\gamma^2+2}} \frac{m^{1-\gamma^2}}{|\mathcal{S}|}\right)^2 |\mathcal{S}| = m - \frac{\gamma^2}{2^{4/\gamma^2+4}} \frac{m^{2-2\gamma^2}}{|\mathcal{S}|}.$$

As Claim 7.13(b) also implies that $|\mathcal{S}| \geq \frac{1}{2^{2/\gamma^2}} m^{1-\gamma^2}$, we have

$$\mathbf{H}_q(Y|Y \in \mathcal{T}) \leq m - \frac{\gamma^2}{2^{2/\gamma^2+4}} m^{1-\gamma^2},$$

which contradicts Claim 7.13(c) if

$$\frac{\gamma^2}{2^{2/\gamma^2+4}} m^{1-\gamma^2} > \gamma^2 \log m + \frac{2}{\gamma^2}.$$

One can check that if $\frac{3}{\log m} \leq \gamma^8 \leq \frac{1}{2^{64}}$, then the left hand side above is always at least $\Omega(\sqrt{m})$ while the right hand side is $O(\log m)$. This proves Lemma 7.5.

Next we turn to proving Claim 7.12 and Claim 7.13. We will need the following proposition.

Proposition 7.14. *For any events $\mathcal{A}$ and $\mathcal{B}$, $p(\mathcal{A}|\mathcal{B}, Y_X = 0) \geq p(\mathcal{A}|\mathcal{B})p(Y_X = 0|\mathcal{A}\mathcal{B})$.*

Proof of Claim 7.12. Let

$$\mathcal{G}_1 = \{r \mid \mathbf{H}_p(X|r, Y_X = 0) \geq (1-\gamma^4)\log m \text{ and } \mathbf{H}_p(Y|r, Y_X = 0) \geq m - \gamma^4 \log m\}, \text{ and}$$
$$\mathcal{G}_2 = \{r \mid p(X \in \mathcal{S}_r|r) \geq 32\gamma\}.$$

As $\mathbf{H}_p(X|r, Y_X = 0) \leq \log m$ and $\mathbf{H}_p(Y|r, Y_X = 0) \leq m$ for every r, Lemma 2.27 and union bound imply that $p(R \notin \mathcal{G}_1|Y_X = 0) \leq 2\left(\gamma^4 + \frac{3}{\gamma^4 \log m}\right) \leq 4\gamma^4$ as $\gamma^8 \geq \frac{3}{\log m}$.

Lemma 2.27 implies that $p(R \in \mathcal{G}_2) \geq 32\gamma$. Furthermore, for any $r \in \mathcal{G}_2$, we have

$$p(Y_X = 0|r) \geq p(X \in \mathcal{S}_r|r)p(Y_X = 0|r, X \in \mathcal{S}_r) \geq 16\gamma.$$

Since $\gamma \leq \frac{1}{2^8}$, using Proposition 7.14,

$$\begin{aligned} p(R \in \mathcal{G}|Y_X = 0) &\geq p(R \in \mathcal{G}_2|Y_X = 0) - p(R \notin \mathcal{G}_1|Y_X = 0) \\ &\geq p(R \in \mathcal{G}_2) \cdot 16\gamma - p(R \notin \mathcal{G}_1|Y_X = 0) \geq 512\gamma^2 - 4\gamma^4 \geq 256\gamma^2. \end{aligned}$$

□

Proof of Claim 7.13. By our choice of $r \in \mathcal{G}$, we have $q(X \in \mathcal{S}_r) \geq 32\gamma$ where $\mathcal{S}_r = \{x|q(Y_x = 0) \geq \frac{1+\gamma}{2}\}$. This also implies:

$$q(Y_X = 0) \geq q(X \in \mathcal{S}_r)q(Y_X = 0|X \in \mathcal{S}_r) \geq 16\gamma. \tag{7.4.1}$$

We define

$$\mathcal{S}_u = \{x \mid q(x|Y_X = 0) \leq \frac{m^{\gamma^2} 2^{1/\gamma^2}}{m}\} \text{ and } \mathcal{S} = \mathcal{S}_r \cap \mathcal{S}_u,$$
$$\mathcal{T}_b = \{y \mid \mathbb{E}_{q(xy|Y=y, X\in\mathcal{S})}[\mathsf{Bias}] \geq \frac{\gamma}{2}\}, \mathcal{T}_u = \{y \mid q(y|Y_X = 0) \leq \frac{m^{\gamma^2} 2^{1/\gamma^2}}{2^m}\}$$
$$\text{and } \mathcal{T} = \mathcal{T}_b \cap \mathcal{T}_u.$$

By definition, we have $\mathbb{E}_{q(xy|(X,Y)\in\mathcal{S}\times\mathcal{T})}[\mathsf{Bias}] \geq \frac{\gamma}{2}$ which establishes Claim 7.13(a).

We shall prove the following propositions.

Proposition 7.15. *For $x \in \mathcal{S}_r$, we have $q(x) \leq 2q(x|Y_X = 0)$. Also, $q(X \in \mathcal{S}|Y_X = 0) \geq 8\gamma$.*

Proposition 7.16. *For $y \in \mathcal{T}_b$, we have $q(y) \leq \frac{1}{4\gamma} q(y|Y_X = 0)$. Also, $q(Y \in \mathcal{T}|Y_X = 0) \geq \gamma^2$.*

Proposition 7.15 and Proposition 7.4.1 then imply that $q(X \in \mathcal{S}) \geq q(Y_X = 0)q(X \in \mathcal{S}|Y_X = 0) \geq 128\gamma^2$, and since $\gamma \leq \frac{1}{2^8}$, we get that

$$q(x|X \in \mathcal{S}) = \frac{q(x)}{q(x \in \mathcal{S})} \leq \frac{2q(x|Y_X = 0)}{q(x \in \mathcal{S})} \leq \frac{m^{\gamma^2} 2^{1/\gamma^2}}{64\gamma^2 m} \leq \frac{2^{2/\gamma^2}}{m^{1-\gamma^2}}.$$

holds for every $x \in \mathcal{S}$ which proves Claim 7.13(b).

With a similar argument, using Proposition 7.16 we get that for $y \in \mathcal{T}$, the following holds

$$q(y|Y \in \mathcal{T}) \leq \frac{m^{\gamma^2} 2^{1/\gamma^2}}{64\gamma^4 2^m} \leq \frac{m^{\gamma^2} 2^{2/\gamma^2}}{2^m}.$$

which implies that $\mathbf{H}_q(Y|Y \in \mathcal{T}) \geq m - \gamma^2 \log m - \frac{2}{\gamma^2}$ which gives Claim 7.13(c). □

Now we turn to proving Proposition 7.15 and Proposition 7.16.

Proof of Proposition 7.15. Recall that $q(Y_x = 0) \geq \frac{1+\gamma}{2}$ for $x \in \mathcal{S}_r$ and $q(X \in \mathcal{S}_r) \geq 32\gamma$. Since $q(XY) = p(XY|r)$, using Proposition Proposition 7.14, it follows that $q(x) \leq 2q(x|Y_X = 0)$ for $x \in \mathcal{S}_r$ and hence $q(X \in \mathcal{S}_r|Y_X = 0) \geq 16\gamma$.

The entropy bound $\mathbf{H}_q(X|Y_X = 0) \geq (1 - \gamma^4) \log m$ implies that $q(X \in \mathcal{S}_u|Y_X = 0) \geq 1 - \gamma^2$ using Lemma 7.9.

Since $\gamma \leq \frac{1}{2^8}$, we get

$$q(X \in \mathcal{S}|Y_X = 0) \geq q(X \in \mathcal{S}_r|Y_X = 0) - q(X \in \overline{\mathcal{S}}_u|Y_X = 0) \geq 16\gamma - \gamma^2 \geq 8\gamma.$$ □

Proof of Proposition 7.16. Note that the condition $\mathbb{E}_{q(xy|Y=y, X \in \mathcal{S})}[\mathsf{Bias}] \geq \gamma/2$ is equivalent to saying that $q(Y_X = 0|y, X \in \mathcal{S}) \geq \frac{1}{2} + \frac{\gamma}{4}$. As $q(XY) = p(XY|r)$ using Proposition 7.14, it follows that $q(y|X \in \mathcal{S}) \leq 2q(y|X \in \mathcal{S}, Y_X = 0)$ for $y \in \mathcal{T}_b$. Since $q(XY)$ is product, Bayes rule and Proposition 7.15 imply that for $y \in \mathcal{T}_b$, the following holds

$$\begin{aligned} q(y) = q(y|X \in \mathcal{S}) &\leq 2q(y|X \in \mathcal{S}, Y_X = 0) \\ &= \frac{2q(y|Y_X = 0)q(X \in \mathcal{S}|y, Y_X = 0)}{q(X \in \mathcal{S}|Y_X = 0)} \leq \frac{1}{4\gamma} q(y|Y_X = 0). \end{aligned}$$

By definition of $\mathcal{S}$, $p(Y_X = 0 | X \in \mathcal{S}) \geq \frac{1}{2} + \frac{\gamma}{2}$, or equivalently, $\mathbb{E}_{q(xy|X\in\mathcal{S})}[\mathsf{Bias}] \geq \gamma$. Lemma 2.27 then implies that $q(Y \in \mathcal{T}_b) \geq \gamma/2$ and hence $q(Y \in \mathcal{T}_b | Y_X = 0) \geq 4\gamma q(Y \in \mathcal{T}_b) \geq 2\gamma^2$.

Also, since $\mathbf{H}_q(Y|Y_X = 0) \geq m - \gamma^4 \log m$, using Lemma 7.9 we have $q(Y \in \mathcal{T}_u | Y_X = 0) \geq 1 - \gamma^2$. Hence, we can conclude

$$q(Y \in \mathcal{T} | Y_X = 0) \geq q(Y \in \mathcal{T}_b | Y_X = 0) - q(Y \in \overline{\mathcal{T}}_u | Y_X = 0) \geq \gamma^2. \qquad \square$$

7.4.2 Non-Negative Rank of Lopsided Unique Disjointness

In this section we prove Theorem 7.3.

Theorem 7.3. *For any non-negative matrix satisfying (7.0.3) and any $\frac{3\cdot 1000^8}{\log(n/k)} \leq \rho^8 \leq 1$, it holds that $\mathsf{rk}_+(A) \geq \binom{n}{\alpha\rho^8 k}$ with $0 < \alpha < 1$ an absolute constant.*

For the proof, it will be convenient and without any loss of generality to assume that n is divisible by k and $\frac{n}{k}$ is a large enough integer. We split the universe $[n]$ into blocks of size $\frac{n}{k}$ where $\{\frac{n}{k}(i-1)+1, \ldots, \frac{n}{k}i\}$ is the i^{th} block for every $i \in [k]$.

Define a distribution on $X \in \binom{[n]}{k}$ and $Y \in \{0,1\}^n$ where the probability of $x \in \binom{[n]}{k}$ and $y \in \{0,1\}^n$ is given by

$$q(xy) = \frac{A_{xy}}{\sum_{x',y'} A_{x'y'}},$$

where A is the lopsided unique disjointness matrix defined in (7.0.3) with parameter ρ and we view $y \in \{0,1\}^n$ as the indicator vector for a subset of $[n]$. Let X^i denote the intersection of X with the elements in the i^{th} block, and let Y^i be the projection of Y onto the coordinates in the i^{th} block. For every $j \in [\frac{n}{k}]$, we will use the notation Y^i_j to denote the jj^{th} coordinate of Y_i, in other words $Y_{\frac{n}{k}(i-1)+j}$.

Let $\mathcal{D}$ denote the event that x and y are disjoint and x^i has exactly one element for every i. Then, we will prove the following key lemma:

Lemma 7.17. *Let R be any random variable satisfying $X - R - Y$. Then, for every $i \in [k]$ it holds that*

$$\mathbf{I}_q\left(R : X^i | X^{<i} Y^{\geq i} \mathcal{D}\right) + \mathbf{I}_q\left(R : Y^i | X^{\leq i} Y^{>i} \mathcal{D}\right) \geq \left(\frac{\rho}{1000}\right)^8 \log\left(\frac{n}{k}\right).$$

With the above lemma in hand, the proof of Theorem 7.3 is straightforward and we present it below. Following that, we will give a proof of Lemma 7.17.

Proof of Theorem 7.3. Using Lemma 2.11 and Proposition 6.5 together with Lemma 7.17, we have

$$2\log \mathsf{rk}_+(A) \geq \sum_{i=1}^{k} \left(\mathbf{I}_q \left(R : X^i | X^{<i} Y^{\geq i} \mathcal{D} \right) + \mathbf{I}_q \left(R : Y^i | X^{\leq i} Y^{>i} \mathcal{D} \right) \right) \geq \left(\frac{\rho}{1000} \right)^8 k \log \left(\frac{n}{k} \right),$$

which proves Theorem 7.3 as $\binom{n}{s} = 2^{\Theta\left(s \log\left(\frac{n}{s}\right)\right)}$. □

Proof of Lemma 7.17. Fix i as in the statement of the lemma. Let $\mathcal{E}$ be the event that X^i has exactly one element for every i and $X \cap Y$ is a subset of the i^{th} block. Note that $\mathcal{D} \subset \mathcal{E}$.

Writing $W = X^{<i} Y^{>i}$, we will prove that for any fixed value w attained by W, we have that

$$\mathbf{I}_q \left(R : X^i | w Y^i \mathcal{D} \right) + \mathbf{I}_q \left(R : Y^i | w X^i \mathcal{D} \right) \geq \left(\frac{\rho}{1000} \right)^8 \log \left(\frac{n}{k} \right), \tag{7.4.2}$$

and the proof is completed by averaging over w.

Note that after fixing w and r, X and Y are independent and they can be checked separately to verify that the event $\mathcal{E}$ occurs as for every block $j \neq i$ either X^j or Y^j is fixed given w. It follows that the distribution $p(XYR) := q(XYR | w\mathcal{E})$ satisfies $p(xy|r) = p(x|r)p(y|r)$ for every x, y, r. Furthermore under the distribution $p(XYR)$, $\mathcal{D}$ is equivalent to the event that $Y^i_{X^i} = 0$. We can compute $p(\mathcal{D}) \geq \frac{1}{1+(1-\rho)} = \frac{1}{2-\rho}$ since the matrix entries are given by

$$A_{xy} = \begin{cases} 1 \text{ when } (x,y) \in \mathsf{supp}(p(XY|\mathcal{D})), \\ \leq 1 - \rho \text{ when } (x,y) \in \mathsf{supp}(p(XY|\overline{\mathcal{D}})), \end{cases} \tag{7.4.3}$$

and the number of entries is the same in both cases.

For the sake of contradiction assume that (7.4.2) does not hold. Then, we are going to show that $p(\mathcal{D})$ must be significantly smaller than what we computed above. Define $\mathcal{B} = \{(x^i, r) \mid p(Y^i_{x^i} =$

$0|r) \geq \frac{1}{2} + \frac{\rho}{2000}\}$. We can upper bound $p(\mathcal{D})$ as follows:

$$\begin{aligned} p(\mathcal{D}) &= \sum_{x^i,r} p(x^i,r) p(Y^i_{x^i} = 0 | r, X^i = x^i) = \sum_{x^i,r} p(x^i,r) p(Y^i_{x^i} = 0|r) \\ &\leq \sum_{(x^i,r)\notin\mathcal{B}} p(x^i,r) \left(\frac{1}{2} + \frac{\rho}{2000}\right) + \sum_{(x^i,r)\in\mathcal{B}} p(x^i,r) p(Y^i_{x^i} = 0|r) \\ &\leq \left(\frac{1}{2} + \frac{\rho}{2000}\right) + p((X^i,R) \in \mathcal{B}), \end{aligned}$$

where the second equality follows since $p(XY|r)$ is product.

We will show that

Claim 7.18. $p((X^i,R) \in \mathcal{B})) \leq \frac{64\rho}{1000}$.

This implies $p(\mathcal{D}) \leq \frac{1}{2} + \frac{\rho}{2000} + \frac{64\rho}{1000} < \frac{1}{2-\rho}$, which contradicts the fact that $p(\mathcal{D}) \geq \frac{1}{2-\rho}$. This finishes the proof. Next we prove Claim 7.18. □

Proof of Claim 7.18. Set $t := \frac{n}{k}$. Conditioned on $x\mathcal{D}$, every coordinate of Y^i other than x^i is uniform and hence $\mathbf{H}_p(Y^i|X^i\mathcal{D}) \geq t-1$. Similarly, conditioned on $y\mathcal{D}$, X^i is uniform on the coordinates of y^i that are zero. We may compute from (7.4.3) that $p(y^i|\mathcal{D}) = \frac{\Delta(y^i)}{t2^{t-1}}$ where $\Delta(y^i)$ is the number of zeros in y^i. The probability of any string y^i with less than $t/4$ zeros is at most $\frac{1}{2^{t+1}}$ and using Proposition 2.26 their total measure under the distribution $p(Y^i|\mathcal{D})$ can be bounded by $\frac{e^{-t/8}}{2}$. Hence, $\mathbf{H}_p(X^i|Y^i\mathcal{D}) \geq (1 - \frac{e^{-t/8}}{2}) \log\left(\frac{t}{4}\right) \geq \log t - 3$.

As $p(XYR) = q(XYR|w\mathcal{E})$, if (7.4.2) is not true, then

$$\mathbf{I}_p\left(R : X^i|Y^i\mathcal{D}\right) = \mathbf{H}_p(X^i|Y^i\mathcal{D}) - \mathbf{H}_p(X^i|RY^i\mathcal{D}) \leq \left(\frac{\rho}{1000}\right)^8 \log t,$$

and a similar statement is obtained by writing $\mathbf{I}_p\left(R : Y^i|X^i\mathcal{D}\right)$ in terms of entropy. It follows that

$$\mathbf{H}_p(X^i|RY^i\mathcal{D}) \geq \left(1 - \left(\frac{\rho}{1000}\right)^8\right) \log t - 3, \text{ and } \mathbf{H}_p(Y^i|RX^i\mathcal{D}) \geq t - \left(\frac{\rho}{1000}\right)^8 \log t - 1.$$

Since entropy can only decrease under conditioning and $\left(\frac{\rho}{1000}\right)^8 \geq \frac{3}{\log t}$, X^i, Y^i and R satisfy the conditions of Lemma 7.5 and the claim follows. □

7.4.3 Non-Negative Rank of the Matching Slack Matrix

Let us recall the definition of the matching slack matrix S from (7.2.1). The entry S_{um} corresponding to cut u and perfect matching m is

$$S_{um} = \frac{|\delta(u) \cap m| + \beta - 1}{2},$$

where $0 < \beta < 1$ is a constant that will be determined by the proof and $\delta(u)$ is the set of edges crossing u. In this section we prove Theorem 7.7.

Theorem 7.7. *For the matrix S defined in (7.2.1), it holds that*

$$\mathsf{rk}_+(S) \geq \binom{n}{\alpha/\varepsilon}$$

where $\frac{c'}{n} \leq \varepsilon \leq 1$ for a large enough constant $c' = c'(\beta)$ and $0 < \alpha < 1$ is an absolute constant.

For convenience in proving the above theorem, we assume without loss of generality that β/ε is an integer and that the graphs are defined on $n = 4n' + 6$ vertices where n' is divisible by β/ε. Set $t = \frac{n'}{\beta/\varepsilon}$ and note that $t \geq c$ for a large constant $c = c(\beta)$.

Using the slack matrix S, define a distribution on cuts U and perfect matchings M where the probability of a cut u and a perfect matching m is given by

$$q(um) = \frac{S_{um}}{\sum_{u',m'} S_{u'm'}}.$$

Fix an arbitrary perfect matching $\mathcal{A}$ and an arbitrary cut $\mathcal{Z}$ that cuts all edges of $\mathcal{A}$. Let $B = (B_0, B_1, \ldots, B_{n'})$ be a uniformly random partition of the set $\mathcal{Z}$ into blocks such that $|B_0| = 3$ and $|B_{i'}| = 2$ for each $i' \in [n']$. For $i \in [\frac{n'}{t}]$, we call $B^i = \{B_{t(i-1)+1}, \ldots, B_{ti}\}$ the i^{th} *chunk* and for $j \in [t]$ use $B^i_j = B_{t(i-1)+j}$ to denote the j^{th} block of the chunk B^i. Let $\mathcal{A}_{i'}$ denote the edges of $\mathcal{A}$ that touch $B_{i'}$.

We say that the cut U is *consistent* with B if $B_0 \subseteq U \subseteq \mathcal{Z}$ and for every $i \in [\frac{n'}{t}]$, $U^i := U \cap (B^i_1 \cup \cdots \cup B^i_t)$ equals B^i_j for some $j \in [t]$. We say that M is *consistent* with B if $\mathcal{A}_0 \subseteq M$ and for each $i' \in [n']$, either $\mathcal{A}_{i'} \subseteq M$ or M matches $B_{i'}$ to itself and matches the neighbors of $B_{i'}$ under $\mathcal{A}_{i'}$ to themselves.

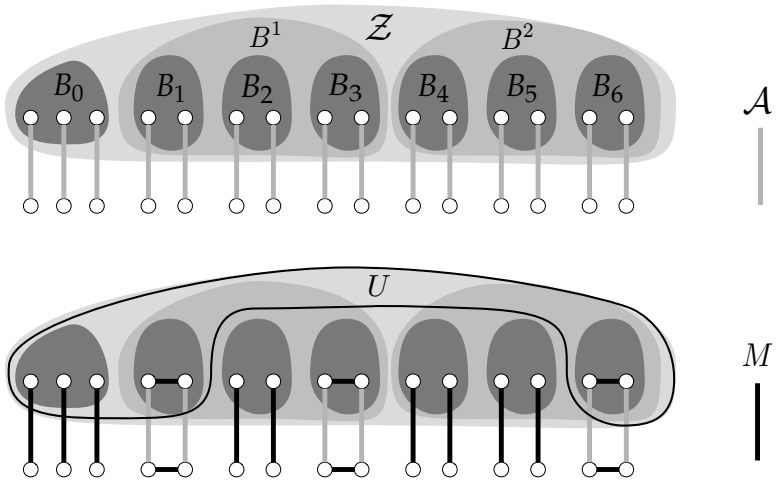


Figure 7.4.1: Top to bottom: An example of $\mathcal{A}, \mathcal{Z}, B$ with $n = 6$ and $t = 3$; U, M conditioned on $\mathcal{D}$.

For $i \in [\frac{n'}{t}]$, we write M^i to denote the edges of M contained in the union of vertices of B^i and neighbors of B^i under $\mathcal{A}$. We write M^i_j (and $\mathcal{A}^i_j$) to denote the edges of M^i (and $\mathcal{A}$) corresponding to the vertices of B^i_j and its neighbors under $\mathcal{A}$.

Let U and M be sampled from q and let $\mathcal{D}$ denote the event that U and M are consistent with B and for every $i \in [\frac{n'}{t}]$, U^i does not cut the edges of M^i. We prove the following key lemma:

Lemma 7.19. *Let R be any random variable satisfying $U - R - M$. Then, for every $i \in [\frac{n'}{t}]$ we have*

$$\mathbf{I}_q\left(R : U^i | BU^{<i}M^{\geq i}\mathcal{D}\right) + \mathbf{I}_q\left(R : M^i | BU^{\leq i}M^{>i}\mathcal{D}\right) \geq \beta^8 \log t,$$

where $\beta > 0$ is an absolute constant.

With the above lemma in hand, the proof of Theorem 7.7 is straightforward and we present it below. Following that, we will give a proof of Lemma 7.19.

Proof of Theorem 7.7. Recall that the number of vertices n in the underlying graph whose cuts and perfect matchings index the slack matrix S satisfies $n = 4n' + 6$ and also recall that $t = \varepsilon n'/\beta$.

Using Lemma 2.11 and Proposition 6.5 together with Lemma 7.19, we have

$$\begin{aligned} 2\log \mathsf{rk}_+(S) &\geq \sum_{i=1}^{n'/t} \left(\mathbf{I}_q \left(R : U^i | BU^{<i} M^{\geq i} \mathcal{D} \right) + \mathbf{I}_q \left(R : M^i | BU^{\leq i} U^{>i} \mathcal{D} \right) \right) \\ &\geq \frac{n'}{t} \cdot \beta^8 \log t = \frac{\beta^8}{\varepsilon} \log \left(\frac{\varepsilon n}{8\beta} \right), \end{aligned}$$

for large enough n. This proves Theorem 7.7 as $\binom{n}{s} = 2^{\Theta\left(s \log\left(\frac{n}{s}\right)\right)}$. □

Proof of Lemma 7.19. Fix a value of i as in the statement of the lemma. Let $\mathcal{E}$ be the event that U, M are consistent with B and for each $i' \neq i$, the edges of $M^{i'}$ are not cut by $U^{i'}$. Note that $\mathcal{D} \subset \mathcal{E}$ but under $\mathcal{E}$ edges of M^i may be cut by U^i.

For any partition b and for any $(u, m) \in \mathsf{supp}(q(UM|b\mathcal{E}))$, the weights of the slack matrix entries are given by

$$S_{um} = \begin{cases} \frac{2+\beta}{2} \text{ when } (u,m) \in \mathcal{D}, \\ \frac{4+\beta}{2} \text{ when } (u,m) \in \overline{\mathcal{D}}. \end{cases} \tag{7.4.4}$$

Also, note that the number of entries in each of the cases above is exactly the same, as after fixing $U^{<i} M^{>i} B^{-i}$, there are exactly $t2^{t-1}$ entries in each case: there are t choices for the blocks in U^i and if $U^i = B^i_j$, then $M^i_{j'}$ has 2 options for every $j' \neq j$.

Writing $W = U^{<i} M^{>i} B^{-i}$, we will prove that for any fixed value w attained by W, the following holds

$$\mathbf{I}_q \left(R : U^i | M^i B w \mathcal{D} \right) + \mathbf{I}_q \left(R : M^i | U^i B w \mathcal{D} \right) \geq \beta^8 \log t, \tag{7.4.5}$$

and the proof is completed by averaging over w. Observe that the partition of the i^{th} chunk B^i and B_0 is still a random variable even after fixing w.

After fixing wrb, U and M are independent and they can be checked separately to verify that the event $\mathcal{E}$ occurs as for every block $i' \neq i$ either $U^{i'}$ or $M^{i'}$ is fixed given wb. It follows that the distribution $p(UMBR)$ defined as $p(UMBR) = q(UMBR|w\mathcal{E})$ satisfies $p(um|rb) = p(u|rb)p(m|rb)$ for every u, m, r, b. A direct computation using (7.4.4) then shows that

$$p(\mathcal{D}) = \frac{2+\beta}{(2+\beta)+(4+\beta)} = \frac{2+\beta}{6+2\beta} \geq \frac{1}{3}.$$

For the sake of contradiction, assume that (7.4.5) does not hold. Then, we are going to argue that $p(\mathcal{D})$ must be significantly smaller than $\frac{1}{3}$. This is reminiscent to the proof of Lemma 7.17.

Define the random variable $J \in [t]$ to be $J = j$ if $U^i = B^i_j$ and note that the distribution $p(JM|rb)$ is product. Furthermore, note that under the distribution $p(UJMRB)$, the event $\mathcal{D}$ is equivalent to $M^i_J \neq \mathcal{A}^i_J$.

For any $(j,r,b) \in \mathsf{supp}(p(JRB))$, we define the set of blocks correlated with j as

$$\mathcal{J}_{jrb} = \left\{ k \;\middle|\; k \in [t], k \neq j, p(M^i_k \neq \mathcal{A}^i_k | rb, M^i_j \neq \mathcal{A}^i_j) \notin \left(\frac{1}{4}, \frac{3}{4}\right) \right\}.$$

Define sets:

$$\mathcal{S}_1 = \left\{ (j,r,b) \;\middle|\; p(M^i_j \neq \mathcal{A}^i_j | rb) \leq \frac{1}{30} \right\}, \quad \mathcal{S}_2 = \left\{ (j,r,b) \;\middle|\; p(M^i_j \neq \mathcal{A}^i_j | rb) \geq \frac{1+\beta}{2} \right\},$$
$$\mathcal{S}_3 = \left\{ (j,r,b) \;\middle|\; p(j|rb\mathcal{D}) \geq \frac{2^{1/\beta^2}}{t^{1-\beta^2}} \right\}, \qquad \mathcal{S}_4 = \left\{ (j,r,b) \;\middle|\; |\mathcal{J}_{jrb}| \geq \frac{\beta}{2} \cdot \frac{t^{1-\beta^2}}{2^{1/\beta^2}} - 1 \right\}.$$

For brevity, when $\mathcal{S} \subseteq [t] \times \mathsf{supp}(p(RB))$ we write $p(\mathcal{S})$ to denote the probability of the event $(j,r,b) \in \mathcal{S}$. Using Lemma 7.5 and other entropy related arguments, we will be able to show that the contribution of $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ and $\mathcal{S}_4$ to $p(\mathcal{D})$ is small.

Claim 7.20. *(a)* $p(\mathcal{S}_1, \mathcal{D}) \leq \frac{1}{30}$, *(b)* $p(\mathcal{S}_2, \mathcal{D}) \leq 64\beta$, *(c)* $p(\mathcal{S}_3, \mathcal{D}) \leq 3\beta^2$, *(d)* $p(\mathcal{S}_4, \mathcal{D}) \leq \beta$.

We continue with the proof of Lemma 7.19 for now and defer the proof of Claim 7.20 until the end of this section.

With Claim 7.20, denoting by $\mathcal{G}$ the complement of the event $\mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3 \cup \mathcal{S}_4$, we can bound

$$p(\mathcal{D}) \leq p(\mathcal{G}, \mathcal{D}) + \frac{1}{30} + 64\beta + 3\beta^2 + \beta \leq p(\mathcal{G}, \mathcal{D}) + \frac{1}{30} + 2^7\beta. \tag{7.4.6}$$

To bound the contribution of $\mathcal{G}$, we need to use the combinatorial structure of matchings. For this, we further split $\mathcal{G}$ into two events. We say $(j,r,b) \in \mathcal{G}_2$ if $(j,r,b) \in \mathcal{G}$ and for all partitions b' such that $(j,r,b') \in \mathcal{G}$ and b' agrees with b on all the blocks except b_0 and b^i_j, it holds that $|b_0 \cap b'_0| \geq 2$. We define $\mathcal{G}_1 = \mathcal{G} \setminus \mathcal{G}_2$. Note that by definition if $(j,r,b) \in \mathcal{G}_1$, then there exists another partition b' such that $(j,r,b') \in \mathcal{G}_1$ and b' agrees with b on all the blocks except b_0 and b^i_j.

We will bound the contribution of $\mathcal{G}_1$ and $\mathcal{G}_2$ to $p(\mathcal{G}, \mathcal{D})$ separately in the next two claims whose proofs we defer until after the completion of the proof of Lemma 7.19.

As discussed in Section 7.1.2, most of the contribution comes from the event $\mathcal{G}_2$ which we can bound by

Claim 7.21.

$$p(\mathcal{G}_2, \mathcal{D}) \leq \frac{4}{10} p(\overline{\mathcal{D}}) + \beta = \frac{4}{10} \cdot \frac{4+\beta}{6+2\beta} + \beta.$$

The contribution of the event $\mathcal{G}_1$ is trickier to bound. We want to relate the contribution of $\mathcal{G}_1$ to the probability of the event $|\delta(U) \cap M| = 1$ under the distribution q. Since, this probability is not zero under the slack matrix S, we will need some more notation to bound it more carefully. We call a matching M *bad* for B if M includes exactly one edge crossing B_0 that is an edge of $\mathcal{A}_0$ and for other blocks $i' \in [n']$ either $\mathcal{A}_{i'} \subseteq M$ or M matches $B_{i'}$ to itself and matches the neighbors of $B_{i'}$ under $\mathcal{A}_{i'}$ to themselves. See Figure 7.4.2.

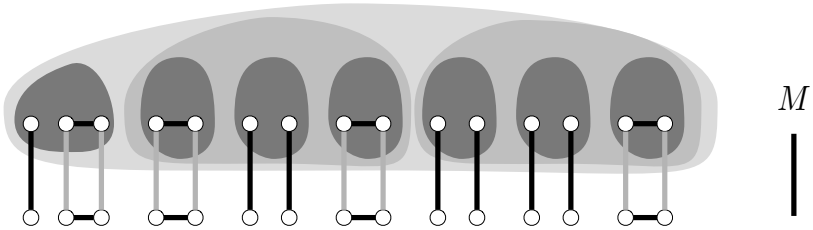


Figure 7.4.2: A bad matching M

Let $\mathcal{F}$ denote the event that U is consistent with B, M is bad for B and for each $i' \neq i$, the edges of $M^{i'}$ are not cut by $U^{i'}$. For any partition b, the entries of the slack matrix when $(u, m) \in \mathsf{supp}(q(UM|b\mathcal{F})$ are given by

$$S_{um} = \begin{cases} \frac{2+\beta}{2} \text{ when } |\delta(u) \cap m| = 3 \\ \frac{\beta}{2} \text{ when } |\delta(u) \cap m| = 1. \end{cases} \tag{7.4.7}$$

Note that the number of entries in the two cases above is exactly $3t2^{t-1}$ since for every fixing of all the edges of M except M_0, there are exactly three matchings that are bad.

With the above, we will be able to show that

Claim 7.22.

$$p(\mathcal{G}_1, \mathcal{D}) \leq 48000q(|\delta(U) \cap M| = 1 \mid \mathcal{E} \cup \mathcal{F}) + \beta \leq 2^{14}\beta.$$

Plugging the above claims in (7.4.6) and choosing β to be a sufficiently small constant, we derive that

$$p(\mathcal{D}) \leq \frac{4}{10} \cdot \frac{4+\beta}{6+2\beta} + \frac{1}{30} + 2^7\beta + 2^{14}\beta \leq \frac{9}{30} + 2^{15}\beta,$$

which is a contradiction as $p(\mathcal{D}) = \frac{2+\beta}{6+2\beta} \geq \frac{1}{3}$. This finishes the proof. Next we turn to proving the claims. □

Proof of Claim 7.21. For any $(j, r, b) \in \mathcal{G}$, we have $p(M^i_j \neq \mathcal{A}^i_j | rb) \leq p(M^i_j = \mathcal{A}^i_j | rb) + \beta$ since

$$p(M^i_j \neq \mathcal{A}^i_j | rb) - p(M^i_j = \mathcal{A}^i_j | rb) = 2p(M^i_j \neq \mathcal{A}^i_j | rb) - 1 \leq 2 \cdot \frac{1+\beta}{2} - 1 = \beta.$$

Recall that $\mathcal{D}$ is equivalent to the event $M^i_J \neq \mathcal{A}^i_J$ under p. As $\mathcal{G}_2 \subset \mathcal{G}$ and $p(JM|rb)$ is product,

$$\begin{aligned} p(\mathcal{D}, \mathcal{G}_2) = p(M^i_J \neq \mathcal{A}^i_J, \mathcal{G}_2) &= \sum_{(j,r,b) \in \mathcal{G}_2} p(jrb) p(M^i_j \neq \mathcal{A}^i_j | rb) \\ &\leq \sum_{(j,r,b) \in \mathcal{G}_2} p(jrb) p(M^i_j = \mathcal{A}^i_j | rb) + \beta = p(M^i_J = \mathcal{A}^i_J, \mathcal{G}_2) + \beta. \end{aligned}$$

Conditioned on the event $J = j, M^i_j = \mathcal{A}^i_j, R = r$, by symmetry the partition B^i is uniform among all partitions that agree on all the blocks except B_0 and B^i_j (recall Figure 7.1.2). For every fixing of the blocks outside B_0 and B^i_j, there are $\binom{5}{3}$ such partitions, out of which at most 4 can be in $\mathcal{G}_2$ by Lemma 7.11. In particular, this means that for any (j, r) we can bound

$$p(\mathcal{G}_2 |\, M^i_j \neq \mathcal{A}^i_j, J = j, r) = \sum_{b:(j,r,b) \in \mathcal{G}_2} p(b | M^i_j = \mathcal{A}^i_j, J = j, r) \leq \frac{4}{\binom{5}{3}} = \frac{4}{10}$$

Averaging over j and r, we get that $p(\mathcal{G}_2 | M^i_J = \mathcal{A}^i_J) \leq \frac{4}{10}$. Finally,

$$\begin{aligned} p(\mathcal{G}_2, \mathcal{D}) &\leq p(M^i_J = \mathcal{A}^i_J, \mathcal{G}_2) + \beta = p(M^i_J = \mathcal{A}^i_J) p(\mathcal{G}_2 | M^i_J = \mathcal{A}^i_J) \\ &\leq \frac{4}{10} p(M^i_J = \mathcal{A}^i_J) + \beta = \frac{4}{10} p(\overline{\mathcal{D}}) + \beta. \end{aligned}$$

□

Proof of Claim 7.22. For every (r,b) fix an ℓ arbitrarily such that $(\ell,r,b) \in \mathcal{G}_1$ (if such an ℓ does not exist then the contribution of r,b to $p(\mathcal{G}_1,\mathcal{D})$ is zero anyway). Then by definition, there exists a partition b' such that $(\ell,r,b') \in \mathcal{G}_1$, b and b' agree on all the blocks except $b_0 \cup b^i_\ell$ and $|b_0 \cap b'_0| = 1$. Note that b' is determined by r,b. Define $\mathcal{T}_{rb} := \mathcal{J}_{\ell rb} \cup \mathcal{J}_{\ell rb'} \cup \{\ell\}$ and let $\mathcal{T}$ denote the event that $J \in \mathcal{T}_{rb}$. We will show that when $(j,r,b) \in \mathcal{G}_1$ and $j \notin \mathcal{T}_{rb}$:

$$p(J=j, M^i_j \neq \mathcal{A}^i_j, r, b) \leq 48000q(J=j, M^i_j \neq \mathcal{A}^i_j, M \text{ is bad}, r, b | \mathcal{E} \cup \mathcal{F}). \tag{7.4.8}$$

The above implies that

$$\begin{aligned} p(\mathcal{G}_1, \overline{\mathcal{T}}, M^i_J \neq \mathcal{A}^i_J) &\leq 48000q(\mathcal{G}_1, \overline{\mathcal{T}}, M^i_J \neq \mathcal{A}^i_J, M \text{ is bad} | \mathcal{E} \cup \mathcal{F}) \\ &\leq 48000q(M^i_J \neq \mathcal{A}^i_J, M \text{ is bad} | \mathcal{E} \cup \mathcal{F}). \end{aligned} \tag{7.4.9}$$

Furthermore, since both (ℓ,r,b) and (ℓ,r,b') are in $\mathcal{G}_1 \subseteq \mathcal{G}$, $|\mathcal{T}_{rb}| \leq \beta \cdot \frac{t^{1-\beta^2}}{2^{1/\beta^2}} - 1 \leq \beta \cdot \frac{t^{1-\beta^2}}{2^{1/\beta^2}}$ and also $p(j|rb,\mathcal{D}) \leq \frac{2^{1/\beta^2}}{t^{1-\beta^2}}$ when $(j,r,b) \in \mathcal{G}_1$. So, we have $p(\mathcal{T},\mathcal{G}_1|\mathcal{D}) \leq \beta$.

When $M^i_J \neq \mathcal{A}^i_J$ and M is bad, then the cut $U^i = B_0 \cup B^i_J$ satisfies $|\delta(U) \cap M| = 1$ (see Figure 7.4.3(b)). So, from (7.4.9) and the above, we get

$$\begin{aligned} p(\mathcal{G}_1,\mathcal{D}) = p(\mathcal{G}_1, M^i_J \neq \mathcal{A}^i_J) &\leq p(\mathcal{G}_1, \overline{\mathcal{T}}, M^i_J \neq \mathcal{A}^i_J) + p(\mathcal{G}_1, \mathcal{T}|\mathcal{D}) \\ &\leq 48000q(M^i_J \neq \mathcal{A}^i_J, M \text{ is bad} | \mathcal{E} \cup \mathcal{F}) + \beta \\ &\leq 48000q(|\delta(U) \cap M| = 1 \mid \mathcal{E} \cup \mathcal{F}) + \beta. \end{aligned}$$

Using (7.4.4) and (7.4.7), the first term on the right hand side is $\frac{3\beta}{(2+\beta)+(4+\beta)+3(2+\beta)+3\beta} \leq \frac{\beta}{4}$ and hence $p(\mathcal{G}_1,\mathcal{D}) \leq 2^{14}\beta$.

All that remains is to prove (7.4.8). Since $(\ell,r,b) \in \mathcal{G}$ and $j \notin \mathcal{T}_{rb}$, $p(M^i_\ell \neq \mathcal{A}^i_\ell | rb) \geq \frac{1}{30}$ and $p(M^i_j \neq \mathcal{A}^i_j | rb, M^i_\ell \neq \mathcal{A}^i_\ell) \geq \frac{1}{4}$, so $p(M^i_j \neq \mathcal{A}^i_j, M^i_\ell \neq \mathcal{A}^i_\ell | rb) \geq \frac{1}{120}$. Since probability is always less than one, trivially

$$p(M^i_j \neq \mathcal{A}^i_j | rb) \leq 120 p(M^i_j \neq \mathcal{A}^i_j, M^i_\ell \neq \mathcal{A}^i_\ell | rb).$$

Using the above and the fact that $p(JM|rb)$ is product, we get

$$p(J=j, M^i_j \neq \mathcal{A}^i_j, r, b) \leq 120 p(J=j, M^i_j \neq \mathcal{A}^i_j, M^i_\ell \neq \mathcal{A}^i_\ell, r, b) \tag{7.4.10}$$

Next we relate the probability on the right hand side above to the probability under q conditioned on the event $\mathcal{E} \cup \mathcal{F}$ as follows:

$$p(J = j, M_j^i \neq \mathcal{A}_j^i, M_\ell^i \neq \mathcal{A}_\ell^i, r, b) = \frac{q(J = j, M_j^i \neq \mathcal{A}_j^i, M_\ell^i \neq \mathcal{A}_\ell^i, \mathcal{E}, r, b \mid \mathcal{E} \cup \mathcal{F})}{q(\mathcal{E} \mid \mathcal{E} \cup \mathcal{F})}$$
$$\leq \frac{20}{6} q(J = j, M_j^i \neq \mathcal{A}_j^i, M_\ell^i \neq \mathcal{A}_\ell^i, \mathcal{E}, r, b \mid \mathcal{E} \cup \mathcal{F})$$

where the last inequality follows since $q(\mathcal{E} \mid \mathcal{E} \cup \mathcal{F}) = \frac{(4+\beta)+(2+\beta)}{(2+\beta)+(4+\beta)+3(2+\beta)+3\beta} = \frac{6+2\beta}{12+8\beta} \geq \frac{6}{20}$ as $\beta < 1$.

Plugging the above in (7.4.10),

$$p(J = j, M_j^i \neq \mathcal{A}_j^i, r, b) \leq 400 q(J = j, M_j^i \neq \mathcal{A}_j^i, M_\ell^i \neq \mathcal{A}_\ell^i, \mathcal{E}, r, b \mid \mathcal{E} \cup \mathcal{F}). \tag{7.4.11}$$

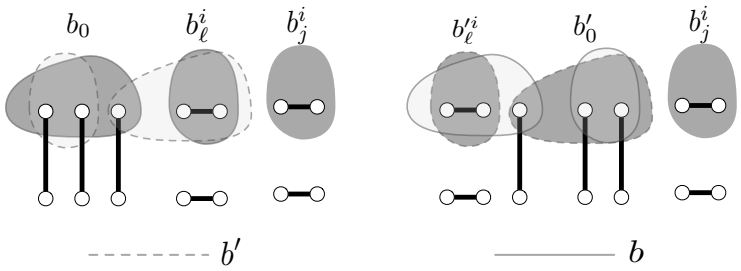
Note that the distribution $q(JM|rb,\ \mathcal{E} \cup \mathcal{F})$ is still product as J and M are independent given rb and given the partition b one can check if the matching m is bad or consistent without knowing the cut. Next, we relate the probability of the right hand side to the probability under the partition b' where b' is the partition guaranteed from (ℓ, r, b) being in $\mathcal{G}_1$. We will show that

$$q(M_j^i \neq \mathcal{A}_j^i, M_\ell^i \neq \mathcal{A}_\ell^i, \mathcal{E} \mid rb,\ \mathcal{E} \cup \mathcal{F}) \leq 120 q(M_j^i \neq \mathcal{A}_j^i, M_\ell^i \neq \mathcal{A}_\ell^i, \mathcal{E} \mid rb', \mathcal{E} \cup \mathcal{F}). \tag{7.4.12}$$

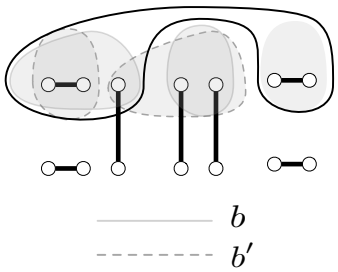
Note that the two events above correspond to different matchings since the partitions are different (see Figure 7.4.3(a)). The probability of any matching that agrees with the event on the right hand side above is zero under $p(M|rb)$, but that is not the case under $q(M|rb, \mathcal{E} \cup \mathcal{F})$. In fact, by symmetry the probability of any such matching is the same under the distribution $q(M|rb, \mathcal{E} \cup \mathcal{F})$ and $q(M|rb', \mathcal{E} \cup \mathcal{F})$. This is because as $q(JM|rb, \mathcal{E} \cup \mathcal{F})$ is product, the probability does not change under conditioning on the event that $J = \ell$ which is the same as conditioning that the cut is $b_0 \cup b_\ell^i$. Then, since the partitions were picked uniformly, both the splits (b_0, b_ℓ^i) and $(b_0', b_\ell'^i)$ are equally likely even when conditioned on $J = \ell, \mathcal{E} \cup \mathcal{F}$ since the matching is either consistent or bad with respect to both of them.

Note that any matching which agrees with the event on the right hand side above is bad for b (see Figure 7.4.3(b)). From this it follows that

$$q(M_j^i \neq \mathcal{A}_j^i, M_\ell^i \neq \mathcal{A}_\ell^i, \mathcal{E} \mid rb,\ \mathcal{E} \cup \mathcal{F}) \leq 120 q(M_j^i \neq \mathcal{A}_j^i, M \text{ is bad} \mid rb, \mathcal{E} \cup \mathcal{F}).$$



(a) Partial Matchings corresponding to the event $M^i_j \neq \mathcal{A}^i_j, M^i_\ell \neq \mathcal{A}^i_\ell$ under partition b(left) and under partition b' (right). Note that b and b' agree in all blocks except b_0 and b^i_ℓ.



(b) A matching agreeing with the event $M^i_j \neq \mathcal{A}^i_j, M^i_\ell \neq \mathcal{A}^i_\ell, \mathcal{E}$ under partition b' and the cut $U^i = b^i_0 \cup b^i_j$. Note that the matching is bad for the partition b.

Plugging the above in (7.4.11) and using that the distribution is product, we get (7.4.8). To complete the proof, we next show that (7.4.12) holds.

For this let us recall that since (ℓ, r, b) and (ℓ, r, b') are both in $\mathcal{G}$, when $j \notin \mathcal{T}_{rb}$, we have $p(M^i_j \neq \mathcal{A}^i_j, M^i_\ell \neq \mathcal{A}^i_\ell | rb) \geq \frac{1}{120}$ and a similar statement also holds with respect to b'.

Since $p(M|rb) = q(M|rb, \mathcal{E})$, we can write

$$\frac{p(M^i_j \neq \mathcal{A}^i_j, M^i_\ell \neq \mathcal{A}^i_\ell | rb)}{p(M^i_\ell = \mathcal{A}^i_\ell | rb)} = \frac{q(M^i_j \neq \mathcal{A}^i_j, M^i_\ell \neq \mathcal{A}^i_\ell, \mathcal{E} |\, rb, \mathcal{E} \cup \mathcal{F})}{q(M^i_\ell = \mathcal{A}^i_\ell, \mathcal{E} |\, rb', \mathcal{E} \cup \mathcal{F})},$$

and a similar statement holds for b' also.

Similar to the argument before, by symmetry $q(M^i_\ell = \mathcal{A}^i_\ell, \mathcal{E} | rb) = q(M^i_\ell = \mathcal{A}^i_\ell, \mathcal{E} | rb')$ and

also $p(M^i_\ell = \mathcal{A}^i_\ell | rb) = p(M^i_\ell = \mathcal{A}^i_\ell | rb')$. It follows that

$$\frac{q(M^i_j \neq \mathcal{A}^i_j, M^i_\ell \neq \mathcal{A}^i_\ell, \mathcal{E} \mid rb, \mathcal{E} \cup \mathcal{F})}{q(M^i_j \neq \mathcal{A}^i_j, M^i_\ell \neq \mathcal{A}^i_\ell, \mathcal{E} \mid rb', \mathcal{E} \cup \mathcal{F})} = \frac{p(M^i_j \neq \mathcal{A}^i_j, M^i_\ell \neq \mathcal{A}^i_\ell | rb)}{p(M^i_j \neq \mathcal{A}^i_j, M^i_\ell \neq \mathcal{A}^i_\ell | rb')} \leq 120,$$

which proves (7.4.12). This completes the proof.

□

Proof of Claim 7.20. We have $\mathbf{H}_p(M^i | U^i B\mathcal{D}) \geq t - 1$ since when $U^i = B^i_j$ then M^i_k either equals $\mathcal{A}^i_k$ or is not equal to $\mathcal{A}^i_k$ with probability $\frac{1}{2}$ independently for each $k \neq j$ conditioned on $\mathcal{D}$. Moreover, conditioned on $\mathcal{D}$, U^i is uniform among blocks B^i_j such that $M^i_j \neq \mathcal{A}^i_j$. We may compute from (7.4.4) that $p(m^i | b\mathcal{D}) = \frac{\Delta(m^i)}{t2^{t-1}}$ where $\Delta(m^i)$ is the number of blocks k of m^i such that $m^i_k \neq \mathcal{A}^i_k$. It follows that the probability of any m^i with $\Delta(m^i) \leq \frac{t}{4}$ is at most $\frac{1}{2^{t+1}}$ and using Proposition 2.26 (by considering the indicator vector for $M^i_k \neq \mathcal{A}^i_k$ for $k \in [t]$) their total measure under the distribution $p(M^i | b\mathcal{D})$ can be bounded by $\frac{e^{-t/8}}{2}$. Hence, $\mathbf{H}_p(U^i | M^i B\mathcal{D}) \geq (1 - \frac{e^{-t/8}}{2}) \log\left(\frac{t}{4}\right) \geq \log t - 3$.

As $p(UMRB) = q(UMRB | w\mathcal{E})$, if (7.4.5) is not true, then

$$\mathbf{I}_p\left(R : U^i | M^i B\mathcal{D}\right) = \mathbf{H}_p(U^i | M^i B\mathcal{D}) - \mathbf{H}_p(U^i | RM^i B\mathcal{D}) \leq \beta^8 \log t,$$

and a similar statement is obtained by writing $\mathbf{I}_p\left(R : M^i | U^i B\mathcal{D}\right)$ in terms of entropy. It follows that

$$\mathbf{H}_p(U^i | RBM^i\mathcal{D}) \geq (1 - \beta^8) \log t - 3, \text{ and } \mathbf{H}_p(M^i | RBU^i\mathcal{D}) \geq t - \beta^8 \log t - 1.$$

In terms of the random variable J (recall that $J = j$ iff $U^i = B^i_j$), we get

$$\mathbf{H}_p(J | RBM^i\mathcal{D}) \geq (1 - \beta^8) \log t - 3, \text{ and } \mathbf{H}_p(M^i | JRB\mathcal{D}) \geq t - \beta^8 \log t - 1.$$

For rest of the proof, it will be helpful to keep in mind that under the distribution p, $\mathcal{D}$ is equivalent to the event that $M^i_J \neq \mathcal{A}^i_J$.

(a) Follows from definition of $\mathcal{S}_1$.

(b) In the probability space p, define random variable $Y \in \{0,1\}^t$ as follows: for $k \in [t]$, $Y_k = \mathbb{1}_{M^i_k = \mathcal{A}^i_k}$. Then, it holds that $J - RB - Y$ and also, we have that

$$\mathbf{H}_p(J | RB, Y_J = 0) \geq (1 - \beta^8) \log t - 3 \text{ and } \mathbf{H}_p(Y | RB, Y_J = 0) \geq t - \beta^8 \log t - 3.$$

Applying Lemma 7.5 gives us $p(\mathcal{S}_2, \mathcal{D}) \leq p(\mathcal{S}_2) \leq 64\beta$.

(c) Since removing conditioning only increases entropy, we have that

$$\mathbf{H}_p(U^i|RB\mathcal{D}) \geq (1-\beta^8)\log t - 3.$$

Let $\mathcal{T} = \{(r,b) \mid \mathbf{H}_p(U^i|rb\mathcal{D}) \geq (1-\beta^4)\log t\}$. As $\mathbf{H}_p(U^i|rb\mathcal{D}) \leq \log t$, Lemma 2.27 then says that $p((R,B) \notin \mathcal{T}|\mathcal{D}) \leq \beta^4 + \frac{3}{\beta^4 \log t} \leq 2\beta^4$ where the last inequality follows as $\beta^8 \log t \geq 3$ when β is a constant and $t \geq c = c(\beta)$ for a sufficiently large constant c. Lemma 7.9 implies that for any $(r,b) \in \mathcal{T}$, $p(\mathcal{S}_3|rb\mathcal{D}) \leq \beta^2$. By a union bound,

$$p(\mathcal{S}_3,\mathcal{D}) \leq p(\mathcal{S}_3|(R,B) \in \mathcal{T},\mathcal{D}) + p((R,B) \notin \mathcal{T}|\mathcal{D}) \leq \beta^2 + 2\beta^4 \leq 3\beta^2.$$

(d) Set $a = \beta \frac{t^{1-\beta^2}}{2^{1/\beta^2+1}} - 1$. Note that for any $k \in \mathcal{J}_{jrb}$, $\mathbf{H}_p(M^i_k|rb, M^i_j \neq \mathcal{A}^i_j) \leq \mathsf{h}\left(\frac{1}{4}\right) \leq \frac{9}{10}$. When $(j,r,b) \in \mathcal{S}_4$, $|\mathcal{J}_{jrb}| \geq a$ and hence $\mathbf{H}_p(M^i|rb, M^i_j \neq \mathcal{A}^i_j) \leq t - \frac{a}{10}$. Since, $p(JM|rb)$ is product, $\mathbf{H}_p(M^i|rb, J=j, M^i_j \neq \mathcal{A}^i_j) = \mathbf{H}_p(M^i|rb, M^i_j \neq \mathcal{A}^i_j)$ and we can say that

$$\mathbf{H}_p(M^i|JRB\mathcal{D}) = \mathbb{E}_{p(jrb|\mathcal{D})}[\mathbf{H}_p(M^i|rb, M^i_j \neq \mathcal{A}^i_j)] \geq t - \beta^8 \log t - 1.$$

Using Lemma 2.27, we get that $p(\mathcal{S}_4,\mathcal{D}) \leq p(\mathcal{S}_4|\mathcal{D}) \leq \frac{\beta^8 \log t + 1}{a/10} \leq \beta$ as $t \geq c = c(\beta)$ for a large enough c.

□

7.5 Common Information and Small Set Disjointness

In this section, we prove a limitation on the common information technique. In fact, we will prove a limitation on the conditional common information technique as defined by Braun and Pokutta [BP16] (see Section 6.4.2). Let Z be another random variable and $\mathcal{E}$ be an event. Recall that the common information between X and Y conditioned on $Z, \mathcal{E}$ is defined as

$$\mathbb{C}(X:Y|Z\mathcal{E}) = \inf_R \mathbf{I}(XY:R|Z\mathcal{E}),$$

where the infimum is taken over all finite random variables R in all extensions of the probability space satisfying the following

1. $X - R - Y$
2. Given XY, we have that $Z\mathcal{E}$ and R are independent: $p(r, z, \mathcal{E}|xy) = p(r|xy)p(z, \mathcal{E}|xy)$.

Braun and Pokutta [BP16] showed that if we consider the distribution obtained by normalizing a non-negative matrix M, then the non-negative rank of M is lower bounded by $\mathbb{C}(X : Y|Z\mathcal{E})$ for any Z and $\mathcal{E}$ satisfying the above conditions. All non-negative rank lower bounds proven using this technique have been tight so far. Here, we give an example showing that the conditional common information can be exponentially smaller than the non-negative rank.

Our example is the small set disjointness matrix where both the rows and columns are indexed by sets of size at most k from the universe $[n]$. For technical convenience, we will work with a structured submatrix of the small set disjointness matrix. To define the submatrix, let us assume that k divides n. Let us divide our universe into k blocks of size $\frac{n}{k}$ and let $\mathcal{X}$ be the collection of subsets of $[n]$ with exactly one element in each block. Consider the small set disjointness matrix $M \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ where $M_{xy} = 1$ iff $|x \cap y| = 0$. It is easy to see that $M = (J - I)^{\otimes k}$ where J is the all ones $\frac{n}{k} \times \frac{n}{k}$ matrix and I is the $\frac{n}{k} \times \frac{n}{k}$ identity matrix. Since $\mathsf{rk}(A^{\otimes k}) \geq k \cdot \mathsf{rk}(A)$ where $\mathsf{rk}(A)$ denotes the rank of a matrix A, it follows that the rank and hence the non-negative rank of M is at least $\Omega\left(\left(\frac{n}{k}\right)^k\right) \geq \binom{n}{ck}$ for an absolute constant $0 < c < 1$.

Now, we show that the common information approach that gave us the lower bound for lopsided unique disjointness cannot prove that the non-negative rank of this matrix is at least $\binom{n}{ck}$ for any constant $c > 0$.

Define a distribution $p(XY)$ where the probability of x, y is given by

$$p(xy) = \frac{M_{xy}}{\sum_{x',y'} M_{x'y'}}.$$

Note that the distribution $p(XY)$ is only supported over sets X and Y that are disjoint.

Lemma 7.23. *For all Z and $\mathcal{E}$ satisfying the conditions in the definition of conditional common information defined above, we have that*

$$\mathbb{C}(X : Y|Z\mathcal{E}) = O(k \log\log n).$$

Taking $k = O(1)$, we get that the non-negative rank of M is $\mathrm{poly}(n)$ while the conditional common information is $\mathrm{poly}(\log n)$. We remark that if in the definition of conditional common information, we allow random variables supported over a countably infinite set, then in Lemma 7.23, we can obtain an upper bound of $O(k)$ making the separation between non-negative rank of M and conditional common information arbitrarily large by choosing $k = O(1)$. Next we prove Lemma 7.23.

Proof of Lemma 7.23. We will construct a finitely supported random variable R such that we have $\mathbf{I}(XY : R|Z\mathcal{E}) = O(k \log\log n)$ and R will also satisfy the conditions in the definition of conditional common information. This will prove the statement of the lemma.

Define the i^{th} block to be $\left\{\frac{n}{k}(i-1)+1, \ldots, \frac{n}{k}i\right\}$ and let X_i (and Y_i) denote the projection of X (and Y) on the i^{th} block. Note that X_i and Y_i, which are both singleton sets, are not the same since X and Y are disjoint. Also noting that the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are all independent, let us define random variables R_i for $i \in [k]$ such that for each $i \in [k]$, the following conditions hold (a) R_i is independent of $X_{-i}Y_{-i}$, (b) $X_i - R_i - Y_i$, and (c) given X_iY_i, we have that $Z\mathcal{E}$ and R_i are independent. Then, it is easily seen that the random variable $R = R_1, \cdots, R_k$ satisfies the conditions in the definition of conditional common information.

Let us fix an i now and define the random variable R_i. Let $\ell = c \log\log n$ for a constant $c > 0$ that will be chosen later. For this let $S^i_1, S^i_2, \ldots, S^i_\ell$ be a sequence of uniform random subsets of the i^{th} block. Let $T_i \in \{0,1\}^{\log \frac{n}{k}}$ be a random variable which is all zeros if there is a $j \in [\ell]$ such that $\mathbb{1}[X \in S^i_j] = \mathbb{1}[Y \in \overline{S^i_j}] = 1$ and is the bit representation of X_i otherwise. Then, we can define $R_i \in \{0,1\}^{2\ell + \log \frac{n}{k}}$ as follows

$$R_i = \mathbb{1}[X \in S^i_1], \cdots, \mathbb{1}[X \in S^i_\ell], \mathbb{1}[Y \in \overline{S^i_1}], \cdots, \mathbb{1}[X \in \overline{S^i_\ell}], T_i.$$

The above R_i satisfies all the conditions we want. Next we upper bound $\mathbf{H}(R_i|Z\mathcal{E})$ by bounding the expected encoding length of R_i given Z and $\mathcal{E}$. Since S^i_j for $j \in [\ell]$ are independent, we have that that for each $j \in [\ell]$, the probability that $X \in S^i_j$ and $Y \in \overline{S^i_j}$ is $1/4$. It follows that T_i is the all zeros string except with probability $(3/4)^\ell$ for every fixed x, y. It follows that, given x, y, the expected encoding length of R_i is $2\ell + (3/4)^\ell \log n = O(\log\log n)$ for a suitably chosen constant c. Since,

R_i is independent of $X_{-i}Y_{-i}$ and given X_iY_i, we have that $Z\mathcal{E}$ and R_i are independent, the same upper bound on encoding length follows given Z and $\mathcal{E}$. Hence, $\mathbf{H}(R_i|Z\mathcal{E}) = O(\log\log n)$.

Now by the chain rule

$$\mathbf{I}\left(XY : R|Z\mathcal{E}\right) \leq \sum_{i=1}^{k} \mathbf{H}(R_i|Z\mathcal{E}) = O(k\log\log n).$$

Furthermore, R satisfies all the conditions in the definition of conditional common information. This completes the proof of the lemma. □

We believe that one should be able to prove a tight lower bound on the non-negative rank of any small set unique disjointness matrix (where the entries corresponding to sets that intersect in more than one element can be arbitrary non-negative reals) by using the hyperplane separation bound (see Section 6.4.3). This will prove an exponential separation between the hyperplane separation bound and the common information bound answering a question posed by Pokutta [Pok15] who asked if these two bounds are always polynomially related. Note that the matrix M considered above is only one matrix in the family of small set unique disjointness matrices.

8 Conclusion

Proving unconditional lower bounds on computation is often a very challenging task. This thesis contributes new information theoretic tools to prove lower bounds for the computational settings of interactive compression and linear programs. Below, we briefly recap our contributions and relate it to some interesting questions that remain unanswered.

8.1 Interactive Models

The first part of this thesis explored whether it is possible to compress the communication when there are multiple parties who receive correlated inputs and interact with each other based on their inputs.

8.1.1 Compressing Two-party Communication

Chapter 4 studied the question of interactive compression for two-party communication: given a protocol with internal information cost I, can it be compressed to $O(I)$ bits of communication? We showed that there is a function which can be computed with internal information cost I, but requires communication $2^{O(I)}$. This ruled out an efficient compression in terms of the internal information cost I, but as discussed in Chapter 4, it is still possible to get an efficient compression, if we allow a polylogarithmic dependence on the communication complexity C of the original protocol:

Open Question 8.1. *Given an input distribution $p(X,Y)$, if a protocol π has $\mathsf{IC}_p(\pi) = I$ and communication complexity C, can it be simulated up to a constant error by a protocol with communication $I \cdot \mathrm{polylog}(C)$?*

Braverman, Ganor, Kol and Raz [BGKR18] present a candidate that might refute the above conjecture. Regardless of whether their candidate example holds up, it will be quite interesting to give an analysis of their example, as it will probably give us useful insights into what is required to construct protocols with better compression guarantees.

A closely related open question, to the one studied in Chapter 4, is whether there is a boolean function which can be computed with small external information cost but requires large communication. The work of Ganor, Kol and Raz [GKR15] already shows that this is true for a search problem but the same is not known for boolean functions.

Open Question 8.2. *Is there a boolean function $f(x,y)$ and an input distribution $p(X,Y)$ satisfying the following two conditions simultaneously:*

(1) On inputs from $p(X,Y)$, the function $f(x,y)$ can be computed with error at most $1/3$ with a protocol π that has $\mathsf{IC}_p^{ext}(\pi) = I_{ext}$, and

(2) Any protocol that computes $f(x,y)$ with error at most $1/3$ on the input distribution $p(X,Y)$ requires communication at least $2^{\Omega(I_{ext})}$?

8.1.2 Direct Sum and Compression for Streaming Algorithms

Chapter 5 looked at the setting of streaming algorithms. We proved a direct sum theorem for the average memory used by streaming algorithms when the input updates were independent, however, the parameters were not optimal. We also defined the notion of cumulative information cost of a streaming algorithm and showed that if one can compress the average memory used by streaming algorithms to their cumulative information content, then an optimal direct sum-would hold if input updates were independent. The obvious open question is if such a compression is possible.

Open Question 8.3. *Let $p(X_1,...,X_n)$ be an input distribution where the updates $X_1,...,X_n$ are independent. Can any streaming algorithm with cumulative information cost I under the distribution $p(X_1,...,X_n)$, be simulated with average memory $O(I)$?*

Even if the above conjecture is true with different parameters, it could still give an interesting

direct sum result for streaming algorithms. Chapter 5 also presented some partial results towards proving the above conjecture.

8.2 Linear Programs

The second part of this thesis studied information theoretic techniques to prove lower bounds on the non-negative rank of matrices and on the size of linear extended formulations of polytopes .

8.2.1 Positive Semidefinite Extension Complexity of the Matching Polytope

Chapter 7 looked at linear extended formulations for the Matching Polytope. The results of Rothvoß [Rot17], as well as the generalizations proved by Braun and Pokutta [BP15] and in Chapter 7 implied that any exact extended formulation for $P_{MAT}(n)$ has size $2^{\Omega(n)}$.

As briefly mentioned in Chapter 6, semidefinite extended formulations or SDP lifts give a way to write a polytope as a suitable projection of the feasible region of a semidefinite program. A major open question is to determine if there are positive semidefinite extended formulations for the Matching Polytope $P_{MAT}(n)$ that are of size polynomial in n.

Open Question 8.4. *Is there a positive semidefinite extended formulation for $P_{MAT}(n)$ that is of size polynomial in n?*

It turns out that this is closely related to proving lower bounds on the positive semidefinite rank of the matching slack matrix. For a non-negative matrix S, the *positive semidefinite rank* of the matrix S is defined to the smallest integer r such that there exist positive semidefinite matrices U_i, V_j of size $r \times r$ such that $S_{ij} = \mathsf{Trace}(U_i V_j)$.

We remark that the question of lower bounding the positive semidefinite rank has close connections to questions in quantum information theory and quantum communication complexity [LRS15, LPdWY16].

8.2.2 Inapproximability of the Max-Knapsack Polytope

Chapter 7 proved a lower bound of $\binom{n}{\Theta(1/\varepsilon)}$ on the size of linear extended formulations that give a $1+\varepsilon$ approximation for the Matching Polytope when $\varepsilon \geq \frac{2}{n}$. This is also tight as there is a $PTAS$-style linear program that gives such an approximation.

A similar situation arises for the Max-Knapsack polytope defined as

$$P_{KNAP}(n) = \mathsf{conv}\left\{ x \in \{0,1\}^n \;\middle|\; \sum_{i=1}^{n} a_i x_i \leq a_0 \right\},$$

where $a_i \geq 0$ for $0 \leq i \leq n$.

For the polytope $P_{KNAP}(n)$, Bienstock [Bieo8] gave an LP of size $\binom{n}{\Theta(1/\varepsilon)})$ which achieves a $(1-\varepsilon)$ approximation for any $\varepsilon \geq \frac{2}{n}$. An exponential lower bound for exact extended formulations for Max-Knapsack follows by a reduction from the unique disjointness matrix [AT13, PV13] (also see [GJW16a] which proves a better lower bound by using a different construction). It is unclear how to extend the aforementioned reduction from unique disjointness to prove a strong lower bound for any $(1-\varepsilon)$ approximation for Max-Knapsack. A lower bound even in the case of $(1-\frac{1}{n})$ approximation remains an interesting open problem.

Open Question 8.5. *Is there an $1-\frac{1}{n}$ approximate extended formulation for $P_{KNAP}(n)$ of size polynomial in n?*

Looking Forward

Information theoretic methods have proven useful in showing unconditional lower bounds for various models of computation, including among others, the settings of interactive compression and linear programs studied in this thesis. We hope that they will find further uses in proving stronger lower bounds for more general models of computation.

■

Vita

Makrand Sinha graduated with the degree of Bachelor of Technology in Computer Science and Engineering from the Indian Institute of Technology Kanpur in 2009. He received the degree of Master of Science in Informatik from ETH Zürich in 2011. Subsequently, he entered the University of Washington as a graduate student in the Paul G. Allen School of Computer Science & Engineering.

Bibliography

[ABC09] Chrisil Arackaparambil, Joshua Brody, and Amit Chakrabarti. Functional monitoring without monotonicity. In Susanne Albers, Alberto Marchetti-Spaccamela, Yossi Matias, Sotiris Nikoletseas, and Wolfgang Thomas, editors, *Automata, Languages and Programming*, volume 5555 of *Lecture Notes in Computer Science*, pages 95–106. Springer Berlin Heidelberg, 2009.

[AIP06] Alexandr Andoni, Piotr Indyk, and Mihai Patrascu. On the optimality of the dimensionality reduction method. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '06, pages 449–458, Washington, DC, USA, 2006. IEEE Computer Society.

[AMS99] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137 – 147, 1999.

[AT13] David Avis and Hans Raj Tiwary. On the Extension Complexity of Combinatorial Polytopes. In *Automata, Languages, and Programming - 40th International Colloquium, ICALP 2013, Riga, Latvia, July 8-12, 2013, Proceedings, Part I*, pages 57–68, 2013.

[ATYY17] Anurag Anshu, Dave Touchette, Penghui Yao, and Nengkun Yu. Exponential Separation of Quantum Communication and Classical Information. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 277–288, 2017.

[BBCR13] Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. How to Compress Interactive Communication. *SIAM J. Comput.*, 42(3):1327–1363, 2013.

[BEO+13] Mark Braverman, Faith Ellen, Rotem Oshman, Toniann Pitassi, and Vinod Vaikuntanathan. A tight bound for set disjointness in the message-passing model. In *FOCS*, pages 668–677, 2013.

[BFPS15a] Abbas Bazzi, Samuel Fiorini, Sebastian Pokutta, and Ola Svensson. No small linear program approximates vertex cover within a factor $2 - \varepsilon$. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 1123–1142, 2015.

[BFPS15b] Gábor Braun, Samuel Fiorini, Sebastian Pokutta, and David Steurer. Approximation Limits of Linear Programs (Beyond Hierarchies). *Math. Oper. Res.*, 40(3):756–772, 2015.

[BG14] Mark Braverman and Ankit Garg. Public vs Private Coin in Bounded-Round Information. In *ICALP*, pages 502–513, 2014.

[BGKR18] Mark Braverman, Anat Ganor, Gillat Kol, and Ran Raz. A candidate for a strong separation of information and communication. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, pages 11:1–11:13, 2018.

[Bie08] Daniel Bienstock. Approximate formulations for 0-1 knapsack sets. *Oper. Res. Lett.*, 36(3):317–320, 2008.

[BK18] Mark Braverman and Gillat Kol. Interactive compression to external information. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 964–977, 2018.

[BM13] Mark Braverman and Ankur Moitra. An information complexity approach to extended formulations. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 161–170, 2013.

[BMY15] Balthazar Bauer, Shay Moran, and Amir Yehudayoff. Internal compression of protocols to entropy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2015, August 24-26, 2015, Princeton, NJ, USA*, pages 481–496, 2015.

[BO03] Brian Babcock and Chris Olston. Distributed top-k monitoring. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, SIGMOD '03, pages 28–39, New York, NY, USA, 2003. ACM.

[BP15] Gábor Braun and Sebastian Pokutta. The Matching Problem Has No Fully Polynomial Size Linear Programming Relaxation Schemes. *IEEE Trans. Information Theory*, 61(10):5754–5764, 2015.

[BP16] Gábor Braun and Sebastian Pokutta. Common Information and Unique Disjointness. *Algorithmica*, 76(3):597–629, 2016.

[BR11] Mark Braverman and Anup Rao. Information Equals Amortized Communication. In *FOCS*, pages 748–757, 2011.

[Bra12] Mark Braverman. Interactive information complexity. In *STOC*, pages 505–524, 2012.

[BRWY13] Mark Braverman, Anup Rao, Omri Weinstein, and Amir Yehudayoff. Direct products in communication complexity. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 746–755, 2013.

[BW12] Mark Braverman and Omri Weinstein. A Discrepancy Lower Bound for Information Complexity. In *APPROX-RANDOM*, pages 459–470, 2012.

[BYJKS02] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An Information Statistics Approach to Data Stream and Communication Complexity. In *FOCS*, pages 209–218, 2002.

[CCZ10] Michele Conforti, Gérard Cornuéjols, and Giacomo Zambelli. Extended formulations in combinatorial optimization. *4OR*, 8(1):1–48, Mar 2010.

[CG05] Graham Cormode and Minos Garofalakis. Holistic aggregates in a networked world: Distributed tracking of approximate quantiles. In *SIGMOD*, pages 25–36, 2005.

[CGFS86] Fan R. K. Chung, Ronald L. Graham, Peter Frankl, and James B. Shearer. Some intersection theorems for ordered sets and graphs. *Journal of Combinatorial Theory, Series A*, 43(1):23–37, 1986.

[CKS03] Amit Chakrabarti, Subhash Khot, and Xiaodong Sun. Near-optimal lower bounds on the multi-party communication complexity of set disjointness. In *18th Annual IEEE Conference on Computational Complexity (Complexity 2003), 7-10 July 2003, Aarhus, Denmark*, pages 107–117, 2003.

[CLRS13] Siu On Chan, James R. Lee, Prasad Raghavendra, and David Steurer. Approximate Constraint Satisfaction Requires Large LP Relaxations. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 350–359, 2013.

[CMM09] Moses Charikar, Konstantin Makarychev, and Yury Makarychev. Near-optimal algorithms for maximum constraint satisfaction problems. *ACM Trans. Algorithms*, 5(3):32:1–32:14, 2009.

[CMY11] Graham Cormode, S. Muthukrishnan, and Ke Yi. Algorithms for distributed functional monitoring. *ACM Trans. Algorithms*, 7(2):21:1–21:20, March 2011.

[CMYZ10] Graham Cormode, S. Muthukrishnan, Ke Yi, and Qin Zhang. Optimal sampling from distributed streams. In *Proceedings of the Twenty-ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '10, pages 77–86, New York, NY, USA, 2010. ACM.

[CMZ06] G. Cormode, S. Muthukrishnan, and Wei Zhuang. What's different: Distributed, continuous monitoring of duplicate-resilient aggregates on data streams. In *Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference on*, pages 57–57, April 2006.

[Cor05] Graham Cormode. Sketching streams through the net: Distributed approximate query tracking. In *VLDB*, pages 13–24, 2005.

[CRR14] Arkadev Chattopadhyay, Jaikumar Radhakrishnan, and Atri Rudra. Topology matters in communication. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, pages 631–640, 2014.

[CSWY01] Amit Chakrabarti, Yaoyun Shi, Anthony Wirth, and Andrew Chi-Chih Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In *FOCS*, pages 270–278, 2001.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[DR98] Pavol Duris and Jose D.P. Rolim. Lower bounds on the multiparty communication complexity. *Journal of Computer and System Sciences*, 56(1):90 – 95, 1998.

[dW03] Ronald de Wolf. Nondeterministic quantum query and communication complexities. *SIAM Journal on Computing*, 32(3):681–699, 2003.

[Edm65] Jack Edmonds. Maximum Matching and a Polyhedron with $0,1$ Vertices. *J. of Res. the Nat. Bureau of Standards*, 69 B:125–130, 1965.

[Edm71] Jack Edmonds. Matroids and the greedy algorithm. *Mathematical Programming*, 1(1):127–136, Dec 1971.

[EJ08] Funda Ergun and Hossein Jowhari. On distance to monotonicity and longest increasing subsequence of a data stream. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '08, pages 730–736, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics.

[FJK+15] Lila Fontes, Rahul Jain, Iordanis Kerenidis, Sophie Laplante, Mathieu Laurière, and Jérémie Roland. Relative discrepancy does not separate information and communication complexity. In *Automata, Languages, and Programming - 42nd International Colloquium, ICALP 2015, Kyoto, Japan, July 6-10, 2015, Proceedings, Part I*, pages 506–516, 2015.

[FMP+15] Samuel Fiorini, Serge Massar, Sebastian Pokutta, Hans Raj Tiwary, and Ronald de Wolf. Exponential Lower bounds for Polytopes in Combinatorial Optimization. *J. ACM*, 62(2):17, 2015.

[FP16] Hamza Fawzi and Pablo A. Parrilo. Self-scaled bounds for atomic cone ranks: applications to nonnegative rank and cp-rank. *Math. Program.*, 158(1-2):417–465, 2016.

[FRPU94] Uriel Feige, Prabhakar Raghavan, David Peleg, and Eli Upfal. Computing with noisy information. *SIAM Journal of Computing*, 23(5):1001–1018, October 1994.

[GG10] Anna Gál and Parikshit Gopalan. Lower bounds on streaming algorithms for approximating the length of the longest increasing subsequence. *SIAM J. Comput.*, 39(8):3463–3479, August 2010.

[GH09] Sudipto Guha and Zhiyi Huang. Revisiting the direct sum theorem and space lower bounds in random order streams. In Susanne Albers, Alberto Marchetti-Spaccamela, Yossi Matias, Sotiris Nikoletseas, and Wolfgang Thomas, editors, *Automata, Languages and Programming*, volume 5555 of *Lecture Notes in Computer Science*, pages 513–524. Springer Berlin Heidelberg, 2009.

[GJW16a] Mika Göös, Rahul Jain, and Thomas Watson. Extension complexity of independent set polytopes. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 565–572, 2016.

[GJW16b] Mika Göös, Rahul Jain, and Thomas Watson. Extension Complexity of Independent Set Polytopes. *CoRR*, abs/1604.07062, 2016.

[GKR14] Anat Ganor, Gillat Kol, and Ran Raz. Exponential Separation of Information and Communication. In *FOCS*, pages 176–185, 2014.

[GKR15] Anat Ganor, Gillat Kol, and Ran Raz. Exponential Separation of Communication and External Information. *Electronic Colloquium on Computational Complexity (ECCC)*, 2015.

[GKR16] Anat Ganor, Gillat Kol, and Ran Raz. Exponential Separation of Information and Communication for Boolean Functions. *J. ACM*, 63(5):46:1–46:31, 2016.

[Goe15] Michel X. Goemans. Smallest compact formulation for the permutahedron. *Mathematical Programming*, 153(1):5–11, Oct 2015.

[Gro09] André Gronemeier. Asymptotically optimal lower bounds on the nih-multi-party information complexity of the and-function and disjointness. In *STACS 2009*, pages 505–516, 2009.

[Hås96] Johan Håstad. Clique is Hard to Approximate Within $n^{1-\varepsilon}$. In *37th Annual Symposium on Foundations of Computer Science, FOCS '96, Burlington, Vermont, USA, 14-16 October, 1996*, pages 627–636, 1996.

[HJMR10] Prahladh Harsha, Rahul Jain, David A. McAllester, and Jaikumar Radhakrishnan. The communication complexity of correlation. *IEEE Transactions on Information Theory*, 56(1):438–449, 2010.

[HRVZ15] Zengfeng Huang, Božidar Radunović, Milan Vojnović, and Qin Zhang. Communication complexity of approximate maximum matching in distributed graph data. In *STACS*, 2015.

[Huf52] David A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the Institute of Radio Engineers*, 40(9):1098–1101, September 1952.

[KCR06] Ram Keralapura, Graham Cormode, and Jeyashankher Ramamirtham. Communication-efficient distributed monitoring of thresholded counts. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, SIGMOD '06, pages 289–300, New York, NY, USA, 2006. ACM.

[Kha79] L. G. Khachiyan. A polynomial algorithm in linear programming. *Doklady Akademii Nauk SSSR*, 244:1093–1096, 1979.

[KLL+12] Iordanis Kerenidis, Sophie Laplante, Virginie Lerays, Jérémie Roland, and David Xiao. Lower Bounds on Information Complexity via Zero-Communication Protocols and Applications. In *FOCS*, pages 500–509, 2012.

[KMR17] Pravesh K. Kothari, Raghu Meka, and Prasad Raghavendra. Approximating rectangles by juntas and weakly-exponential lower bounds for LP relaxations of csps. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 590–603, 2017.

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

[Kol16] Gillat Kol. Interactive compression for product distributions. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 987–998, 2016.

[KS92] Bala Kalyanasundaram and Georg Schnitger. The Probabilistic Communication Complexity of Set Intersection. *SIAM Journal on Discrete Mathematics*, 5(4):545–557, 1992.

[LPdWY16] Troy Lee, Anupam Prakash, Ronald de Wolf, and Henry Yuen. On the Sum-of-Squares Degree of Symmetric Quadratic Functions. In *31st Conference on Computational Complexity, CCC 2016, May 29 to June 1, 2016, Tokyo, Japan*, pages 17:1–17:31, 2016.

[LRS15] James R. Lee, Prasad Raghavendra, and David Steurer. Lower Bounds on the Size of Semidefinite Programming Relaxations. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 567–576, 2015.

[Mar91] R.Kipp Martin. Using separation algorithms to generate mixed integer model reformulations. *Operations Research Letters*, 10(3):119 – 128, 1991.

[MNSW98] Peter Bro Miltersen, Noam Nisan, Shmuel Safra, and Avi Wigderson. On data structures and asymmetric communication complexity. *Journal of Computer and System Sciences*, 57(1):37 – 49, 1998.

[MSDO05] Amit Manjhi, Vladislav Shkapenyuk, Kedar Dhamdhere, and Christopher Olston. Finding (recently) frequent items in distributed data streams. In *Proceedings of the 21st International Conference on Data Engineering*, ICDE '05, pages 767–778, Washington, DC, USA, 2005. IEEE Computer Society.

[MWY13] Marco Molinaro, David P. Woodruff, and Grigory Yaroslavtsev. Beating the direct sum theorem in communication complexity with implications for sketching. In *SODA*, pages 1738–1756, 2013.

[Nis94] Noam Nisan. The communication complexity of threshold gates. In *In Proceedings of "Combinatorics, Paul Erdos is Eighty"*, pages 301–315, 1994.

[NR15] Sivaramakrishnan Natarajan Ramamoorthy and Anup Rao. How to Compress Asymmetric Communication. In *30th Conference on Computational Complexity, CCC 2015, June 17-19, 2015, Portland, Oregon, USA*, pages 102–123, 2015.

[NS15] Sivaramakrishnan Natarajan Ramamoorthy and Makrand Sinha. On the Communication Complexity of Greater-Than. In *53rd Annual Allerton Conference on Communication, Control and Computing*, pages 442–444, 2015.

[Pas12] Kanstantsin Pashkovich. *Extended Formulations for Combinatorial Polytopes*. PhD thesis, Magdeburg Universität, 2012.

[Pat11] Mihai Patrascu. Unifying the landscape of cell-probe lower bounds. *SIAM J. Comput.*, 40(3):827–847, 2011.

[Pok15] Sebastian Pokutta. Information theory and polyhedral combinatorics. In *53rd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2015, Allerton Park & Retreat Center, Monticello, IL, USA, September 29 - October 2, 2015*, pages 1119–1126, 2015.

[PR82] Manfred W. Padberg and M. R. Rao. Odd minimum cut-sets and b-matchings. *Mathematics of Operations Research*, 7(1):67–80, 1982.

[PV13] Sebastian Pokutta and Mathieu Van Vyve. A note on the extension complexity of the knapsack polytope. *Oper. Res. Lett.*, 41(4):347–350, 2013.

[PVZ12] Jeff M. Phillips, Elad Verbin, and Qin Zhang. Lower bounds for number-in-hand multiparty communication complexity, made easy. In *Proceedings of the Twenty-third Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '12, pages 486–501. SIAM, 2012.

[Rad52] Richard Rado. An inequality. *Journal of the London Mathematical Society*, 27:1 – 6, 1952.

[Rad03] Jaikumar Radhakrishnan. Entropy and Counting. In *Computational Mathematics, Modelling and Algorithms (Ed. J.C. Misra)*, pages 146–168. Narosa Publishing House, New Delhi, 2003.

[Raz92] A.A. Razborov. On the Distributional Complexity of Disjointness. *Theoretical Computer Science*, 106(2):385 – 390, 1992.

[Rot13] Thomas Rothvoß. Some 0/1 polytopes need exponential size extended formulations. *Math. Program.*, 142(1-2):255–268, 2013.

[Rot17] Thomas Rothvoss. The matching polytope has exponential extension complexity. *J. ACM*, 64(6):41:1–41:19, 2017.

[RS15] Anup Rao and Makrand Sinha. Simplified Separation of Information and Communication. *Electronic Colloquium on Computational Complexity (ECCC)*, 22:57, 2015.

[RS16] Anup Rao and Makrand Sinha. A Direct-Sum Theorem for Read-Once Branching Programs. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2016, September 7-9, 2016, Paris, France*, pages 44:1–44:15, 2016.

[RY18] Anup Rao and Amir Yehudayoff. *Communication Complexity*. Textbook (Draft), 2018.

[Sha48] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:623–666, October 1948.

[She16] Alexander A. Sherstov. Compressing interactive communication under product distributions. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:81, 2016.

[Sin18] Makrand Sinha. Lower Bounds for Approximating the Matching Polytope. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pages 1585–1604, 2018.

[SSK08] Izchak Sharfman, Assaf Schuster, and Daniel Keren. Shape sensitive geometric monitoring. In *Proceedings of the Twenty-seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '08, pages 301–310, New York, NY, USA, 2008. ACM.

[SSK10] Izchak Sharfman, Assaf Schuster, and Daniel Keren. Ubiquitous knowledge discovery. chapter A Geometric Approach to Monitoring Threshold Functions over Distributed Data Streams, pages 163–186. Springer-Verlag, Berlin, Heidelberg, 2010.

[Tur36] Alan M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 2(42):230–265, 1936.

[Vio15] Emanuele Viola. The communication complexity of addition. *Combinatorica*, 35(6):703–747, December 2015.

[Wil84] Richard M. Wilson. The exact bound in the Erdős-Ko-Rado theorem. *Combinatorica*, 4(2):247–257, 1984.

[Woo04] David Woodruff. Optimal space lower bounds for all frequency moments. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '04, pages 167–175, Philadelphia, PA, USA, 2004. Society for Industrial and Applied Mathematics.

[Wyn75] A.D. Wyner. The common information of two dependent random variables. *IEEE Transactions on Information Theory*, 21(2):163–179, 1975.

[WZ12] David P. Woodruff and Qin Zhang. Tight bounds for distributed functional monitoring. In *STOC*, pages 941–960, 2012.

[WZ14] David P. Woodruff and Qin Zhang. An optimal lower bound for distinct elements in the message passing model. In *SODA*, pages 718–733, 2014.

[Yan91] Mihalis Yannakakis. Expressing Combinatorial Optimization Problems by Linear Programs. *J. Comput. Syst. Sci.*, 43(3):441–466, 1991.