KTH Computer Science
and Communication

Conditional Inapproximability and Limited Independence

PER AUSTRIN

Doctoral Thesis
Stockholm, Sweden 2008

TRITA-CSC-A 2008:18
ISSN-1653-5723
ISRN-KTH/CSC/A--08/18--SE
ISBN 978-91-7415-179-4

KTH Datavetenskap och kommunikation
SE-100 44 Stockholm
SWEDEN

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen i datalogi fredagen den 28 november 2008 klockan 13.00 i D3, Lindstedtsvägen 5, Kungl Tekniska högskolan, Stockholm.

Tryck: Universitetsservice US AB

Abstract

Understanding the theoretical limitations of efficient computation is one of the most fundamental open problems of modern mathematics. This thesis studies the *approximability* of intractable optimization problems. In particular, we study so-called Max CSP problems. These are problems in which we are given a set of constraints, each constraint acting on some k variables, and are asked to find an assignment to the variables satisfying as many of the constraints as possible.

A predicate $P : [q]^k \to \{0, 1\}$ is said to be *approximation resistant* if it is intractable to approximate the corresponding CSP problem to within a factor which is better than what is expected from a completely random assignment to the variables. We prove that if the Unique Games Conjecture is true, then a sufficient condition for a predicate $P : [q]^k \to \{0, 1\}$ to be approximation resistant is that there exists a pairwise independent distribution over $[q]^k$ which is supported on the set of satisfying assignments $P^{-1}(1)$ of P.

We also study predicates $P : \{0, 1\}^2 \to \{0, 1\}$ on two boolean variables. The corresponding CSP problems include fundamental computational problems such as Max Cut and Max 2-Sat. For any P, we give an algorithm and a Unique Games-based hardness result. Under a certain geometric conjecture, the ratios of these two results are shown to match exactly. In addition, this result explains why additional constraints beyond the standard "triangle inequalities" do not appear to help when solving these problems. Furthermore, we are able to use the generic hardness result to obtain improved hardness for the special cases of Max 2-Sat and Max 2-And. For Max 2-Sat, we obtain a hardness of $\alpha_{LLZ} + \epsilon \approx 0.94016$, where α_{LLZ} is the approximation ratio of the algorithm due to Lewin, Livnat and Zwick. For Max 2-And, we obtain a hardness of 0.87435. For both of these problems, our results surprisingly demonstrate that the special case of balanced instances (instances where every variable occurs positively and negatively equally often) is not the hardest. Furthermore, the result for Max 2-And also shows that Max Cut is not the hardest 2-CSP.

Motivated by the result for k-CSP problems, and their fundamental importance in computer science in general, we then study t-wise independent distributions with random support. We prove that, with high probability, $\mathrm{poly}(q) \cdot n^2$ random points in $[q]^n$ can support a pairwise independent distribution. Then, again with high probability, we show that $(\mathrm{poly}(q) \cdot n)^t \log(n^t)$ random points in $[q]^n$ can support a t-wise independent distribution. For constant t and q, we show that $\Omega(n^t)$ random points are necessary in order to be able to support a t-wise independent *balanced* distribution with non-negligible probability. Also, we show that *every* subset of $[q]^n$ with size at least $q^n(1 - \mathrm{poly}(q)^{-t})$ can support a t-wise independent distribution.

Finally, we prove a certain noise correlation bound for low-degree functions with small Fourier coefficients. This type of result is generally useful in hardness of approximation, derandomization, and additive combinatorics.

Acknowledgments

Many years have passed since I started as a PhD student, not being sure what it meant or what I was supposed to do, but enticed by all the glory and attention received by computer scientists (or maybe it was something else—I can't remember). Now, several years and many more failures later, I am a very different person. A little less naive, a few beers heavier, and several hours of missed sleep more tired. There are many people without whom this would not have happened.

A $1 - \epsilon$ fraction of the acknowledgments go to my advisor, and now co-author, Johan Håstad. Johan is one of few people of whom I am in true awe, partly due to his great no-nonsense personality and vast knowledge of many things, but mostly due to his uncanny ability to immediately understand any idea, however complicated it might be. Should I ever find myself being a tenth of the scientist Johan is, I will consider myself lucky.

Many thanks also to my co-author Elchanan Mossel at UC Berkeley and the Weizmann Institute. I have learned a lot from Elchanan, who is a great person with something interesting to say on virtually any subject, and I have found his enthusiasm for "abstract nonsense" highly contagious.

Thanks to Luca Trevisan at UC Berkeley for having me as a visitor during the spring semester of 2008. I really enjoyed my time there, and already miss the foggy San Francisco mornings. Thanks also to other people who I have visited for shorter periods of time: Subhash Khot at New York University, Avner Magen at the University of Toronto and Rafael Pass at Cornell University.

It has been great fun to work with all the nice people in the theory group at KTH: I have had lots of inane arguments, pencil wars, and general fun with office mate and co-author Gunnar Kreitz, together with whom I have more than once driven our poor next door neighbor Mikael Goldmann nuts. There was also an office mate called Fredrik Niemelä, but nobody knows what happened to him (rumor has it he was eaten by industry). I am also encouraged to see the spirit of room 1445 live on in the recent additions to the theory group, Torbjörn Granlund and Douglas Wikström, both being almost as obstinate as myself.

Some people actually took the time to look at one or more of the many preliminary versions of this thesis. I am very grateful to Mikael Goldmann, Johan Håstad, Gunnar Kreitz, Elchanan Mossel and Jakob Nordström for their valuable comments.

I am very grateful to my family for letting me go my own way in life and not questioning why on earth I would choose to be a PhD student rather than make money, and for not asking "when will you be done?" too often. Non-academic friends are also jolly good to have, and I am particularly happy with the ones I have, and grateful to them for all the great fun I have had with them.

Finally, thanks to Frida for being who she is, for putting up with my sometimes very tenuous connection with reality and general absentmindedness, and for her relentless support.

Contents

Part I

Introduction

This thirsts for knowledge, yet how is it bought?
With many a sleeplesse night and racking thought?

John Hall – On An Hourglass

Chapter 1

What is This Thesis About?

This chapter gives a non-technical introduction to the subjects of this thesis, intended to be understandable for people not necessarily sharing the same background in—and enthusiasm for—theoretical computer science as the author.

1.1 Computing

A large part of this thesis involves, in some way or another, the concept of *computing*. What exactly does this word mean? While most people are familiar with it, few probably know an exact definition. Thus, let us begin this thesis with the perhaps somewhat boring task of looking up *computing* in Webster's dictionary.

How do we perform this task? There are several different ways. One way, which most people would refer to as naive, or maybe even stupid, would be to start at page one and then scan the pages of the dictionary one by one until the sought word is found. A second way, which is similar to what we actually do in practice, is to open the dictionary somewhere roughly in the middle. On the page we open, we might find words like, "mathematics" or "metaphysical", and deduce that "computing" must appear somewhere in the earlier part of the dictionary, before that page. Then, we can flip to a new page, this time somewhere roughly in the middle of that first part. Based on the words on this new page, we can again decide whether "computing" occurs earlier or later in the dictionary. We then continue in this fashion until we come across a page where we find the word we're looking for. This way, we can fairly quickly find any word we seek in any dictionary, even a dictionary containing several million words.

After flipping some pages in Webster's dictionary, we come across the following:

> **compute** *vt* to determine mathematically; to calculate by means of a computer.

Hence, *computing* is the process of mathematically determining some fact. Such a fact may be almost anything, from the fact that $23 \times 17 = 391$, to the fact that the

shortest path from point A to point B goes via point C. The specific process by which one determines that something is a fact, is known as an *algorithm*. The two different ways described earlier for finding a word in the dictionary are examples of two different *algorithms* for the *computational problem* of searching in a sorted list.

Computational complexity is a branch of computer science in which one studies how much *computational resources* are required to solve a problem. The most important computational resource is *time*—the more time you have at hand, the more problems you can solve. Another important resource is *memory*, though it will not be relevant in this thesis. In order to determine how much time is required to solve a problem, one can either exhibit an *upper bound*, by giving an algorithm using only this or that much time, or one can exhibit a *lower bound*, by proving that *every* possible algorithm for solving the problem must use at least this or that much time.

A first natural question to ask might be: "can every computational problem be solved?" The perhaps somewhat surprising answer is: "no". For instance, suppose that you would like to write a computer program which analyzes other computer programs in order to determine whether they can crash or not. Clearly, such a program must be quite difficult to build, since otherwise someone would have done so by now and we would not have to put up with the incessant crashing of our computers any more. However, it turns that it is not only difficult, but even mathematically impossible to actually write such a program! Somewhat informally, the reason for this is that one can prove that, for any such program, there would be cases when it would need to run for an infinite amount of time, which clearly is a lot longer than we are willing to wait.

1.2 The Million Dollar Question

A second natural question to ask might be: "given some problem which *can* be solved, can it be solved using a reasonable amount of resources?". This brings us to one of the most fundamental concepts in computer science, that of *efficient computation*.

Let us now consider a different problem. Suppose that you are student and need to decide which courses to take for the next semester. There are some courses you are interested in, but unfortunately, the schedules of some of these collide, so you will not be able to take all of them. In order to keep CSN[1] happy, you need to take at least k courses, otherwise they will stop giving you money. Is it possible for you to take only courses that you are interested in, or are you going to have to take some additional courses which you are not interested in, just in order to meet CSN's requirements?

Can we construct an efficient algorithm which is guaranteed to find the best set of courses to take? It may sound surprising, but answering this question is actually worth one million dollars! In computer science jargon, this question is known as

[1] The Swedish National Board of Student Aid.

the "P vs. NP" question. Who are these "P" and "NP", why are they fighting each other, and why are they worth enough money to buy a big apartment in central Stockholm?

To answer this, we must first elaborate on what we mean by "efficient algorithm". We say that an algorithm $\mathcal{A}$ is efficient, if there is some number x such that, if we increase the size of the input to $\mathcal{A}$ by 1%, the running time of $\mathcal{A}$ increases by at most x%. Another way of characterizing this kind of performance is to say that the running time grows *polynomially* in the size of the input.

P is the family of all decision problems for which there exists an efficient algorithm. The name P is simply an acronym for "Polynomial time". This includes problems such as deciding whether there is a short path from point A to point B in a map, and deciding whether a given number n is a prime or not.

The definition of NP, on the other hand, is a bit tricksier[2]. It is *not*, as one might be tempted to guess in light of the definition of P, simply an acronym of "Not Polynomial". Rather, NP is the family of all decision problems for which, *if the answer is "yes", then that fact can be efficiently verified*. By "verifying" that the answer is "yes", we mean that there is some "certificate" of this fact which we can look at to convince ourselves that the answer is indeed "yes". For example, our course selection problem is in NP: if the answer is "yes", i.e., if there are k courses which do not collide, then a list of those courses constitutes a certificate. We can efficiently verify it by checking that there are at least k courses in the list, and that no two courses in the list collide. Formulated in a mathematical terminology, one can think of P as the class of statements which are easy to prove, and of NP as the class of all statements for which, if they are true, there is a proof which can be easily checked. A great introduction to P vs. NP for mathematicians can be found in [104].

Clearly, every problem in P is also in NP—if it is easy to compute whether the answer is "yes" or "no", then one can verify that the answer is "yes" simply by computing the answer and ignoring whatever certificate is given. The P vs. NP question asks whether the other direction holds, i.e., whether every problem in NP is also in P. P vs. NP is one of the seven so-called Millennium Problems announced by the Clay Math Institute in 2000[3].

Why then, is the question about efficient solutions to our course selection problem the same as the P vs. NP question? The course selection problem is in fact a disguised formulation of a well-known problem called the Independent Set problem. This problem is one of the so-called NP-*hard* problems. To understand the importance of NP-hardness, one has to understand the computer scientist's love for *reductions*. It turns out that, as one might guess from the 1 million dollar prize purse of the P vs. NP problem, it is quite difficult to prove that there does not exist any efficient algorithm for a given problem. The way computer scientists cope

[2]Like hobbitses.

[3]`http://www.claymath.org/millennium/`. It should be pointed out that one of the problems, the *Poincaré Conjecture* from topology, was recently solved by Grigori Perelman.

with this, is by reductions. Loosely speaking, reductions are ways of relating the difficulty of one problem to the difficulty of another, rather than to the amount of resources needed to solve the problem. In particular, we are very good at saying things of the form "If this problem can be solved efficiently, then all these other problems can too", or conversely, "If any of these problems can not be solved efficiently, then this problem can not be solved either". If an NP-hard problem can be efficiently solved, then *every* problem in NP can be efficiently solved, and P = NP. Today, literally hundreds of problems are known to be NP-hard, many of them problems which are quite important in practice, such as different types of scheduling and routing problems. The general consensus is that P ≠ NP, but currently, we are very far away from proving such a thing.

1.3 Solutions for the Inaccurately Minded

The Max Independent Set problem is the variant of Independent Set where we are asked to find an independent set (i.e., a list of non-colliding courses) which is as large as possible. In the previous section, we were only asking if there was a set of size larger than some given number k. Clearly, Max Independent Set is even harder than Independent Set—if we are not able to determine whether the maximum is smaller or larger than k, then we can not hope to compute it. Given that the Max Independent Set problem is NP-hard, it is natural to ask whether the problem becomes feasible if we content ourselves with finding an independent set which is not necessarily of maximum size, but just within, say, 90%, of maximum size. This type of algorithm, which does not necessarily find the best answer, but at least finds something which is *guaranteed* to be close to the best answer, is known as an *approximation algorithm*.

It turns out that Max Independent Set is not only NP-hard, it is even NP-hard to approximate within 90%, or even 1%, or even 0.0001% (or even, for those familiar with the terminology, within $2^{(\log n)^{3/4+\epsilon}}/n$).

It turns out that different NP-hard problems have very different characteristics when it comes to approximability. Some problems, such as Maximum Independent Set, can almost not be approximated at all, whereas others can be approximated to within almost arbitrarily small error. Understanding the approximability of different natural combinatorial optimization problems has been a very active area of computer science in the last 15 years, after the discovery of something known as the PCP Theorem[4]. The work of this thesis is, either directly or indirectly, related to this search for understanding of the theoretical limitations of efficient computation.

[4]No, the name is not related to the kind of PCP which sometimes appears in the movies.

1.4 Organization and Contents

This thesis has two main parts. Part II gives several inapproximability results for a certain type of constraint satisfaction problems. Part III is more loosely connected, and contains two results which both relate to k-wise independent probability distributions. As large portions of the background material necessary for the two parts are disjoint, each of these parts has a separate "Preliminaries" chapter, giving the background necessary for that part. In addition, the next chapter, Chapter 2, gives preliminaries required for both parts, and introduces much of the notation used throughout the thesis.

Some of the results in this thesis have appeared previously in a different form. In particular, the results in Part II are based on three papers. Chapter 5 is based on "Approximation Resistant Predicates From Pairwise Independence" [9], coauthored with Elchanan Mossel, which appeared at the IEEE Conference on Computational Complexity, in 2008. Chapter 6 is based on two papers. The first is "Balanced Max 2-Sat Might Not Be the Hardest" [7], which appeared at the ACM Symposium on Theory of Computing in 2007. The second is "Towards Sharp Inapproximability for Any 2-CSP" [8], which appeared at the IEEE Symposium on Foundations of Computer Science in 2007.

The results in Part III are more recent, and have not yet been published elsewhere. Chapter 8 is based on a collaboration with my advisor, Johan Håstad. Chapter 9 is based on a collaboration with Elchanan Mossel.

Chapter 2

Preliminaries

This chapter introduces notation, contains some background material, and describes some results that will be useful for us. As the chapter covers a fairly large amount of material in a quite small number of pages, the reader who does not want to become saturated with definitions may choose to skip ahead and return to it later when the need arises. To assist this, here are some pointers to when the different parts of this chapter will be needed. Sections 2.1 and 2.2 contain notation used throughout the entire thesis, though reading Section 2.2.3 may be postponed until one starts reading the latter parts of this chapter, i.e., Section 2.3 and onwards. Section 2.3 is primarily used in Chapter 4 and in Part III. Sections 2.4 and 2.5 are first used towards the end of Chapter 4, and will then be used in the remaining chapters of Part II as well as in Chapter 9.

2.1 Sets, Vectors, Functions, Inner and Outer Products

We use the following notation for various frequently occurring sets.

Symbol	Meaning
$\mathbb{R}$	The real numbers
$\mathbb{Z}$	The integers
$\mathbb{N}$	The natural numbers, i.e., $\{ n \in \mathbb{Z} : n \geq 1 \}$
$[n]$	The integers from 1 to n, i.e., $\{1, 2, \ldots, n\}$
$\mathbb{Z}_n$	The integers from 0 to $n-1$, i.e., $\{0, 1, \ldots, n-1\}$
$\mathbb{F}_q$	The finite field with q elements (for q a prime power)

Table 2.1: Standard sets

For two sets X and Y, X^Y denotes the set of all vectors over X indexed by Y. For the case when $Y = [n]$, we write X^n rather than $X^{[n]}$. We will, in general, make no distinction between X^Y and the set of functions $f : Y \to X$, and will use whichever notation we find most convenient for the task at hand.

For a vector $v \in X^Y$ and $S \subseteq Y$, $v_S \in X^S$ is the *projection* of v to the coordinates in S (i.e., for every $i \in S$, the $v_S(i) = v(i)$). We make no distinction between $x_{\{i\}}$ and x_i, even though, syntactically, $x_{\{i\}}$ is a function from $\{i\}$ to X, whereas x_i is an element of X.

For X^Y and $x \in X$, we often use $\mathbf{x}$ to denote the vector in X^Y in which all entries are identically equal to x, i.e., $\mathbf{x}_y = x$ for every $y \in Y$. In particular, $\mathbf{0} \in \mathbb{R}^Y$ denotes the all-zeros vector, and $\mathbf{1} \in \mathbb{R}^Y$ denotes the all-ones vector. Note that this is not quite well-defined since $\mathbf{x}$ has no reference to the index set Y, but this will always be clear from the context. We sometimes also use $\mathbf{1}_{[\mathsf{statement}]}$ to denote an indicator function of whether "statement" is true. For instance, if $f : X \to Y$ is a function from X to Y and $a \in Y$ is an element of Y, $\mathbf{1}_{[f=a]} : X \to \{0,1\}$ is the indicator function

$$\mathbf{1}_{[f=a]}(x) = \begin{cases} 1 & \text{if } f(x) = a \\ 0 & \text{otherwise.} \end{cases} .$$

For two vectors $u, v \in \mathbb{R}^X$, we denote by $\langle u, v \rangle_{\mathbb{R}}$ the standard *inner product* of u and v,

$$\langle u, v \rangle_{\mathbb{R}} = \sum_{x \in X} u_x \cdot v_x.$$

For functions $f : A \to \mathbb{R}$ and $g : B \to \mathbb{R}$, $f \otimes g : A \times B \to \mathbb{R}$ denotes the (outer) *tensor product* of f and g, defined by

$$(f \otimes g)(a, b) = f(a) \cdot g(b).$$

We use $f^{\otimes n} : A^n \to \mathbb{R}$ to denote the n-fold tensor product of f with itself,

$$f^{\otimes n} = \underbrace{f \otimes f \otimes \ldots \otimes f}_{n \text{ times}}.$$

For functions $f : X \to Y$ and $g : Y \to Z$, we denote by $g \circ f : X \to Z$ the *composition* of f with g, $(g \circ f)(x) = g(f(x))$. In particular, if $x \in X^n$ is a string of length n, and $\pi : [n] \to [n]$ is a permutation, $x \circ \pi \in X^n$ denotes x permuted by π, i.e.,

$$x \circ \pi = x_{\pi(1)} x_{\pi(2)} \ldots x_{\pi(n)}.$$

2.2 Probability Theory

In this thesis we will concern ourselves with two "types" of probability spaces: distributions over some finite domain Ω, and the standard Gaussian distribution over $\mathbb{R}^d$. In this section we will describe the basic notation and facts that we will use about such distributions.

2.2.1 A Small Formal Note

We shall not be completely formal in our treatment of these spaces, and in particular we shall not talk about the underlying σ-algebras of the spaces, as these will always be the "standard" σ-algebras associated with the domain—the complete σ-algebra in the case of finite domains, and the Borel σ-algebra in the case of $\mathbb{R}^n$.

Consequently, for a domain Ω and probability density function $\mu : \Omega \to [0, 1]$, we will use (Ω, μ) (or sometimes even Ω when μ is clear from the context) to denote the space $(\Omega, \mathscr{A}, \mathbf{P})$, where $\mathscr{A}$ is the "standard" σ-algebra of Ω, and $\mathbf{P} : \mathscr{A} \to [0, 1]$ is given by

$$\mathbf{P}(S) = \int_{x \in S} \mu(x),$$

the integral being with respect to the "standard" measure over Ω—the Lebesgue measure in the case of $\mathbb{R}^n$, and the counting measure when Ω is finite.

2.2.2 Basic definitions

Let (Ω, μ) be a probability space. A random variable over (Ω, μ) is a function $f : \Omega \to \mathbb{R}$. In most parts of this thesis, the latter view will be the most convenient one, and we will explicitly talk about functions rather than random variables, but we will still use some of the notation used for random variables and, e.g., write $\mathbb{E}[f]$ rather than the more cumbersome $\mathbb{E}_{x \in (\Omega,\mu)}[f(x)]$ for the expected value of f.

For $1 \leq p < \infty$ we define the ℓ_p norm of $f : \Omega \to \mathbb{R}$ as

$$||f||_p = \left(\mathbb{E}[|f|^p]\right)^{1/p}.$$

For $p = \infty$, we define $||f||_\infty = \max_{\mu(x)>0} |f(x)|$. A basic fact about ℓ_p norms is that they are increasing in p.

Fact 2.2.1. For $1 \leq p \leq q \leq \infty$ and $f : \Omega \to \mathbb{R}$,

$$||f||_p \leq ||f||_q.$$

We use $L^2(\Omega, \mu)$ to denote the set of all functions $f : \Omega \to \mathbb{R}$ such that $||f||_2 < \infty$ (for Ω finite, $L^2(\Omega, \mu)$ consists of all functions $f : \Omega \to \mathbb{R}$). We endow $L^2(\Omega, \mu)$ with the inner product

$$\langle f, g \rangle_\mu = \mathop{\mathbb{E}}_{x \in (\Omega,\mu)}[f(x) \cdot g(x)].$$

In many cases, the probability space (Ω, μ) will be clear from the context, and in this case we will drop the subscript μ and simply write $\langle f, g \rangle$.

When (Ω, μ) is finite, we denote by $\alpha(\mu)$ the minimum non-zero probability of any atom, i.e.,

$$\alpha(\mu) = \min_{\substack{x \in \Omega \\ \mu(x)>0}} \mu(x).$$

We denote by

$$\mathrm{Cov}[f,g] = \mathbb{E}[(f - \mathbb{E}[f])(g - \mathbb{E}[g])] = \mathbb{E}[fg] - \mathbb{E}[f]\,\mathbb{E}[g]$$

the *covariance* between f and g, by

$$\mathrm{Var}[f] = \mathrm{Cov}[f,f] = \mathbb{E}[f^2] - \mathbb{E}[f]^2 = ||f - E[f]||_2^2$$

the *variance* of f, and by

$$\tilde{\rho}_{f,g} = \frac{\mathrm{Cov}[f,g]}{\sqrt{\mathrm{Var}[f]\,\mathrm{Var}[g]}}$$

the *correlation coefficient* between f and g.

Theorem 2.2.2 (Hölder's Inequality)**.** *Let $1 \leq p \leq q \leq \infty$ be such that $1/p + 1/q = 1$. Then*

$$\langle f, g \rangle \leq ||f||_p \cdot ||g||_q.$$

Two easy but important corollaries of Hölder's Inequality are the "Repeated Hölder's Inequality" and the Cauchy-Schwarz inequality.

Corollary 2.2.3 (Repeated Hölder's Inequality)**.** *Let $f_1, \ldots, f_k \in L^2(\Omega, \mu)$, and $\sum_{i=1}^k 1/p_i = 1$, where $1 \leq p_i \leq \infty$ for each i. Then*

$$\mathbb{E}\left[\prod_{i=1}^k f_i\right] \leq \prod_{i=1}^k ||f_i||_{p_i}.$$

Corollary 2.2.4 (Cauchy-Schwarz' Inequality)**.** *For every $f, g \in L^2(\Omega, \mu)$ we have*

$$\langle f, g \rangle \leq ||f||_2 \cdot ||g||_2.$$

2.2.3 Product Spaces and Correlation

In most parts of this thesis, we will be working with probability spaces (Ω, μ) in which the domain Ω is the Cartesian product of n domains $\Omega = \Omega_1 \times \ldots \times \Omega_n$. We will refer to these spaces as product spaces.

For a product space (Ω, μ), and a set $S \subseteq [n]$, we denote by $\mu|_S$ the marginal measure of μ, restricted to $\prod_{i \in S} \Omega_i$. Formally, for Ω finite and $x \in \prod_{i \in S} \Omega_i$,

$$\mu|_S(x) = \sum_{\substack{y \in \Omega^n \\ y_S = x}} \mu(y).$$

When $S = \{i\}$, we write $\mu|_i$ rather than $\mu|_S$.

Many times, the distribution μ over Ω^n will simply be a product distribution $\mu = \bigotimes_{i=1}^n \mu_i$ for some distributions μ_i over Ω_i. I.e., $(\Omega_1 \times \ldots \times \Omega_n, \mu_1 \otimes \ldots \otimes \mu_n)$

is the probability space over Ω in which the density of an atom $(x_1, \ldots, x_n) \in \Omega$ is given by $\prod_{i=1}^{n} \mu_i(x_i)$. These distributions are called product distributions.

We will also need to work with product spaces (Ω, μ) where μ is not a product distribution and there is some dependence between the coordinates of Ω. In the remainder of this section will introduce some terminology for such spaces.

Let $(\Omega_1 \times \Omega_2, \mu)$ be a product space and $f \in L^2(\Omega_1, \mu|_1)$, $g \in L^2(\Omega_2, \mu|_2)$ be two functions. We define the correlation coefficient $\tilde{\rho}_{f,g}$ by viewing f as a function on $(\Omega_1 \times \Omega_2, \mu)$ which depends only on the first coordinate, and g as a function on $(\Omega_1 \times \Omega_2, \mu)$ which depends only on the second coordinate. Formally, define $\tilde{f}, \tilde{g} \in L^2(\Omega_1 \times \Omega_2, \mu)$ by $\tilde{f}(x, y) = f(x)$ and $\tilde{g}(x, y) = g(y)$. Then we define

$$\tilde{\rho}_{f,g} = \tilde{\rho}_{\tilde{f},\tilde{g}} = \frac{\mathbb{E}_{(x,y)\in(\Omega_1\times\Omega_2,\mu)}[f(x)g(y)] - \mathbb{E}[f]\,\mathbb{E}[g]}{\sqrt{\mathrm{Var}[f]\,\mathrm{Var}[g]}}.$$

A notion which will be very important for us is that of the correlation of a product space, introduced in [78], which is defined as follows.

Definition 2.2.5. Let $(\Omega_1 \times \Omega_2, \mu)$ be a product space. The *correlation* $\tilde{\rho}(\Omega_1, \Omega_2, \mu)$ of $(\Omega_1 \times \Omega_2, \mu)$ is defined by

$$\tilde{\rho}(\Omega_1, \Omega_2, \mu) = \sup_{\substack{f \in L^2(\Omega_1, \mu|_1) \\ g \in L^2(\Omega_2, \mu|_2)}} \tilde{\rho}_{f,g}.$$

Suppose (x, y) is a sample from the space $(\Omega_1 \times \Omega_2, \mu)$. Intuitively, the correlation $\tilde{\rho}(\Omega_1, \Omega_2, \mu)$ measures how much information you can get about y by being given x, or vice versa. In particular, if $\tilde{\rho} = 0$, x and y are completely independent (i.e., μ is a product distribution). On the other hand, if $\tilde{\rho} = 1$, there exist non-trivial partitions $S_1 \cup \overline{S}_1 = \Omega_1$ and $S_2 \cup \overline{S}_2 = \Omega_2$ such that whenever $x \in S_1$, we also have $y \in S_2$ (seeing this is not quite trivial, but it is a consequence of Lemma 2.2.7 below).

The definition of $\tilde{\rho}$ is extended to product spaces on many coordinates as follows.

Definition 2.2.6. Let $(\Omega_1 \times \ldots \times \Omega_n, \mu)$ be a product space. The *correlation* $\tilde{\rho}(\Omega_1, \ldots, \Omega_n, \mu)$ is

$$\tilde{\rho}(\Omega_1, \ldots, \Omega_n, \mu) = \max_{1 \le i \le n} \tilde{\rho}(\Omega_i, \prod_{j \neq i} \Omega_j, \mu).$$

A useful condition for $\tilde{\rho}$ being strictly smaller than 1, which is usually needed, is the following, which follows from [78], Lemma 2.9.

Lemma 2.2.7. *Let (Ω, μ) be a finite product space with $\Omega = \Omega_1 \times \ldots \times \Omega_k$, and consider the graph $G = (V, E)$ defined as follows. The vertices are the elements $V = \{\, a \in \Omega : \mu(a) > 0 \,\}$ in Ω with positive probability, and there is an edge from $a = (a_1, \ldots, a_k)$ to $a' = (a'_1, \ldots, a'_k)$ if a and a' differ in exactly one coordinate. Then, if G is connected, we have*

$$\tilde{\rho}(\Omega_1, \ldots, \Omega_k, \mu) \le 1 - \alpha(\mu)^2/2.$$

(b_1, b_2)	$\mu(b_1, b_2)$
$(1, 1)$	$\frac{1+\xi_1+\xi_2+\rho}{4}$
$(1, -1)$	$\frac{1+\xi_1-\xi_2-\rho}{4}$
$(-1, 1)$	$\frac{1-\xi_1+\xi_2-\rho}{4}$
$(-1, -1)$	$\frac{1-\xi_1-\xi_2+\rho}{4}$

Table 2.2: The distribution μ

In particular, if $\mu(x) > 0$ for every $x \in \Omega$, $\tilde{\rho}(\Omega_1, \ldots, \Omega_k, \mu) < 1$.

In Chapter 6, we will work with the special case of $\Omega_1 = \Omega_2 = \{-1, 1\}$. In particular, let $(\{-1, 1\}^2, \mu)$ be a probability space on pairs of bits such that

- The expected value of the first bit is ξ_1.
- The expected value of the second bit is ξ_2.
- The expected value of the product of the bits is ρ.

The parameters ξ_1, ξ_2, and ρ completely determine any distribution μ over $\{-1, 1\}^2$ (see Table 2.2).

Proposition 2.2.8. *Let $(\{-1, 1\}^2, \mu)$ be as in Table 2.2. Then*

$$\tilde{\rho}(\{-1, 1\}, \{-1, 1\}, \mu) = \left| \frac{\rho - \xi_1 \xi_2}{\sqrt{1 - \xi_1^2}\sqrt{1 - \xi_2^2}} \right|.$$

Proof. Since correlation coefficients are invariant under translation and scaling, we can without loss of generality take $\tilde{\rho}(\{-1, 1\}, \{-1, 1\}, \mu)$ as the supremum over $\mathbb{E}[fg]$ for $f \in L^2(\{-1, 1\}, \mu|_1)$ and $g \in L^2(\{-1, 1\}, \mu|_2)$ with $\mathbb{E}[f] = \mathbb{E}[g] = 0$ and $\mathrm{Var}[f] = \mathrm{Var}[g] = 1$. But any function on $\{-1, 1\}$ is determined uniquely (up to sign) by its expectation and variance. In particular, the only two functions on $(\{-1, 1\}, \mu|_1)$ with expectation 0 and variance 1 are f and $-f$, where

$$f(x_1) = \frac{x_1 - \mathbb{E}[x_1]}{\sqrt{\mathrm{Var}[x_1]}} = \frac{x_1 - \xi_1}{\sqrt{1 - \xi_1^2}} = \begin{cases} \sqrt{-\frac{1+\xi_1}{1-\xi_1}} & \text{if } x_1 = -1 \\ \sqrt{\frac{1-\xi_1}{1+\xi_1}} & \text{if } x_1 = 1 \end{cases},$$

and similarly for $(\{-1, 1\}, \mu|_2)$. Thus,

$$\mathbb{E}[fg] = \pm \frac{\mathbb{E}[(x_1 - \xi_1)(x_2 - \xi_2)]}{\sqrt{1 - \xi_1^2}\sqrt{1 - \xi_2^2}} = \pm \frac{\rho - \xi_1 \xi_2}{\sqrt{1 - \xi_1^2}\sqrt{1 - \xi_2^2}}.$$

and hence the supremum over all f and g is as claimed. □

Another important notion is that of k-wise independence. The study of k-wise independent variables goes back at least 30 years [69, 59, 81]. They were first used in computer science in the work of Alon et al. [2], and have since seen many applications, in particular in derandomization. See [74] for a survey.

Definition 2.2.9. A product space (Ω^n, μ) is *k-wise independent with marginals* η (for some probability distribution η over Ω), if, for every subset $S \subseteq [n]$ of at most k indices, we have that $\mu|_S = \eta^{\otimes |S|}$ (up to an appropriate identification of the indices).

Put differently, (Ω^n, μ) is k-wise independent if, for every t indices $i_1 < i_2 < \ldots < i_t$, and $a_1, \ldots, a_t \in \Omega$, we have that

$$\Pr_{x \in (\Omega^n, \mu)}[x_{i_1} = a_1, x_{i_2} = a_2, \ldots, x_{i_t} = a_t] = \prod_{i=1}^{t} \eta(a_i).$$

When the marginal distribution η is the uniform distribution over Ω, we say that (Ω^n, μ) is *balanced* k-wise independent. We can of course define k-wise independence more generally for an arbitrary product space $\Omega_1 \times \ldots \times \Omega_n$ with some specified marginal distributions $\eta_1, \ldots, \eta_n$, but to keep the exposition simple, we restrict ourselves to the case Ω^n with all marginals equal.

2.2.4 Gaussian Space

For $x \in \mathbb{R}$, we denote by $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and $\Phi(x) = \int_{t=-\infty}^{x} \phi(t) \mathrm{d}t$ the density and distribution functions of a standard normal variable. A *standard normal vector* $r \in \mathbb{R}^n$ is an n-dimensional vector in which every entry is an independent standard normal variable.

Fact 2.2.10. Let $v_1, v_2 \in \mathbb{R}^n$, and let r be standard normal vector in $\mathbb{R}^n$. Then $x_1 = \langle v_1, r \rangle_{\mathbb{R}}$ and $x_2 = \langle v_2, r \rangle_{\mathbb{R}}$ are jointly normal variables with covariance matrix

$$\begin{pmatrix} \langle v_1, v_1 \rangle_{\mathbb{R}} & \langle v_1, v_2 \rangle_{\mathbb{R}} \\ \langle v_2, v_1 \rangle_{\mathbb{R}} & \langle v_2, v_2 \rangle_{\mathbb{R}} \end{pmatrix}.$$

We make the following definition for bivariate normal distributions, which undoubtedly looks somewhat cumbersome, but will be convenient for us to work with.

Definition 2.2.11. Let $\rho \in [-1, 1]$ and let X_1, X_2 be jointly normal variables with $E[X_1] = 0$ and $E[X_2] = 0$, and covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. For $\mu_1, \mu_2 \in [-1, 1]$, we define

$$\Gamma_\rho(\mu_1, \mu_2) = \Pr\left[X_1 \leq \Phi^{-1}\left(\frac{1-\mu_1}{2}\right) \wedge X_2 \leq \Phi^{-1}\left(\frac{1-\mu_2}{2}\right)\right].$$

In Section 3.6, we will study Γ_ρ in more detail and give some of its properties which are used in Chapter 6.

2.3 Harmonic Analysis

Informally, harmonic analysis is a branch of mathematics in which one seeks to decompose functions into sums of some "nicely behaved" functions. An example is the classic Fourier transform of a periodic function over $\mathbb{R}$, in which these "nice" functions are wave functions. In this thesis, the functions of interest will be random variables over some product space $(\Omega^n, \mu^{\otimes n})$, and the "nice" basis functions will be functions that one can think of as multilinear monomials on n variables.

2.3.1 Fourier Decomposition

Let (Ω, μ) be a finite probability space with $|\Omega| = q$, which is non-degenerate in the sense that $\mu(x) > 0$ for every $x \in \Omega$. Let $\chi_0, \ldots, \chi_{q-1} : \Omega \to \mathbb{R}$ be an orthonormal basis for the space $L^2(\Omega, \mu)$ w.r.t. the scalar product $\langle \cdot, \cdot \rangle_\mu$. Furthermore, let this basis be such that $\chi_0 = \mathbf{1}$, i.e., the function that is identically 1 on every element of Ω.

For $\sigma \in \mathbb{Z}_q^n$, define $\chi_\sigma : \Omega^n \to \mathbb{R}$ as $\bigotimes_{i \in [n]} \chi_{\sigma_i}$, i.e.,

$$\chi_\sigma(x_1, \ldots, x_n) = \prod_{i \in [n]} \chi_{\sigma_i}(x_i).$$

Fact 2.3.1. The functions $\{\chi_\sigma\}_{\sigma \in \mathbb{Z}_q^n}$ form an orthonormal basis for the product space $L^2(\Omega^n, \mu^{\otimes n})$.

Proof. For $\sigma, \sigma' \in \mathbb{Z}_q^n$, we have

$$\langle \chi_\sigma, \chi_{\sigma'} \rangle_{\mu^{\otimes n}} = \mathop{\mathbb{E}}_{x \in (\Omega^n, \mu^{\otimes n})} \left[\prod_{i=1}^n \chi_{\sigma_i}(x_i) \chi_{\sigma'_i}(x_i) \right] = \prod_{i=1}^n \left\langle \chi_{\sigma_i}, \chi_{\sigma'_i} \right\rangle_\mu ,$$

which if $\sigma_i \neq \sigma'_i$ for some i, equals 0, and otherwise equals 1, by the orthonormality of $\chi_0, \ldots, \chi_{q-1}$. Finally, it is clear that $|\{\chi_\sigma \,|\, \sigma \in \mathbb{Z}_q^n\}| = q^n = \dim(L^2(\Omega^n, \mu^{\otimes n}))$, and hence they form a basis. □

Thus, every function $f \in L^2(\Omega^n, \mu^{\otimes n})$ can be written as

$$f(x) = \sum_{\sigma \in \mathbb{Z}_q^n} \hat{f}(\sigma) \chi_\sigma(x),$$

where $\hat{f} : \mathbb{Z}_q^n \to \mathbb{R}$ is defined by $\hat{f}(\sigma) = \langle f, \chi_\sigma \rangle_{\mu^{\otimes n}}$. The most basic properties of $\hat{f}$ are summarized by Fact 2.3.2, which is an immediate consequence of the orthonormality of $\{\chi_\sigma\}_{\sigma \in \mathbb{Z}_q^n}$.

Fact 2.3.2. We have

$$\mathbb{E}[fg] = \sum_\sigma \hat{f}(\sigma)\hat{g}(\sigma) \qquad \mathbb{E}[f] = \hat{f}(\mathbf{0}) \qquad \mathrm{Var}[f] = \sum_{\sigma \neq \mathbf{0}} \hat{f}(\sigma)^2.$$

An example of this transform which is widely used in computer science is the Fourier-Walsh transform (for which there are many different names—the names Hadamard transform or simply Fourier transform are also commonly used). Here, $\Omega = \{-1, 1\}$ and μ is the uniform distribution (and hence, $(\Omega^n, \mu^{\otimes n})$ is the n-dimensional boolean hypercube with the uniform distribution). In this case, we have $\chi_1(x) = x$, and every function $f : \{-1, 1\}^n \to \mathbb{R}$ can be decomposed as

$$f(x) = \sum_{S \subseteq [n]} \hat{f}(S) \prod_{i \in S} x_i,$$

i.e., the basis functions in the decomposition are exactly the 2^n multilinear monomials over the variables $x_1, \ldots, x_n$.

As far as we are aware, there is no standard name for the transform $f \mapsto \hat{f}$ for general product spaces and bases. Since it is in some sense a very general type of Fourier transform, we are simply going to refer to it as the Fourier transform, and $\hat{f}$ as the Fourier coefficients of f. We remark that the article "the" is somewhat inappropriate, since the transform and coefficients in general depend on the choice of basis $\{\chi_i\}_{i \in \mathbb{Z}_q}$. However, in this thesis, we will always be working with some fixed (albeit arbitrary) basis, and hence there should be no ambiguity in referring to the Fourier transform as if it were unique. Furthermore, as we shall see, most of the important properties of $\hat{f}$ are actually basis-independent.

Before proceeding, let us introduce some useful notation for the index set $\mathbb{Z}_q^n$ of the Fourier coefficients.

Definition 2.3.3. A *multi-index* is a vector $\sigma \in \mathbb{Z}_q^n$, for some q and n. The *active set* of a multi-index is $S(\sigma) = \{\, i : \sigma_i > 0 \,\}$. We extend notation defined for $S(\sigma)$ to σ in the natural way, and write e.g. $|\sigma|$ instead of $|S(\sigma)|$, $i \in \sigma$ instead of $i \in S(\sigma)$, and so on.

Another fact which is sometimes useful is the following trivial bound on the ℓ_∞ norm of χ_σ (recall that $\alpha(\mu)$ is the minimum non-zero probability of any atom in μ).

Fact 2.3.4. Let $(\Omega^n, \mu^{\otimes n})$ be a product space with Fourier basis $\{\chi_\sigma\}_{\sigma \in \mathbb{Z}_q^n}$. Then for any $\sigma \in \mathbb{Z}_q^n$,

$$||\chi_\sigma||_\infty \leq \left(\frac{1}{\alpha(\mu)}\right)^{|\sigma|/2}.$$

To see this, note that $||\chi_\sigma||_\infty = \prod_{i \in \sigma} ||\chi_{\sigma_i}||_\infty$, and that $\chi_{\sigma_i}(x) \leq 1/\sqrt{\alpha}$ for every $x \in \Omega$, since otherwise $||\chi_{\sigma_i}||_2$ would exceed 1.

2.3.2 Efron-Stein Decomposition

In this section, we describe a somewhat "coarser" decomposition of $f \in L^2(\Omega^n, \mu^{\otimes n})$ than the Fourier decomposition.

Theorem 2.3.5. *Any $f \in L^2(\Omega^n, \mu^{\otimes n})$ can be* uniquely *decomposed as a sum of functions*

$$f(x) = \sum_{S \subseteq [n]} f_S(x),$$

where

- *$f_S(x)$ depends only on $x_S = (x_i : i \in S)$*
- *For every $S \subseteq [n]$, for every S' which does not contain S and $y_{S'} \in \Omega^{S'}$, it holds that*

  $$\mathbb{E}[f_S(x) \,|\, x_{S'} = y_{S'}] = 0.$$

  In other words, whenever we condition on some variables $x_{S'}$, the expected value of f_S is going to be 0 as long as we have not conditioned on all the variables that f_S depend on.

This decomposition is known as the Efron-Stein decomposition [29] (see also [78], Definition 2.10). It is easily verified that it relates to the Fourier decomposition as follows.

Proposition 2.3.6. *Fix an arbitrary Fourier basis $\{\chi_\sigma\}_{\sigma \in \mathbb{Z}_q^n}$ for $(\Omega^n, \mu^{\otimes n})$. Then, for any $f \in L^2(\Omega^n, \mu^{\otimes n})$, the Efron-Stein decomposition $f = \sum f_S$ of f can be written as*

$$f_S(x) = \sum_{\substack{\sigma \in \mathbb{Z}_q^n \\ S(\sigma) = S}} \hat{f}(\sigma) \chi_\sigma(x). \tag{2.1}$$

Proving this is just a matter of verifying that the functions f_S defined as in Equation (2.1) satisfy the conditions in Theorem 2.3.5, and that they sum up to f.

This means that, in general, properties of f defined in terms of its Fourier coefficients in a "nice" way, will be independent of the choice of Fourier basis. For instance, any expression of the form $\sum_{S(\sigma) \in \mathcal{F}} \hat{f}(\sigma)\chi_\sigma$, where $\mathcal{F}$ is some family of subsets of $[n]$, is independent of the choice of Fourier basis, as it equals $\sum_{S \in \mathcal{F}} f_S$. In particular also the sums of squares of Fourier coefficients involved in such sums are invariant under the choice of basis.

2.3.3 Degree and Influences

Definition 2.3.7. The *degree* $\deg(f)$ of $f \in L^2(\Omega^n, \mu^{\otimes n})$ is the infimum of all $d \in \mathbb{Z}$ such that $\hat{f}(\sigma) = 0$ for all σ with $|\sigma| > d$.

The degree of f is one of its most important properties. In general, the smaller $\deg(f)$ is, the more nicely behaved f is. When $\deg(f) \leq d$, we will refer to f as a *degree-d polynomial* in $L^2(\Omega^n, \mu^{\otimes n})$.

Definition 2.3.8. For $f : \Omega^n \to \mathbb{R}$ and $d \in \mathbb{Z}$, the function $f^{\leq d} : \Omega^n \to \mathbb{R}$ is defined by

$$f^{\leq d} = \sum_{|\sigma| \leq d} \hat{f}(\sigma)\chi_\sigma.$$

We define $f^{<d}$, $f^{=d}$, $f^{>d}$ and $f^{\geq d}$ analogously.

Next, we define the important notion of influence. As the name suggests, the influence of the ith variable on $f \in L^2(\Omega^n, \mu^{\otimes n})$ measures how much f can change if the value of the ith variable is changed and all other variables are fixed.

Definition 2.3.9. The *influence* of i on $f \in L^2(\Omega^n, \mu^{\otimes n})$ is

$$\mathrm{Inf}_i(f) = \mathop{\mathbb{E}}_{x_{[n]\setminus i}}\left[\mathop{\mathrm{Var}}_{x_i}[f(x)]\right].$$

We sometimes refer to variables with "large" influence (where the exact value of "large" can differ but usually means bounded from below by some constant independent of n) as influential, and to functions without influential variables as low-influence functions.

It turns out that the influence of a function has a particularly nice characterization in terms of its Fourier coefficients.

Proposition 2.3.10. *For every $f \in L^2(\Omega^n, \mu^{\otimes n})$,*

$$\mathrm{Inf}_i(f) = \sum_{\substack{\sigma \in \mathbb{Z}_q^n \\ i \in \sigma}} \hat{f}(\sigma)^2.$$

Proof. Define $f_0, f_1 : \Omega^n \to \mathbb{R}$ as

$$f_0 = \sum_{\substack{\sigma \in \mathbb{Z}_q^n \\ i \notin \sigma}} \hat{f}(\sigma)\chi_\sigma \qquad\qquad f_1 = \sum_{\substack{\sigma \in \mathbb{Z}_q^n \\ i \in \sigma}} \hat{f}(\sigma)\chi_\sigma,$$

i.e., f_0 is the part of f which does not depend on x_i, and f_1 is the part which depends on x_i. For $x \in \Omega^n$, we can then write

$$\mathop{\mathrm{Var}}_{x_i}[f(x) \,|\, x_{[n]-i}] = \mathop{\mathrm{Var}}_{x_i}[f_1(x) \,|\, x_{[n]-i}] = \mathop{\mathbb{E}}_{x_i}[f_1(x)^2 \,|\, x_{[n]-i}],$$

where the first equality holds since f_0 does not depend on x_i, and the second equality holds since $\mathbb{E}_{x_i}[f_1(x) \,|\, x_{[n]-i}] = 0$. Thus, averaging over all values of $x_{[n]-i}$, we have

$$\mathrm{Inf}_i(f) = \mathbb{E}[f_1^2] = \sum_{\sigma \in \mathbb{Z}_q^n} \hat{f}_1(\sigma)^2 = \sum_{\substack{\sigma \in \mathbb{Z}_q^n \\ i \in \sigma}} \hat{f}(\sigma)^2. \qquad \square$$

While the influences of a function are an important property, they will not be a central part in this thesis. However, the closely related property of low-degree influence, defined next, is going to play a crucial role in the inapproximability results obtained in Part II.

Definition 2.3.11. The *d-degree influence* of i on $f \in L^2(\Omega^n, \mu^{\otimes n})$ is defined by

$$\mathrm{Inf}_i^{\leq d}(f) = \mathrm{Inf}_i(f^{\leq d}).$$

We often omit the explicit reference to d and simply refer to d-degree influence as *low-degree influence*.

Note that, by Proposition 2.3.10, we can write

$$\mathrm{Inf}_i^{\leq d}(f) = \sum_{\substack{\sigma \in \mathbb{Z}_q^n \\ i \in \sigma \\ |\sigma| \leq d}} \hat{f}(\sigma)^2.$$

The key property of low-degree influence which makes it useful in the context of hardness of approximation is that the number of variables with large low-degree influence is always bounded.

Proposition 2.3.12. *For any $f \in L^2(\Omega^n, \mu^{\otimes n})$, the number of variables $i \in [n]$ such that*

$$\mathrm{Inf}_i^{\leq d}(f) \geq \tau$$

is at most $\frac{d}{\tau}\mathrm{Var}[f]$.

Proof. The total low-degree influence in all variables can be written as

$$\sum_{i=1}^{n} \mathrm{Inf}_i^{\leq d}(f) = \sum_{\substack{\sigma \in \mathbb{Z}_q^n \\ |\sigma| \leq d}} \sum_{i \in \sigma} \hat{f}(\sigma)^2 = \sum_{k=1}^{d} k \cdot ||f^{=k}||_2^2 \leq d\,\mathrm{Var}[f]. \qquad \square$$

In particular, if $f : \Omega^n \to [-1, 1]$, we have $\mathrm{Var}[f] \leq 1$ and hence the number of variables with d-degree influence at least τ is bounded by d/τ.

2.4 Noise Correlation

In this section we introduce the notion of noise correlation.

Various special cases of noise correlation has been the focus of much work, as we discuss below. Informally, the noise correlation between two functions f and g measure how much $f(x)$ and $g(y)$ correlate on random inputs x and y which are correlated. We remark that the name "noise correlation" is a slight misnomer and that "correlation under noise" would be a more descriptive name—we are not looking at how well a random variable correlates with noise, but rather how well two random variables correlate with each other in the presence of noise.

Definition 2.4.1. Let (Ω, μ) be a product space with $\Omega = \Omega_1 \times \ldots \times \Omega_k$, and let $f_1, \ldots, f_k$ be functions with $f_i \in L^2((\Omega_i)^n, (\mu|_i)^{\otimes n})$. The *noisy inner product*, or *noise correlation*, of $f_1, \ldots, f_k$ with respect to μ is

$$\langle f_1, f_2, \ldots, f_k \rangle_{\mathscr{N}} = \mathbb{E}\left[\prod_{i=1}^{k} f_i\right].$$

As it can take some time to get used to Definition 2.4.1, let us write out $\langle f_1, \ldots, f_k \rangle_{\mathscr{N}}$ more explicitly. Let $f_i : \Omega_i^n \to \mathbb{R}$ be functions on the product space Ω_i^n, and let μ be some probability distribution on $\Omega = \Omega_1 \times \ldots \times \Omega_k$. Then,

$$\langle f_1, \ldots, f_k \rangle_{\mathscr{N}} = \mathop{\mathbb{E}}_{X}\left[\prod_{i=1}^{k} f_i(X_i)\right],$$

where X is a $k \times n$ random matrix such that each column of X is a sample from (Ω, μ), independently of the other columns, and X_i refers to the ith row of X.

The notation $\langle f_1, \ldots, f_k \rangle_{\mathscr{N}}$ is new for this thesis, but such quantities arise naturally in many different settings. They are also of central interest in the recent work of Mossel [78] and its applications [9, 87]. In the remainder of this section, we will briefly mention two particularly interesting special cases from two different areas of mathematics.

One important special case of noise correlation is *noise sensitivity*, introduced by Benjamini et al. [14]. The noise sensitivity $\mathsf{NS}_\epsilon(f)$ of f at ϵ is the answer to the following question: suppose we pick a uniformly random point $x \in \{-1, 1\}^n$, and then perturb x by flipping each bit with probability ϵ, obtaining a point $y \in \{-1, 1\}^n$. What is the probability that $f(x) \neq f(y)$? Noise sensitivity is closely related to *noise stability*. The noise stability $\mathbb{S}_\rho(f)$ of f at $\rho \in [-1, 1]$ is $\mathbb{S}_\rho(f) = \mathbb{E}_{x,y}[f(x)f(y)]$, where x is a uniformly random string, and y is obtained by flipping each bit of x with probability $(1 - \rho)/2$, independently (so that the expected value of each bit $x_i y_i$ is ρ). It is easily verified that

$$\mathsf{NS}_\epsilon(f) = \frac{1 - \mathbb{S}_{1-2\epsilon}(f)}{2}.$$

Also, for an appropriate choice of (Ω, μ), we have $\mathbb{S}_\rho(f) = \langle f, f \rangle_{\mathscr{N}}$. There has been a lot of work on noise sensitivity, partly because of applications in computer science and the theory of social choice [82, 63, 79], but perhaps even more so because of applications to the study of so-called *crossing probabilities* in percolation theory [14, 97, 40].

A second important special case of noise correlation is the *Gowers norm* from additive combinatorics, introduced by Gowers [44] in a Fourier-analytic proof of a seminal theorem by Szemerédi [100]. Let $f : \{0, 1\}^n \to \mathbb{R}$ be a function on the boolean hypercube, and let $d \geq 1$ be an integer. Then, the degree-d Gowers norm

of f is defined by

$$||f||_{U^d}^d = \mathop{\mathbb{E}}_{x,x_1,\ldots,x_d} \left[\prod_{S \subseteq [d]} f\left(x + \sum_{i \in S} x_i \right) \right],$$

where $x, x_1, \ldots, x_d$ are independent uniformly random elements of $\{0,1\}^n$, and "+" in $\{0,1\}^n$ is interpreted as componentwise addition in $\mathbb{Z}_2^n$ (i.e., XOR). The Gowers norm, which is indeed a norm, enjoys many interesting properties, and since its introduction there has been much work aimed at obtaining a better understanding of it [46, 47, 72]. In our notation, letting $k = 2^d$, and $(\Omega_1 \times \ldots \times \Omega_k, \mu)$ be a suitably chosen product space, the Gowers norm can be written as $\langle f, \ldots, f \rangle_{\mathcal{N}}^{1/k}$.

2.5 Noise Correlation Bounds

In this section, we review a family of powerful results which have been discovered in recent years. These results give good bounds on noisy inner products of functions in various settings. The first such result was the Majority Is Stablest Theorem by Mossel et al. [79], sometimes also called the MOO Theorem after its authors. This theorem, which essentially deals with $\mathbb{S}_\rho(f)$, was first conjectured by Khot et al. in [63], where it was shown that it implied that the famous Goemans-Williamson approximation algorithm for the MAX CUT problem is optimal under the Unique Games Conjecture (in Chapter 3 we will talk more about these matters). Subsequently, various minor extensions of the MOO Theorem to slightly more general settings appeared in different applications [28, 27, 8]. Recently, Mossel [78] gave a wider generalization which already has found very interesting applications in hardness of approximation, e.g., the result in Chapter 5 and Raghavendra's result [87] connecting integrality gaps of semidefinite programs to hardness under the Unique Games Conjecture. These noise correlation bounds also have interesting applications in the theory of social choice (in the context of so-called Condorcet paradoxes) and in additive combinatorics.

For the result in Chapter 5, we will need the result stated in Theorem 2.5.1 below. It essentially says that if $f_1, \ldots, f_k$ do not have influential variables, then $\langle f_1, \ldots, f_k \rangle_{\mathcal{N}}$ under some pairwise independent distribution μ, is close to what it would be if μ was completely uniform. Put differently, low-influence functions can not "distinguish" pairwise independence from true independence.

Theorem 2.5.1 ([78], Theorem 6.6)**.** *Let (Ω, μ) be a finite probability space over $\Omega = \prod_{i=1}^k \Omega_i$ with the following properties:*

(a) μ is pairwise independent.

(b) For every $a \in \Omega$, $\mu(a) > 0$.

Then for every $\epsilon > 0$ there exists constants $\tau > 0$ and $d > 0$, depending only on ϵ and the minimum probability $\alpha(\mu)$, such that the following holds. Let $f_1, \ldots, f_k$ be

functions $f_i \in L^2(\Omega_i^n, (\mu|_i)^{\otimes n})$ *with* $f_i(x) \in [0,1]$ *for all* x*, satisfying that, for all* $1 \leq j \leq n$*,*

$$|\{\, i \,:\, \mathrm{Inf}_j^{\leq d}(f_i) \geq \tau \,\}| \leq 2.$$

Then

$$\left| \langle f_1, \ldots, f_k \rangle_{\mathscr{N}} - \prod_{i=1}^{k} \mathbb{E}\left[f_i\right] \right| \leq \epsilon.$$

The statement of [78] is somewhat stronger, it does not require $\mu(a) > 0$ for every $a \in \Omega$, only that $\rho(\Omega_1, \ldots, \Omega_k, \mu) < 1$ (which, by Lemma 2.2.7 is a weaker condition). However, the form of Theorem 2.5.1 will be sufficient for our applications.

For the result in Chapter 6, we use the following statement, bounding $\langle f_1, f_2 \rangle_{\mathscr{N}}$ in terms of estimates for normally distributed variables with the same correlation. This theorem can be viewed as a slight generalization of the Majority is Stablest Theorem to arbitrary product distributions over $\{-1,1\}^2$, and follows from the original MOO Theorem. See [8], Theorem 2.18 and Corollary 2.19 for proofs. It can also be viewed as a special case of [78], Theorem 6.3.

Theorem 2.5.2 ([8], Corollary 2.19)**.** *Let* $(\{-1,1\}^2, \mu)$ *be a finite probability space, and let* $\tilde{\rho} := \tilde{\rho}(\{-1,1\}, \{-1,1\}, \mu) < 1$*. Then for every* $\epsilon > 0$ *there exist constants* $\tau > 0$ *and* $d > 0$ *depending only on* ϵ *and* $\alpha(\mu)$ *such that the following holds.*

Let f_1, f_2 *be functions* $f_i \in L^2(\{-1,1\}^n, (\mu|_i)^{\otimes n})$ *with* $f_i(x) \in [-1,1]$ *for all* x*, satisfying that, for all* $1 \leq j \leq n$*,*

$$\max(\mathrm{Inf}_j^{\leq d}(f_1), \mathrm{Inf}_j^{\leq d}(f_2)) \leq \tau.$$

Then

$$\Gamma_{-\tilde{\rho}}(\mathbb{E}[f], \mathbb{E}[g]) - \epsilon \leq \langle f_1, f_2 \rangle_{\mathscr{N}} + \mathbb{E}[f] + \mathbb{E}[g] - 1 \leq \Gamma_{\tilde{\rho}}(\mathbb{E}[f], \mathbb{E}[g]) + \epsilon.$$

In Chapter 9, we obtain a different noise correlation bound, for functions with no large Fourier coefficients (note that any function with small influences also has all Fourier coefficients small). It is our hope that this result may find applications in fields where [78, 79] have been useful, such as inapproximability and additive combinatorics. Unfortunately, the settings in which this result works are significantly more limited than those of [78, 79], so it is currently not clear whether such applications are possible.

At the heart of both Theorem 2.5.1 and Theorem 2.5.2 lies what is known as an invariance principle. An invariance principle is, very loosely speaking, a statement to the effect that some function A of some random variables $f_1, \ldots, f_n$ behaves "the same" if $f_1, \ldots f_n$ are replaced by some other random variables $g_1, \ldots g_n$ having a different distribution. To make this concrete, an example of this which one encounters in a first course on probability is the Central Limit Theorem (CLT). The CLT asserts that if $x_1, \ldots, x_n$ is a sequence of i.i.d. random variables with finite mean μ and variance σ^2, then $A(x_1, \ldots, x_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i$ converges in distribution

to $A(g_1, \ldots, g_n)$ of n i.i.d. normal random variables $g_1, \ldots, g_n$ with mean μ and variance σ^2, as n tends to ∞, which in turn is simply a normal random variable g with mean μ and variance σ^2. The invariance principles behind Theorem 2.5.1 and Theorem 2.5.2 can be viewed as generalizations of CLT to low-degree polynomials with no small influences (or rather of quantitative versions of the CLT such as the Berry-Esséen Theorem). Note that the average $A(x_1, \ldots, x_n)$ is an extreme case of such a polynomial, as it has degree 1 and tiny influences. We remark that an invariance principle similar to the one of [78] was discovered already in 1979 by Rotar [91], though with worse error bounds and without the truncation arguments that make it possible to say something about arbitrary functions and not just low-degree polynomials.

Part II

Some Conditional Inapproximability Results

Are you familiar with the old robot saying, "`Does not compute`*"?*

Bender Bending Rodríguez – Futurama Season 5 Episode 4

Chapter 3

Preliminaries

In this section, we give some background material necessary for the results of Chapter 5 and Chapter 6. Most of this material is about approximation algorithms and hardness of approximation.

3.1 Constraint Satisfaction Problems

We assume some familiarity with combinatorial constraint satisfaction problems, see e.g. Chapter 1 of [85] for an extensive treatment. For combinatorial optimization problems in general, we always write $\mathrm{Opt}(\Psi)$ to denote the optimum value of an instance Ψ, and $\mathrm{Val}_\Psi(a)$ to denote the value of a feasible solution a for Ψ. When the instance Ψ is clear from context as it usually is, we omit the subscript Ψ from Val_Ψ and simply write $\mathrm{Val}(a)$.

Broadly speaking, a constraint satisfaction problem is a problem in which one is given a set of constraints acting on a set of variables, and seeks to find an assignment to the variables so as to maximize the number of satisfied constraints.

A very basic example of a constraint satisfaction problem is the Max Cut problem, in which we are given an undirected graph $G = (V, E)$, and seek to find a partition $V = U \cup \overline{U}$ such that the number of edges cut by U, i.e., the number of edges between U and $\overline{U}$, is maximized. In this problem there is a variable $x_v \in \{0, 1\}$ for each vertex $v \in V$, and for each edge (u, v) there is a constraint on x_u and x_v which is satisfied if $x_u \neq x_v$.

A very important class of constraint satisfaction problems are the Max k-CSP$_q$ problems. We will discuss various special cases of them in more detail later, but first, let us define the problem in its full generality.

Definition 3.1.1. Let k and q be positive integers. Then, an instance Ψ of the Max k-CSP$_q$ problem is a tuple $\Psi = (\mathcal{C}, \mathrm{wt})$, where $\mathcal{C}$ is a set of constraints and $\mathrm{wt} : \mathcal{C} \to [0, 1]$ assigns a weight to each constraint in $\mathcal{C}$. We assume that wt is normalized so that $\sum_{C \in \mathcal{C}} \mathrm{wt}(C) = 1$.

Each constraint is a function $C : [q]^S \to [0,1]$, where $S := S(C) \subseteq [n]$ is a set of $|S| = k$ variables, which this particular constraint acts on. The value of an assignment $a \in [q]^n$ is given by

$$\mathrm{Val}_\Psi(a) = \sum_{C \in \mathcal{C}} \mathrm{wt}(C) C(a_S).$$

The optimum of Ψ is the maximum value of any assignment,

$$\mathrm{Opt}(\Psi) = \max_{a \in [q]^n} \mathrm{Val}_\Psi(a).$$

The Max k-CSP$_q$ problem is NP-hard whenever $k \geq 2$ and $q \geq 2$. In the case when $q = 2$, we drop the subscript q and simply call it the Max k-CSP problem. We remark that it is common to define a constraint as a function into $\{0,1\}$ rather than $[0,1]$ as we do. We take the more general route, as our hardness results apply also in this setting. We sometimes also refer to a constraint function $C : [q]^S \to [0,1]$ as an *objective function*.

An important special case of Max k-CSP$_q$ are the Max CSP(P) problems.

Definition 3.1.2. Let $P : \{0,1\}^k \to [0,1]$ be an objective function.

The Max CSP(P) problem is the special case of the Max k-CSP problem in which each constraint $C : \{0,1\}^S \to \{0,1\}$ is of the form $P(l_1, \ldots, l_k)$ for some literals $l_1, \ldots, l_k$, where each literal is either a variable or a negated variable. In other words, C is completely specified by the set $S \subseteq [n]$ and a "sign" vector $s \in \{0,1\}^n$, viz.

$$C(x_S) = P(x_S \oplus s),$$

where $\oplus$ denotes coordinate-wise Xor.

The Max CSP$^+$(P) problem is the special case of the Max CSP(P) problem in which only positive literals are used, i.e., where the sign vector is always $\mathbf{0}$. In other words, it is the special case of the Max k-CSP problem in which each constraint C is of the form $P(x_1, \ldots, x_k)$ for some set of k variables.

Many fundamental computational problems can be cast as Max CSP problems. For instance, Max Cut is exactly the Max CSP$^+$($\oplus_2$) problem, where $\oplus_2 : \{0,1\}^2 \to \{0,1\}$ is the Xor predicate on two variables. Another famous example is the Max 3-Sat problem, which is exactly the Max CSP($\vee_3$) problem, where $\vee_3 : \{0,1\}^3 \to \{0,1\}$ is the Or predicate on three variables. Let us give some formal definitions.

Definition 3.1.3. We define the following special cases of Max k-CSP.

Max k-Xor is the Max CSP($\oplus_k$) problem, where $\oplus_k(x_1, \ldots, x_k)$ is the Xor predicate.

Max k-And is the Max CSP($\wedge_k$) problem, where $\wedge_k(x_1, \ldots, x_k)$ is the And predicate.

Max k-Sat is the Max CSP($\vee_k$) problem, where $\vee_k(x_1, \ldots, x_k)$ is the Or predicate.

For objective functions $P : [q]^k \to [0,1]$ on larger domains, we can define MAX CSP(P) and MAX CSP$^+$(P) similarly. For the definition of MAX CSP(P), there are several natural ways of generalizing the notion of a literal. One possible definition is to say that a literal l is of the form $\pi(x_i)$, for some variable x_i and permutation $\pi : [q] \to [q]$. A more restrictive definition is to say that a literal is of the form $x_i + b$, where, again, x_i is a variable, $b \in [q]$ is some constant, and $+$ is interpreted as taking place modulo q. In this thesis, we use the second, more restrictive, definition. As this is a special case of the first definition, our hardness results apply also to the first definition.

A special case of the MAX 2-CSP$_q$ problem which is very important in hardness of approximation is the Label Cover problem. It is important because it often provides a very good starting point when one wants to prove that some other problem is hard to approximate. We will elaborate further on this point later, towards the end of Section 3.3.

Definition 3.1.4. For an integer $L > 0$, an instance Ψ of the L-LABEL COVER problem is a tuple $\Psi = (X, Y, E, \Pi)$, where $(X \cup Y, E)$ is a bipartite graph, and $\Pi = \{\pi_e\}_{e \in E}$ associates to each edge $e \in E$ a function $\pi_e : [L] \to [L]$.

A *labeling* of Ψ is a function $\ell : X \cup Y \to [L]$. An edge $e = (x, y)$ is *satisfied* by ℓ if $\ell(y) = \pi_e(\ell(x))$, and the value of ℓ is the fraction of edges satisfied by ℓ,

$$\mathrm{Val}(\ell) = \frac{1}{|E|} |\{\, e \in E \mid \ell \text{ satisfies } e \,\}|.$$

The optimum of Ψ is the maximum value of any labeling,

$$\mathrm{Opt}(\Psi) = \max_{\ell : X \cup Y \to [L]} \mathrm{Val}(\ell).$$

3.2 Approximation and Inapproximability

Almost since the discovery of NP-completeness, there has been an interest in how well NP-hard problems can be approximated. A classic example of this is the TRAVELING SALESPERSON PROBLEM (TSP) in a metric space, in which one seeks the minimum total distance one has to travel to visit some specified set of points. It is an easy exercise to prove that in a metric space, the minimum cost of a TSP tour is at least the cost c of a minimum spanning tree, and it is easy to construct a tour of cost at most $2c$ by simply traversing a minimum spanning tree. One of the classic results in approximation algorithms, *Christofides' algorithm* [25], improves upon this: by being careful when taking shortcuts in the spanning tree, it is possible to always construct a tour which has cost at most $1.5c$. This algorithm, while more than three decades old, is the best one known today for TSP in a general metric space.

A more fine-grained approach to approximation, which has become more common in recent years, is to look at not just the worst case ratio between optimum

value and the value found by the algorithm, but at the entire "approximability curve" of a problem: given that the optimum is at least x, how good solution $r(x)$ can we find? Many papers, e.g. [42, 22, 64], study the "high" or "low" ends of this curve, i.e., given that the optimum is $1-\epsilon$, or $\gamma+\epsilon$ (where γ is the smallest possible value for the optimum), how good solutions can we find? In recent papers such as [83, 87], the entire curve is studied.

Thus, we define approximation algorithms as follows.

Definition 3.2.1. Let $r : \mathbb{R} \to \mathbb{R}$ be a function. An algorithm $\mathcal{A}$ for a maximization problem $\mathcal{P}$ is an r-approximation algorithm if, for every instance $\Psi \in \mathcal{P}$,

$$\mathrm{Val}_\Psi(\mathcal{A}(\Psi)) \geq r(\mathrm{Opt}(\Psi)).$$

In the case when $\mathcal{A}$ is a randomized algorithm, the left hand side in the above equation is taken to be the expected value of $\mathrm{Val}_\Psi(\mathcal{A}(\Psi))$.

For $\alpha \in [0,1]$, we say that $\mathcal{A}$ is an α-approximation algorithm if it is an r-approximation algorithm with $r(x) = \alpha x$, i.e., if it is guaranteed to always find a solution which has a value within a factor α of the optimal value. Note that this definition only makes sense if $\mathrm{Opt}(\Psi) \geq 0$ for every Ψ, since if $\mathrm{Opt}(\Psi) < 0$, it is impossible to find a solution with value at least $\alpha\,\mathrm{Opt}(\Psi)$.

As a simple example of an approximation algorithm, consider the algorithm for Max 3-Sat which simply picks a random assignment to the variables. It is not hard to verify that this is an r-approximation algorithm for $r(x) = \frac{7}{8}x$, i.e., that it has an approximation ratio of $7/8$.

We remark that in general, r (and α) should be a function not only of $\mathrm{Opt}(\Psi)$ but also of the instance size $n = |\Psi|$, since it may be the case that the approximation that can be guaranteed will depend on n—e.g., the best approximation algorithm known for the Maximum Independent Set problem on n vertices has approximation ratio $\Omega\left(\frac{(\log n)^3}{n(\log\log n)^2}\right)$ [32]. However, for the problems studied in this thesis, Definition 3.2.1 will suffice.

We make a similar definition for *hardness* of approximation, which is what the subsequent chapters will mainly be about.

Definition 3.2.2. Let $0 < s < c$. A maximization problem $\mathcal{P}$ is (s,c)-hard if is NP-hard to distinguish between $\Psi \in \mathcal{P}$ with $\mathrm{Opt}(\Psi) \geq c$, and $\Psi \in \mathcal{P}$ with $\mathrm{Opt}(\Psi) < s$.

For $\alpha \in [0,1]$, we say that $\mathcal{P}$ is NP-hard to approximate within a factor α if there exists some c such that $\mathcal{P}$ is $(\alpha c, c)$-hard. If $\mathcal{P}$ is (s,c)-hard, then for any $r : \mathbb{R} \to \mathbb{R}$ such that $r(c) \geq s$, there does not exist a deterministic polynomial time r-approximation algorithm for $\mathcal{P}$ unless $\mathsf{P} = \mathsf{NP}$ (and in particular there does not exist an s/c-approximation algorithm), since such an algorithm would be able to distinguish $\mathrm{Opt}(\Psi) < s$ from $\mathrm{Opt}(\Psi) \geq c$.

It is easy to verify that if $\mathcal{P}$ is (s,c)-hard it is also $(s+x,c)$-hard and $(s,c-x)$-hard for every $x < c - s$.

3.3 Probabilistically Checkable Proofs

The most fundamental result in the field of inapproximability is the so-called PCP Theorem [6, 5]. The acronym PCP stands for *Probabilistically Checkable Proofs*. Informally, these are proofs which can be very efficiently verified, in the sense that one can look at only a constant number of bits of the proof and then with good probability know whether the proof is correct or not.

A *verifier* for a language L is a deterministic algorithm $\mathcal{V}$ which takes as input a string x and a "proof" Σ that $x \in L$, and is such that

- if $x \in L$ there is some Σ such that $\mathcal{V}(x, \Sigma)$ accepts.
- if $x \notin L$ then there is no Σ such that $\mathcal{V}(x, \Sigma)$ accepts.

A language is in NP if and only if it has a verifier which runs in time polynomial in x. For the current discussion, we can think of the "proof" Σ as a binary string, but in general, it may be a string over some larger alphabet Ω.

A (q, r)-*restricted verifier* is a probabilistic polynomial time algorithm $\mathcal{V}$ which, just like a regular verifier, is supposed to determine whether a given x is in some language L, with the help of some auxiliary information Σ. However, a (q, r)-restricted verifier is not allowed to look at the entire proof Σ, but only at q entries of Σ. In addition, $\mathcal{V}$ is allowed to use up to r random bits. We think of both r and q as functions of $n = |x|$, the size of the instance. Clearly, when q is small, such a verifier has to fail sometimes, unless L is in P. We say that $\mathcal{V}$ has *completeness* c and *soundness* s, if the following holds:

- If $x \in L$ there is some Σ such that $\mathcal{V}(x, \Sigma)$ accepts with probability at least c.
- If $x \notin L$ then for every Σ, the probability that $\mathcal{V}(x, \Sigma)$ accepts is at most s.

We denote by $\mathsf{PCP}_{c,s}[r, q]$ the class of all languages having an (r, q)-restricted verifier with completeness c and soundness s. It is not hard to see that for every $s < c$ and every q,

$$\mathsf{PCP}_{c,s}[\mathcal{O}(\log n), q] \subseteq \mathsf{NP}\,.$$

This holds because if $\mathcal{V}$ uses at most $\mathcal{O}(\log n)$ random bits, then we can construct a deterministic verifier which enumerates all possible choices of the random bits and then computes the exact probability that $\mathcal{V}$ accepts.

The PCP Theorem [6, 5] asserts that every language in NP has a $(\mathcal{O}(\log n), \mathcal{O}(1))$-restricted verifier with completeness 1 and soundness bounded away from 1.

Theorem 3.3.1 (The PCP Theorem)**.** *There exists a constant $\delta < 1$ such that*

$$\mathsf{NP} \subseteq \mathsf{PCP}_{1,\delta}[\mathcal{O}(\log n), \mathcal{O}(1)].$$

What does this have to do with hardness of approximation? Well, the theorem is in fact equivalent to the following hardness of approximation result.

Theorem 3.3.2 (The PCP Theorem, equivalent formulation)**.** *There exists a constant $\delta < 1$ such that* Max 3-Sat *is $(\delta, 1)$-hard.*

Let us sketch why these two formulations are equivalent. Theorem 3.3.1 can be equivalently stated as saying that there exists an $(\mathcal{O}(\log n), q)$-restricted verifier $\mathcal{V}$ for 3-Sat with completeness 1 and soundness δ, for some constant q. Consider the behavior of $\mathcal{V}$ on input a 3-CNF formula Ψ_0. $\mathcal{V}$ will randomly choose q bits $i_1, \ldots, i_q$ of Σ, and then accept if $\phi(\Sigma(i_1), \ldots, \Sigma(i_q))$ for some function $\phi : \{0,1\}^q \to \{0,1\}$ which also depends on the random bits. Now, let us enumerate all possible random strings for $\mathcal{V}$, and write down the corresponding constraints $\phi(\Sigma(i_1), \ldots, \Sigma(i_q))$. This gives an instance Ψ of the Max q-CSP problem on $2^{\mathcal{O}(\log n)} = \text{poly}(n)$ constraints. Furthermore, if Ψ_0 was satisfiable, then Ψ is satisfiable, whereas if Ψ_0 was not satisfiable, then $\text{Opt}(\Psi) \leq \delta$. Hence, we have that Max q-CSP is $(\delta, 1)$-hard. We can write each q-ary constraint ϕ as a 3-CNF formula on at most $\mathcal{O}(q)$ clauses. Doing this to Ψ gives a Max 3-Sat instance Ψ' such that if $\text{Opt}(\Psi) = 1$ then $\text{Opt}(\Psi') = 1$, whereas if $\text{Opt}(\Psi) \leq 1 - \epsilon = \delta$ then $\text{Opt}(\Psi') \leq 1 - \epsilon/\mathcal{O}(q) = 1 - \Omega(\epsilon) = \delta'$. Hence Max 3-Sat is $(\delta', 1)$-hard. This shows that Theorem 3.3.1 implies Theorem 3.3.2. The other direction is easier. Theorem 3.3.2 asserts that there exists a reduction R from 3-Sat to Max 3-Sat such that if x is a 3-CNF formula which is satisfiable, then $R(x)$ is also satisfiable, and if $R(x)$ is not satisfiable, then $\text{Opt}(R(x)) \leq \delta$. We can then construct a verifier $\mathcal{V}$ which expects as proof Σ a satisfying assignment to $R(x)$, and then verifies this proof by simply picking a random clause of $R(x)$ and checking that it is satisfied by Σ. We omit the details.

The PCP Theorem can also be viewed as stating that L-Label Cover is $(\delta, 1)$-hard for some $\delta < 1$ and some constant L, by the following reduction from Max 3-Sat to L-Label Cover: given a Max 3-Sat instance, construct a bipartite graph G where the vertex sets are the variables and clauses. The label for a clause ϕ is supposed to be one of the 7 satisfying assignments to the variables of ϕ, and the label for a vertex x is supposed to be the value of that vertex. A clause ϕ and a vertex x is connected by an edge if x occurs in ϕ, and the constraint π associated with the edge checks that the label of x is as claimed by the label of ϕ.

A natural way to boost this $(\delta, 1)$-hardness, is to apply *parallel repetition*. For a L-Label Cover instance $\Psi = (X, Y, E, \Pi)$, the n-fold parallel repetition of Ψ, denoted Ψ^n, is the L^n-Label Cover instance (X^n, Y^n, E', Π'), where there is an edge between $(x_1, \ldots, x_n) \in X^n$ and $(y_1, \ldots, y_n) \in Y^n$ whenever all of $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are edges in E. The label set L^n is identified with $[L]^n$, and we think of the label $\ell(x) \in [L]^n$ of a vertex $x = (x_1, \ldots, x_n) \in X^n$ as assigning labels in $[L]$ to each of $x_1, \ldots, x_n$. An edge $(x, y) \in E'$ is satisfied by a labeling $\ell : X^n \cup Y^n \to [L]^n$ if the edges $(x_1, y_1), \ldots, (x_n, y_n)$ are all satisfied by the labelings of $x_1, \ldots, x_n, y_1, \ldots, y_n$ induced by $\ell(x)$ and $\ell(y)$.

It is not hard to see that any labeling of Ψ with value δ can be lifted to a labeling of Ψ' with value δ^n. However, the converse does not necessarily hold. In other words, there can be labelings of Ψ' with significantly higher value than

δ^n. A natural question is whether the optimum value of Ψ^n at least goes down exponentially with n. This was answered in the affirmative in a celebrated result by Raz [89], who proved that if $\mathrm{Opt}(\Psi) \leq 1 - \epsilon$ with $\epsilon > 0$, then there exists some $\epsilon' > 0$ depending only on ϵ, such that

$$\mathrm{Opt}(\Psi^n) \leq (1 - \epsilon')^{\Omega(n/\log L)}$$

(recall that L is the number of labels in the L-Label Cover problem). Together with the PCP Theorem, this implies the following strong inapproximability result for label cover.

Theorem 3.3.3. *For every $\gamma > 0$ there exists an L such that the L-Label Cover problem is $(\gamma, 1)$-hard.*

As mentioned earlier, Label Cover often provides a useful starting point for proving strong inapproximability results. The reason for this is exactly the very strong hardness given by Theorem 3.3.3. Let us explain how the proofs of such results often proceed. Suppose that $P : \{0,1\}^k \to \{0,1\}$ is a predicate and that we want to prove that Max CSP(P) is (s, c)-hard for some s and c. We can do this in the same fashion as our sketch of the proof that Theorem 3.3.1 implies Theorem 3.3.2: suppose we have an $(\mathcal{O}(\log n), k)$-restricted PCP verifier $\mathcal{V}$ for L-Label Cover which only reads k bits $b_1, \ldots, b_k$ of the proof, and then accepts if $P(b_1, \ldots, b_k)$ is true. Then, to prove that Max CSP(P) is (s, c)-hard, it suffices to prove that there exists some $\gamma > 0$ *independent* of L such that

- If $\mathrm{Opt}(\Psi) = 1$, then there is a proof Σ such that $\Pr[\mathcal{V}(\Psi, \Sigma) \text{ accepts}] \geq c$.
- If $\mathrm{Opt}(\Psi) \leq \gamma$ then for every proof Σ it holds that $\Pr[\mathcal{V}(\Psi, \Sigma) \text{ accepts}] \leq s$.

Here, the probabilities are taken over the randomness of $\mathcal{V}$. To see this, note that if we simply write down all possible checks made by the verifier (with according weights, based on the probabilities of different checks) we obtain a Max CSP(P) instance Ψ' such that if $\mathrm{Opt}(\Psi) = 1$ then $\mathrm{Opt}(\Psi') \geq c$ whereas if $\mathrm{Opt}(\Psi) \leq \gamma$ then $\mathrm{Opt}(\Psi') \leq s$. Furthermore, this reduction is polynomial (since $\mathcal{V}$ uses only a logarithmic amount of randomness). Hence if L is taken large enough so that L-Label Cover is $(\gamma, 1)$-hard, we get that Max CSP(P) is (s, c)-hard.

3.4 The Unique Games Conjecture

In 2002, Khot introduced a conjecture known as the *Unique Games Conjecture* (UGC). In the short time since it was introduced, this conjecture has established itself as one of the single most important open problems in theoretical computer science. The reason for this is that it has been shown that the UGC implies a slew of optimal or near-optimal inapproximability results which are beyond the reach of "traditional" PCP techniques. Examples include [65, 66, 64, 63, 28, 94, 56, 7, 8, 9, 83, 75, 67, 87]. In many cases, the UGC not only enables us to prove far stronger

results than the best unconditional results, they also enable us to construct results which *exactly* match the performance ratio obtained by the best algorithms.

To formulate the conjecture, we need to define a special case of the Label Cover problem known as the Unique Label Cover problem.

Definition 3.4.1. The L-Unique Label Cover problem is the special case of the L-Label Cover problem in which each constraint $\pi_e : [L] \to [L]$ is a permutation on $[L]$.

How hard is L-Unique Label Cover? It is an easy exercise to check that if the optimum $\mathrm{Opt}(\Psi)$ of an L-Unique Label Cover instance is 1, i.e., if there is a labeling satisfying all edges of Ψ, then such a labeling can be efficiently found in time $\mathcal{O}(L \cdot |E|)$. Hence, in sharp contrast to the state of affairs for Label Cover, we can not hope that L-Unique Label Cover is $(s, 1)$-hard for any $s < 1$. The Unique Games Conjecture asserts that L-Unique Label Cover is very hard to approximate as soon as we move to almost-satisfiable instances.

Conjecture 3.4.2 (Unique Games Conjecture)**.** *For every $\gamma > 0$, there exists an L such that the L-Unique Label Cover problem is $(\gamma, 1-\gamma)$-hard.*

Whenever a problem is hard under the UGC, we say that it is Unique Games-hard, or more often simply UG-hard. For instance, for every $\gamma > 0$, L-Unique Label Cover is, by definition, $(\gamma, 1-\gamma)$-UG-hard for sufficiently large L.

There have been plenty of works which, either directly or indirectly, has helped improve our understanding of the UGC. Feige and Reichman [36] showed that for every γ there exists a $\delta > 0$ and L such that L-Unique Label Cover is $(\gamma\delta, \delta)$-hard. However, the constant δ tends to 0 quite rapidly with γ, and their technique is inherently limited to getting that type of bounds.

On the algorithmic side, there have been several results [62, 102, 20, 4]. The best general algorithm is due to Charikar et al. [20], who gave an algorithm which on input a L-Unique Label Cover instance Ψ with $\mathrm{Opt}(\Psi) \geq 1-\gamma$, finds a labeling with value at least $1 - \mathcal{O}(\sqrt{\gamma \log L})$. This implies that in order for the UGC to be true, L needs to be at least $2^{\Omega(1/\gamma)}$. This result is nicely complemented by results of Khot et al. [63], who prove that, if the UGC is true then in the non-bipartite version one can take $L = 2^{\pi/(2\gamma)}$. In other words, any significant improvement of the algorithm of Charikar et al. would disprove the UGC.

Recently, Arora et al. [4] gave an algorithm for L-Unique Label Cover on constraint graphs with good *expansion* (measured in terms of the spectral gap of the constraint graph). In particular, this has the interesting consequence that L-Unique Label Cover is *not* hard on random graphs, as these are likely to be good expanders. This is contrary to many other constraint satisfaction problems such as Max 3-Sat, where there are indications that random instances are as hard as the worst case instances with respect to approximability [95].

Another very interesting line of research related to the UGC is the investigation of parallel repetition. Recall that Raz [89] proved that for any Label Cover

instance Ψ with value $\mathrm{Opt}(\Psi) \leq 1-\epsilon$, the value of the n-fold repetition Ψ^n is at most $\mathrm{Opt}(\Psi^n) \leq (1-\epsilon')^{\Omega(n/\log L)}$ for some $\epsilon' > 0$ depending only on ϵ (one could take $\epsilon' = \epsilon^{32}$). The proof of Raz was simplified by Holenstein [57], who proved that $\mathrm{Opt}(\Psi^n) \leq (1-\epsilon^3)^{\Omega(n/\log L)}$. This was then further improved by Rao [88], who proved that $\mathrm{Opt}(\Psi^n) \leq (1-\epsilon^2)^{\Omega(n)}$ (note the lack of dependency on L in the exponent). We remark that the first two results hold also for a more general class of Label Cover problems than the ones defined in Definition 3.1.4, where each π_e is an arbitrary relation on $[L] \times [L]$ rather than a function from $[L]$ to $[L]$. Rao's result implies the following equivalent formulation of the UGC:

Theorem 3.4.3. *Suppose that there exists a constant $t > 2$ such that for every $\gamma > 0$ there exists an L such that L-Unique Label Cover is $(1-\gamma^{1/t}, 1-\gamma)$-hard. Then the Unique Games Conjecture is true.*

In other words, in order for the Unique Games Conjecture to be true, it suffices that Unique Games are hard on almost-satisfiable instances with just a very small gap. The obvious question of whether the condition $t > 2$ in Theorem 3.4.3 can be improved is known as the Strong Parallel Repetition Conjecture, introduced by Feige et al. [34]. The conjecture is particularly interesting as it would imply that proving the UGC is equivalent to proving that Max Cut is $(1-\Theta(\sqrt{\epsilon}), 1-\epsilon)$-hard for every $\epsilon > 0$ (it is already known that the UGC implies such hardness [63]), and this problem could potentially be much easier to attack than the UGC in its original form. However, as recently proved by Raz [90], the Strong Parallel Repetition Conjecture is false, so in some sense, Theorem 3.4.3 is the "weakest possible" characterization of the UGC.

Despite all the results mentioned above, there are still many aspects of Unique Label Cover which are not properly understood. For instance, it is not known how easy or hard the problem is in the special case where the constraint graph is the boolean hypercube $\{0,1\}^n$.

3.5 Constructions of Pairwise Independence

In Chapter 5, we will be interested in pairwise independent distributions with as small support as possible. There is a rich literature on constructions of t-wise independence. In this section, we present some such constructions.

3.5.1 Large Alphabets

The following well-known lemma is useful for constructing pairwise independent distributions.

Lemma 3.5.1. *Let R be a finite commutative ring, and let R^* denote the set of units (i.e., the elements having a multiplicative inverse) of R. Let $u, v \in R^n$ be*

two vectors over R such that $u_i v_j - u_j v_i \in R^*$ for some i, j.[1] *Let $X \in R^n$ be a uniformly random vector over R^n and let μ be the probability distribution over R^2 of $(\langle u, X\rangle, \langle v, X\rangle) \in R^2$. Then μ is a balanced pairwise independent distribution.*

Proof. Without loss of generality, assume that $i = 1$ and $j = 2$. It suffices to prove that, for all $(a, b) \in R^2$ and any choice of values of $X_3, \ldots, X_n$, we have

$$\Pr[(\langle u, X\rangle, \langle v, X\rangle) = (a, b) \mid X_3, \ldots, X_n] = 1/|R|^2.$$

For this to be true, we need that the system

$$\begin{cases} u_1 X_1 & + & u_2 X_2 & = & a' \\ v_1 X_1 & + & v_2 X_2 & = & b' \end{cases}$$

has exactly one solution, where $a' = a - \sum_{i=3}^{n} u_i X_i$ and similarly for b'. This in turn follows directly from the condition on u and v. □

Consequently, given a set of m vectors in R^n such that any pair of them satisfy the condition of Lemma 3.5.1, we can construct a pairwise independent distribution over R^m with support size $|R|^n$.

From this, it is easy construct small pairwise independent distributions over $[q]^k$ for q a prime power.

Theorem 3.5.2. *Let q be a prime power. Then, there exists a pairwise independent distribution μ over $[q]^k$ with support size*

$$|\operatorname{Supp}(\mu)| \leq k(q-1)q.$$

In the special case that $k = \frac{q^r - 1}{q-1}$ for some r, the bound on the support improves to

$$|\operatorname{Supp}(\mu)| \leq k(q-1) + 1.$$

The construction we use is essentially the same as that of [81], though in a somewhat different language.

Proof. Let $r = \lceil \log_q(k(q-1)+1) \rceil$, and $n = (q^r - 1)/(q-1) \geq k$.

Let $\mathbb{P}(\mathbb{F}_q^r)$ be the projective space over $\mathbb{F}_q^r$, i.e.,

$$\mathbb{P}(\mathbb{F}_q^r) = (\mathbb{F}_q^r \setminus 0)/\!\sim.$$

Here $\sim$ is the equivalence relation defined by $(x_1, \ldots, x_r) \sim (y_1, \ldots, y_r)$ if there exists a $c \in \mathbb{F}_q^*$ such that $x_i = c y_i$ for all i, i.e., if $(x_1, \ldots, x_r)$ and $(y_1, \ldots, y_r)$ are linearly dependent. We then have

$$|\mathbb{P}(\mathbb{F}_q^r)| = (q^r - 1)/(q-1) = n.$$

[1] *In the case that R is a field, the condition is equivalent to saying that u and v are linearly independent.*

Choose n vectors $u_1, \ldots, u_n \in \mathbb{F}_q^r$ as representatives from each of the equivalence classes of $\mathbb{P}(\mathbb{F}_q^r)$. Then any pair u_i, u_j satisfy the condition of Lemma 3.5.1. Thus, we have that $(\langle u_i, X\rangle)_{1\le i\le n}$ for a uniformly random $X \in \mathbb{F}_q^r$ induces a balanced pairwise independent distribution over $\mathbb{F}_q^n$ (and hence over $[q]^k$) with support size q^r.

When $k = (q^r - 1)/(q - 1)$, this gives a support of size $k(q - 1) + 1$, and for general k, in particular

$$k = (q^{r-1} - 1)/(q - 1) + 1,$$

we lose almost a factor q in the support size. □

3.5.2 The Binary Alphabet

Let us now look closer at the special case of pairwise independence over binary strings, i.e., the case $q = 2$.

An Hadamard matrix is an $n \times n$ matrix over ± 1 such that $HH^T = nI$, i.e., each pair of rows, and each pair of columns, are orthogonal. Let $\mathrm{Had}(n)$ denote the smallest $n' \ge n$ such that there exists an $n' \times n'$ Hadamard matrix. It is a well-known fact that Hadamard matrices give pairwise independent distributions. To be specific, we have the following proposition:

Proposition 3.5.3. *For every k, there exists a pairwise independent distribution μ over $\{-1, 1\}^k$ with support size*

$$|\operatorname{Supp}(\mu)| = \mathrm{Had}(k + 1).$$

Proof. Let $n = \mathrm{Had}(k+1)$ and let A be an $n \times n$ Hadamard matrix, normalized so that one column contains only ones. Remove $n-k$ of the columns, including the all-ones column, and let A' be the resulting $n \times k$ matrix. Let $\mu : \{-1, 1\}^k \to [0, 1]$ be the probability distribution which assigns probability $1/n$ to each row of A'. Then μ is a balanced pairwise independent distribution with $|\operatorname{Supp}(\mu)| = \mathrm{Had}(k+1)$. □

It is well known that Hadamard matrices can only exist for $n = 1$, $n = 2$, and $n \equiv 0 \pmod 4$. The famous *Hadamard Conjecture* [99, 50, 84] asserts that Hadamard matrices exist for all n which are divisible by 4, in other words, that $\mathrm{Had}(n) = 4\lceil n/4\rceil \le n+3$. The smallest value for which the conjecture is not known to hold is $n = 668$. It is also possible to get useful unconditional bounds on $\mathrm{Had}(n)$. We now give one such bound, which is an easy consequence of the two following theorems.

Theorem 3.5.4 ([84])**.** *For every odd prime p and integers $e, f \ge 0$, there exists an $n \times n$ Hadamard matrix H_n where $n = 2^e(p^f + 1)$, whenever this number is divisible by 4.*

Theorem 3.5.5 ([10])**.** *There exists an integer n_0 such that for every $n \ge n_0$, there is a prime p between n and $n + n^{0.525}$.*

Corollary 3.5.6. *For every n, it holds that* $\mathrm{Had}(n) \leq n + \mathcal{O}(n^{0.525})$.

Proof. Let p be the smallest prime larger than $n/2$, and let $n' = 2(p+1) \geq n$. Then, Theorem 3.5.4 asserts that there exists an $n' \times n'$ Hadamard matrix, so $\mathrm{Had}(n) \leq n'$. If n is sufficiently large ($n \geq 2n_0$), then by Theorem 3.5.5, $p \leq n/2 + (n/2)^{0.525}$ and $n' \leq n + 2n^{0.525}$, as desired. □

To summarize the discussion in this section, we have the following theorem on pairwise independence over $\{0,1\}^k$.

Theorem 3.5.7. *There exists a balanced pairwise independent distribution μ over $\{0,1\}^k$ with support size*

$$|\operatorname{Supp}(\mu)| \leq k + \mathcal{O}(k^{0.525}).$$

Furthermore, if the Hadamard Conjecture is true, there exists μ with

$$|\operatorname{Supp}(\mu)| = \lceil (k+1)/4 \rceil \leq k + 4.$$

We remark that it should be possible to get a stronger unconditional bound on $\mathrm{Had}(n)$ than the one given by Corollary 3.5.6 by using stronger construction techniques than the one of Theorem 3.5.4.

3.6 Properties of the Bivariate Normal Distribution

In this section, we prove some basic facts about Γ_ρ (Definition 2.2.11). The first is the following symmetry observation.

Proposition 3.6.1. *For all $\rho \in [-1,1]$, $\mu_1, \mu_2 \in [-1,1]$, we have*

$$\Gamma_\rho(-\mu_1, -\mu_2) = \Gamma_\rho(\mu_1, \mu_2) + \mu_1/2 + \mu_2/2.$$

Proof. Let $t_i = \Phi^{-1}\left(\frac{1-\mu_i}{2}\right)$, and let X and Y be two ρ-correlated $N(0,1)$ variables. Clearly, $\Gamma_\rho(-\mu_1, -\mu_2) = \Pr[X \leq -t_1 \wedge Y \leq -t_2]$. Assume that $\mu_1 < 0, \mu_2 < 0$ (implying $t_1 > 0$ and $t_2 > 0$). We have

$$\begin{aligned}
\Gamma_\rho(\mu_1,\mu_2) - \Gamma_\rho(-\mu_1,-\mu_2) &= \Pr[X \leq t_1 \wedge Y \leq t_2] - \Pr[X \leq -t_1 \wedge Y \leq -t_2] \\
&= \Pr[X \leq 0 \wedge |Y| \leq t_2] + \\
&\quad \Pr[0 \leq X \leq t_1 \wedge -t_2 \leq Y \leq 0] + \\
&\quad \Pr[0 \leq X \leq t_1 \wedge 0 \leq Y \leq t_2] + \\
&\quad \Pr[|X| \leq t_1 \wedge Y \leq -t_2].
\end{aligned}$$

Note that $\Pr[0 \leq X \leq t_1 \wedge 0 \leq Y \leq t_2] = \Pr[-t_1 \leq X \leq 0 \wedge -t_2 \leq Y \leq 0]$ and that $\Pr[X \leq 0 \wedge |Y| \leq t_2] = \Pr[X \geq 0 \wedge |-Y| \leq t_2] = \Pr[|Y| \leq t_2]/2 = -\mu_2/2$. Thus,

$$\begin{aligned}
\Gamma_\rho(\mu_1,\mu_2) - \Gamma_\rho(-\mu_1,-\mu_2) &= \Pr[X \leq 0 \wedge |Y| \leq t_2] + \\
&\quad \Pr[|X| \leq t_1 \wedge -t_2 \leq Y \leq 0] + \\
&\quad \Pr[|X| \leq t_1 \wedge Y \leq -t_2] \\
&= -\mu_1/2 - \mu_2/2,
\end{aligned}$$

as desired. The other three sign combinations for μ_1 and μ_2 are taken care of. □

Next, we compute the derivative of Γ_ρ. For the rest of this section, let $t(x) = \Phi^{-1}\left(\frac{1-x}{2}\right)$.

Proposition 3.6.2. *For $\rho \in (-1,1)$, we have*

$$\frac{\partial \Gamma_\rho}{\partial \mu_1}(\mu_1,\mu_2) = -\frac{1}{2}\Phi\left(\frac{t(\mu_2)-\rho t(\mu_1)}{\sqrt{1-\rho^2}}\right).$$

Proof. This follows from the fact that $\Gamma_\rho(\mu_1,\mu_2)$ can be written as

$$\Gamma_\rho(\mu_1,\mu_2) = \int_{x=-\infty}^{t(\mu_1)} \phi(x)\Phi\left(\frac{t(\mu_2)-\rho x}{\sqrt{1-\rho^2}}\right) \mathrm{d}x,$$

giving

$$\frac{\partial \Gamma_\rho}{\partial \mu_1}(\mu_1,\mu_2) = t'(\mu_1)\phi(t(\mu_1))\Phi\left(\frac{t(\mu_2)-\rho t(\mu_1)}{\sqrt{1-\rho^2}}\right).$$

Using $t'(x) = -\frac{1}{2\phi(t(x))}$, we obtain the desired result. □

As a simple corollary, we get the following result.

Corollary 3.6.3. *For $\rho \in (-1,1)$, we have*

$$\frac{\partial \Gamma_\rho}{\partial \mu}(\mu,\mu) = -\Phi\left(\sqrt{\frac{1-\rho}{1+\rho}}t(\mu)\right).$$

Note that Corollary 3.6.3 implies that $\frac{\partial^2 \Gamma_\rho}{\partial \mu^2}(\mu,\mu) > 0$ for all μ, i.e. that $\Gamma_\rho(\mu,\mu)$ is a convex function in μ.

Proof. Indeed,

$$\begin{aligned}
\frac{\partial \Gamma_\rho}{\partial \mu}(\mu,\mu) &= \frac{\partial \Gamma_\rho}{\partial \mu_1}(\mu,\mu) + \frac{\partial \Gamma_\rho}{\partial \mu_2}(\mu,\mu) \\
&= 2\cdot\left(-\frac{1}{2}\Phi\left(\frac{(1-\rho)t(\mu)}{\sqrt{1-\rho^2}}\right)\right) \\
&= -\Phi\left(\sqrt{\frac{1-\rho}{1+\rho}}t(\mu)\right).
\end{aligned}$$

Here, we used the fact that $\Gamma_\rho(\mu_1,\mu_2) = \Gamma_\rho(\mu_2,\mu_1)$, so the derivative of Γ_ρ with respect to μ_2 can also be computed using Proposition 3.6.2. □

Another simple but useful corollary of Proposition 3.6.2 is that Γ_ρ is "Lipschitz continuous".

Corollary 3.6.4. *For any $\mu_1, \mu_1', \mu_2, \mu_2' \in [-1,1]$ and $\rho \in (-1,1)$, we have*

$$|\Gamma_\rho(\mu_1,\mu_2) - \Gamma_\rho(\mu_1',\mu_2')| \le \frac{|\mu_1-\mu_1'| + |\mu_2-\mu_2'|}{2}.$$

Chapter 4

Hardness by Testing Dictators

This chapter gives a review of a, by now fairly standard, method to obtain inapproximability results. In particular, we describe certain algorithms known as dictatorship tests, and their intimate connection with PCPs. The connections between these two types of objects have been known for a long time. In recent years it has been realized that, when constructing PCPs for Unique Label Cover, this connection is particularly strong. It turns out that with a "sufficiently general" definition of what a dictatorship test is, such tests can be used in a black-box fashion to prove Unique Games-based inapproximability for virtually any constraint satisfaction problem. However, the "sufficient generality" needed in order for us to be able to derive both the results in Chapter 5 for k-CSPs and the results in Chapter 6 for 2-CSPs turns out to be fairly large, making the definitions quite cumbersome to work with. Thus, rather than formally stating and proving this black-box conversion from dictatorship tests to hardness results, we will in this chapter give the subject an informal (but still relatively detailed) treatment, and defer the formal proofs to Chapter 5 and Chapter 6.

This chapter is not a prerequisite for Chapter 5 and Chapter 6, in the sense that they are both (mostly) self-contained. However, a reader not familiar with these types of results, wanting to obtain some intuition for how they work, is encouraged to first read this chapter.

4.1 Dictators

Throughout this chapter, let Ω be a finite domain and n a positive integer. The reader is encouraged to think of n as being very large, and of $|\Omega|$ as being a relatively small constant. A function $f : \Omega^n \to \Omega$ is called a *dictator* if it is defined by

$$f(x) = x_i$$

for every $x \in \Omega^n$. We use $\textsc{Dict}_i$ to denote the dictator function returning the i:th input coordinate (formally, we should also index Dict by Ω and n, but these will

always be clear from the context, so we omit them for the sake of brevity).

In the "traditional" inapproximability literature, dictators are commonly also called *long codes*, often with a different notation than our notation above. Long codes were first defined in the context of approximability by Bellare, Goldreich and Sudan [13], and have since become ubiquitous in the design of strong PCPs. The name long code comes from the fact that the mapping $i \mapsto \textsc{Dict}_i$ can be seen as a highly redundant error-correcting code, which maps elements from a set of size n to a set of size $|\Omega|^{|\Omega|^n}$ (the set of all functions $f : \Omega^n \to \Omega$). This code has very good error-correcting properties. In fact, dictatorship tests, which are defined in the next section and which play a crucial role in many hardness results, can be viewed as so-called local error-detection procedures for the long code.

4.2 Dictatorship Testing

Dictatorship testing is a particular case of a subject in theoretical computer science known as *property testing*. Loosely speaking, in property testing one seeks to test whether a given (possibly huge) object has a given property by just looking at a very small portion of the object, and then making an "educated guess". As an example, suppose that we are given the adjacency matrix A of a graph G, and want to test if G is *triangle-free*[1] by only examining a small number of entries A. This problem is a special case of the more general and very interesting problem of testing whether a graph G has some fixed graph H as a subgraph, which has fascinating applications in additive combinatorics.

The critical performance characteristics of such a test are the *completeness* and *soundness* of the test. If our test accepts every G which has the property (in this case, triangle-freeness) with probability at least c, we say that it has completeness c. If our test accepts every G which is "far" from having the property with probability at most s, we say that it has soundness s. The exact notion of what it means to be "far" from having a property can vary in different settings as we shall see in the remainder of this chapter. For the triangle-freeness testing problem, a reasonable definition is to say that G is ϵ-far from every triangle-free graph if at least ϵn^2 entries of the adjacency matrix A have to be changed in order to make G triangle-free.

For the problem of testing triangle-freeness, it turns out that there is a function $f : (0, 1] \to \mathbb{R}$ such that for every $\epsilon > 0$ there is a testing algorithm looking only at a constant number $f(\epsilon)$ of entries of A, which accepts every triangle-free graph with probability 1, and rejects every graph which is ϵ-far from being triangle-free with probability at least $1/2$ (to readers familiar with the *Triangle Removal Lemma* of additive combinatorics, this should not come as a surprise).

Informally, dictatorship testing is the property testing problem in which we are given access to some arbitrary function $f : \Omega^n \to \Omega$, and should determine whether or not $f = \textsc{Dict}_i$ for some $i \in [n]$, by only evaluating f on a small number of

[1] Recall that a graph G is triangle-free if it contains no triangles, i.e., if there are no $u, v, w \in V$ such that all three of (u, v), (v, w) and (w, u) are edges.

inputs. To make this testing problem concrete, we need to specify what we mean by a function f being "far" from every dictator. For the application that we are interested in, hardness of approximation, we need to define "far" in such a way that dictatorship tests can be used to construct PCPs. Then, as the performance of the PCP in general depends directly on the soundness and completeness properties of tests, we should choose a notion of distance which is as "weak" as possible, in the sense that it allows us to construct tests with good completeness and soundness, while at the same time being strong enough so that we can convert such test into PCPs.

4.3 Hardness

As we shall see in this section, the most important property of a definition of "distance" is that it should, in some sense, induce a good "code" on the set of all functions from $\Omega^n \to \Omega$. In particular, suppose that the notion of distance is such that every function f is close to at most R different dictators, where R is some number which may depend on Ω but is independent of n.

We will illustrate how a dictatorship test, under any such definition of distance, can be used to obtain Unique Games-based hardness results. Suppose for concreteness that we have a dictatorship test $\mathcal{T}$ with completeness c and soundness s for $f : \{-1,1\}^n \to \{-1,1\}$ which evaluates f in 2 points x and y and then accepts if $f(x) \neq f(y)$. Let us use this test to construct a PCP verifier for L-Unique Label Cover. We remark that when working with a binary domain, as we do here, we almost always choose $\Omega = \{-1,1\}$ rather than $\{0,1\}$ because it tends to make computations a lot cleaner.

The format of the proof is as follows. Given a L-Unique Label Cover instance $\Psi = (X, Y, E, \Pi)$, the verifier will expect as a proof a set $\Sigma = \{f_v\}_{v \in X \cup Y}$ of functions $f_v : \{-1,1\}^L \to \{-1,1\}$, one for each vertex of Ψ. The verifier expects that $f_v = \textsc{Dict}_{\ell(v)}$, for some good labeling $\ell : X \cup Y \to [L]$ of Ψ.

Given an arbitrary proof Σ, how can we verify it efficiently? Consider the following verification process. First pick a random vertex v, and two random neighbors w_1, w_2 of v. Now, if the two edges $e_1 = (v, w_1)$ and $e_2 = (v, w_2)$ are satisfied by the labeling which Σ is supposed to encode, then we should have $f_v = \textsc{Dict}_i$, $f_{w_1} = \textsc{Dict}_{\pi_{e_1}(i)}$ and $f_{w_2} = \textsc{Dict}_{\pi_{e_2}(i)}$ for some $i \in [L]$. In particular, checking whether the two edges selected are satisfied amounts to checking whether

$$h_1 = h_2 = \textsc{Dict}_i \tag{4.1}$$

for some $i \in [L]$, where $h_i(x) = f_{w_i}(x \circ \pi_{e_i}^{-1})$ (recall the definition of composition $\circ$ from Section 2.1). Checking this is almost the same as a dictator test, except that we have two different functions which should be the same dictator. It turns out that this is not really an important difference, and that in general, the dictatorship test $\mathcal{T}$ can be used in this setting as well.

We thus have a verification procedure which only reads two single bits out of the gigantic proof Σ, and then checks that these bits are not equal. How well does this verifier work? It is not too hard to see that if $\mathrm{Opt}(\Psi) \geq 1-\gamma$, then the intended proof is accepted with probability at least $c-2\gamma$, simply because with probability at least $1-2\gamma$, Equation (4.1) will hold and in this case we accept with probability at least c.

The soundness analysis is more interesting. Suppose that the verifier accepts with probability at least $s+\epsilon$. Then a simple averaging argument shows that for at least an $\epsilon/2$ fraction of all $v \in X$ it must be the case that the verifier accepts this particular v with probability at least $s+\epsilon/2$. Let us call these vertices "good", and fix some "good" vertex v. Another averaging argument shows that for at least an $\epsilon/4$ fraction of all pairs (w_1, w_2) of neighbors of v, $\mathcal{T}$ accepts this pair with probability at least $s+\epsilon/4$. Consider the neighbor w^* which participates as w_1 in the maximum number of such pairs. We then have that for a fraction $\epsilon/4$ of all neighbors w of v, $\mathcal{T}$ accepts (w^*, w) with probability $s+\epsilon/4$, which implies that there is some i such that f_{w^*} is close to $\textsc{Dict}_{\pi_{e^*}(i)}$ and f_w is close to $\textsc{Dict}_{\pi_e(i)}$.

Consider now the following random labeling ℓ of Ψ. For each $w \in Y$, pick $\ell(w)$ uniformly at random such that f_w is close to $\textsc{Dict}_{\ell(w)}$ (or arbitrarily if f_w is not close to any dictator). For a "good" $v \in X$ with neighbor w^* as above, simply let $\ell(v) = \pi_{e^*}^{-1}(\ell(w^*))$. Assign the other labels of X arbitrarily. Now, we claim that the expected fraction of satisfied edges by ℓ is at least $\frac{\epsilon^2}{8R^2}$. Recall that R is the upper bound on the number of different dictators that a function can be simultaneously close to. To prove the bound on $\mathbb{E}[\mathrm{Val}(\ell)]$, let v be one of the good vertices with neighbor w^*, and let w be one of the neighbors such that (w^*, w) is accepted by $\mathcal{T}$ with good probability. Then since w is close to $\textsc{Dict}_{\pi_e(i)}$, the probability that $\ell(w) = \pi_e(i)$ is at least $1/R$, and since w^* is close to $\textsc{Dict}_{\pi_{e^*}(i)}$, the probability that $\ell(w^*) = \pi_{e^*}(i)$ is at least $1/R$. Furthermore these two events are independent. But if $\ell(w^*) = \pi_{e^*}(i)$ then by definition $\ell(v) = i$ and hence this edge is satisfied with probability at least $1/R^2$. In total, this type of edge constitutes an $\epsilon^2/8$ fraction of all edges, implying the bound on $\mathbb{E}[\mathrm{Val}(\ell)]$.

To summarize, this shows that if the verifier accepts with probability at least $s+\epsilon$, it must be the case that $\mathrm{Opt}(\Psi) \geq \frac{\epsilon^2}{8R^2}$. This in turn implies that if we take γ smaller than $\frac{\epsilon^2}{8R^2}$ and then take L large enough so that L-Unique Label Cover is $(\gamma, 1-\gamma)$-UG-hard, it will be UG-hard to determine whether there is a proof that makes the PCP verifier accept with probability at least $c-\epsilon$, or whether every proof is accepted with probability at most $s+\epsilon$. This in turn implies that Max Cut is $(s+\epsilon, c-\epsilon)$-UG-hard to approximate for every $\epsilon > 0$.

We only used one single property of the notion of "distance"[2]: that a function $f : \Omega^n \to \Omega$ could be close to at most R different dictators, where R is independent on n.

[2]This is not quite true, since we also required that the test $\mathcal{T}$ should work even when we plug in two different functions, and whether or not this is possible to achieve may also depend on the definition of distance. However, this issue usually turns out to be a minor one.

4.4 Folding

In many hardness applications, it is crucial that the supposed dictator encodings of the labels in a good labeling are *balanced*. A function $f : \Omega^n \to \Omega$ is balanced if, for every $a \in \Omega$, the equation $f(x) = a$ has exactly $|\Omega|^{n-1}$ solutions, i.e., if f takes every value equally often. We will see one example where this is crucial shortly, in Section 4.6, and it will also be needed both in Chapter 5 and Chapter 6.

How can we enforce this condition on an arbitrary proof $\Sigma = \{f_v\}_{v \in X \cup Y}$? We do this by a technique called *folding*, which enforces a special type of balance. A function $f : \Omega^n \to \Omega$ is said to be *folded*, if, for every $x \in \Omega^n$ and $a \in \Omega$

$$f(x_1 + a, x_2 + a, \ldots, x_n + a) = f(x_1, \ldots, x_n) + a,$$

where "+" in Ω is defined so that $(\Omega, +)$ is an Abelian group. Note that a dictator is always folded, and that a folded function is completely specified by its values on the $|\Omega|^{n-1}$ points of Ω^n of the form $(x_1, \ldots, x_{n-1}, 0)$. Thus, we can enforce that f_v is folded by saying that the proof should consist of the values of $f_v(x_1, \ldots, x_{n-1}, 0)$. Then, when the verifier wants to evaluate $f_v(x_1, \ldots, x_n)$, it instead evaluates

$$f_v(x_1 - x_n, \ldots, x_{n-1} - x_n, 0) + x_n.$$

This successfully enforces all supposed dictator encodings to be balanced, but it has a cost. Suppose that our PCP verifier uses some predicate $P : [q]^k \to \{0, 1\}$ when deciding whether or not to accept. When we use folding, the verifier will read some k entries y_1, y_2, …, $y_k \in [q]$ of the proof, and then accept if $P(y_1 + a_1, \ldots, y_k + a_k)$ is true. This means that when we use the PCP verifier as a reduction to a CSP, the resulting instance will have constraints where P is applied to k-tuples of literals, and not just k-tuples of variables. In other words, we will get hardness for Max CSP(P), rather than for Max CSP$^+$(P).

To summarize: if we want to show hardness for Max CSP(P) we can assume that the supposed dictator encodings in the PCP are balanced, but if we want to show hardness for Max CSP$^+$(P) we can not, in general, make this assumption.

4.5 The BLR Test

Now that we understand how dictatorship tests are used to obtain hardness results, let us try to make a more explicit definition of distance. The arguably most natural definition of distance between functions would be the *normalized Hamming distance*, defined for two functions $f, g : \Omega^n \to \Omega$ as

$$\mathrm{ham}(f, g) = \frac{|\{\, x \in \Omega^n \mid f(x) \neq g(x) \,\}|}{|\Omega|^n}.$$

Note that this was the notion of distance we used in the triangle-freeness case, when we said that the distance between two graphs G and G' is ϵ if we need to change

at least ϵn^2 entries of the adjacency matrix of G to obtain the adjacency matrix of G'.

The perhaps most classic result in this setting is the BLR linearity test by Blum, Luby and Rubinfeld [16]. As the name indicates, this is not a dictatorship test, but rather a *linearity* test. In particular, for any groups G and H, [16] shows how to test if a given function $f : G \to H$ is close to a homomorphism (i.e., if $f(ab) = f(a)f(b)$ for every $a, b \in G$) using only 3 queries. We will briefly describe their test in the setting $G = \{-1, 1\}^n$ (with coordinatewise multiplication) and $H = \{-1, 1\}$ (with multiplication). The classic Fourier-analytic proof for this setting given below is due to Bellare et. al [12].

Given a function $f : \{-1, 1\}^n \to \{-1, 1\}$, the BLR test works as follows: pick two random $x, y \in \{-1, 1\}^n$, and check that $f(xy) = f(x)f(y)$. It is clear that if f is linear, then the test accepts with probability 1. What may be more surprising is the fact that if f is ϵ-far from linear, the test accepts with probability at most $1 - \epsilon$, for any $\epsilon \in [0, 1/2]$ (note that every f is at most $1/2$-far from linear). Let us prove this. Arithmetizing the acceptance condition of the test, the probability that the test accepts can be written as

$$\Pr_{x,y}[f(x)f(y) = f(xy)] = \mathop{\mathbb{E}}_{x,y}\left[\frac{1 + f(x)f(y)f(xy)}{2}\right].$$

Expanding $f = \sum_{S\subseteq[n]} \hat{f}(S)\chi_S$ in terms of its Fourier coefficients and using linearity of expectation, this gives

$$\Pr_{x,y}[f(x)f(y) = f(xy)] = \frac{1}{2} + \frac{1}{2} \sum_{S,T,U\subseteq[n]} \hat{f}(S)\hat{f}(T)\hat{f}(U) \mathop{\mathbb{E}}_{x,y}[\chi_S(x)\chi_T(y)\chi_U(xy)].$$

However, by the linearity of $\chi_S(x) = \prod_{i\in S} x_i$, we have $\chi_U(xy) = \chi_U(x)\chi_U(y)$, i.e.,

$$\chi_S(x)\chi_T(y)\chi_U(xy) = \chi_S(x)\chi_U(x)\chi_T(y)\chi_U(y).$$

Hence

$$\mathop{\mathbb{E}}_{x,y}[\chi_S(x)\chi_T(y)\chi_U(xy)] = \mathop{\mathbb{E}}_{x}[\chi_S(x)\chi_U(x)] \mathop{\mathbb{E}}_{y}[\chi_T(y)\chi_U(y)] = \langle \chi_S, \chi_U \rangle \cdot \langle \chi_T, \chi_U \rangle,$$

which, by orthogonality of the χ functions, is 0 unless $S = T = U$, in which case it equals 1. Thus, the probability that the test accepts equals

$$\Pr_{x,y}[f(x)f(y) = f(xy)] = \frac{1}{2} + \frac{1}{2} \sum_{S\subseteq[n]} \hat{f}(S)^3 \leq \frac{1}{2} + \frac{1}{2}||\hat{f}||_\infty \cdot ||\hat{f}||_2 = \frac{1}{2} + \frac{1}{2} \max_{S\subseteq[n]} |\hat{f}(S)|.$$

But $|\hat{f}(S)|$ is exactly the maximum correlation between f and χ_S or $-\chi_S$, hence $\frac{1+|\hat{f}(S)|}{2}$ is exactly the maximum agreement of f with χ_S or $-\chi_S$, and the maximum of this quantity over all S is exactly the maximum agreement of f with any linear function (since any linear function from $\{-1, 1\}^n$ to $\{-1, 1\}$ is either of the form χ_S or of the form $-\chi_S$).

4.6 A Dictatorship Test Based on Distance

While the BLR test is a linearity test rather than a dictatorship test, it can be turned into a dictatorship test with some small modifications. The test we describe in this section was introduced by Håstad to prove $(1/2+\epsilon, 1-\epsilon)$-hardness for MAX 3-XOR in his seminal paper [54].

First, it turns out that it is too restrictive to require that every function f which is far from a dictator in Hamming distance should be rejected with high probability. For instance, the functions $f_{ij} : \{-1,1\}^n \to \{-1,1\}$ defined by $f_{ij}(x) = \chi_{ij}(x) = x_i x_j$ are very far from dictators (the hamming distance is $1/2$). But intuitively, it seems very difficult for a test to distinguish between these functions and dictators.

Let us then weaken the notion of distance somewhat, and instead ask for a test which is only required to reject f if it is not close to any linear function depending on a small number of variables. I.e., we only ask that the test rejects with good probability if $|\hat{f}(S)| \leq \delta$ for every $S \subseteq [n]$ with $|S| \leq d$ for some constants δ and d. Conversely, we say that f is close to DICT_i if $|\hat{f}(S)| > \delta$ for some S with $i \in S$ and $|S| \leq d$. Note that, with this definition, f can be close to at most $R = d/\delta^2$ different dictators, since there can be at most $1/\delta^2$ Fourier coefficients of size δ (since $\sum_S \hat{f}(S)^2 = 1$ for any $f : \{-1,1\}^n \to \{-1,1\}$).

Such a test can be constructed by simply adding a small amount of noise to the BLR test. The intuition behind this is that a dictator is likely not to be affected by noise, whereas a parity of a large number of variables have a high probability of being affected by the noise. Hence, we test f as follows: pick x, y uniformly in $\{-1,1\}^n$, and pick z in $\{-1,1\}^n$ such that each bit of z is -1 with probability η (which we think of as being very small), independently. Then check that $f(x)f(y) = f(xyz)$. Let us now analyze this test. First, we no longer have perfect completeness, but if f is a dictator, it is easy to see that it will be accepted with probability exactly $1-\eta$. For the soundness analysis, an analysis similar to the one for the BLR test will show that the acceptance probability of this test is

$$\begin{aligned} \Pr_{x,y,z}[f(x)f(y) = f(xyz)] &= \frac{1}{2} + \frac{1}{2}\sum_{S\subseteq[n]} (1-2\eta)^{|S|}\hat{f}(S)^3 \\ &\leq \frac{1}{2} + \frac{1}{2}\max_{S\subseteq[n]}(1-2\eta)^{|S|}|\hat{f}(S)|. \end{aligned}$$

This implies that if $|\hat{f}(S)| \leq \delta$ for every $|S| \leq d$, then the test accepts with probability at most

$$\frac{1 + \max(\delta, (1-2\eta)^d)}{2},$$

which is dominated by $(1+\delta)/2$ if d is a large enough constant compared to η.

There is still a problem with this test, which is the fact that the test accepts the two constant functions $\mathbf{1}$ and $-\mathbf{1}$ with probability 1, which is not acceptable since these are not close to any dictators. This is however easily remedied by *folding*,

which, as described earlier, allows us to assume that the function f is balanced and in particular that $\hat{f}(\emptyset) = 0$.

With an appropriate choice of the parameters δ, η and d, this test, plugged into a PCP verifier similar to the one described in Section 4.3, shows that for every $\epsilon > 0$, MAX 3-XOR is $(1/2+\epsilon, 1-\epsilon)$-UG-hard to approximate. To get unconditional $(1/2 + \epsilon, 1 - \epsilon)$-NP-hardness, as [54] does, is considerably more involved.

4.7 Influence-based Testing

Unfortunately, even the notion of distance used in the previous section is too strong for our purposes, in the sense that there are functions that are very far from every low-degree linear function in Hamming distance, but share important characteristics with some dictator DICT_i, making it hard to distinguish them from DICT_i.

In particular, consider a function $f : \{-1, 1\}^n \to \{-1, 1\}$ in which the variable i has large influence, but every single Fourier coefficient is small. Intuitively, f is similar to DICT_i, in the sense that when we flip the ith bit, f has a good probability of changing value (over a random choice of the other bits). Nevertheless, in the previous notion of distance, we would be required to reject it with good probability.

To illustrate that this is a problem, suppose we want to construct the test $\mathcal{T}$ from Section 4.3 for functions $f : \{-1, 1\}^n \to \{-1, 1\}$, using only 2 queries, but still having a good gap between completeness and soundness. The most natural way to construct such a test (and it is hard to imagine any radically different test which still makes sense) would be to pick a random string $x \in \{-1, 1\}^n$, and then flip each bit with some probability p, obtaining a string y. For concreteness, let us take $p = 9/10$. Then, we check that $f(x) \neq f(y)$, since in a dictator, this should happen with probability $9/10$. Similarly to the analysis of the previous tests, one can here prove that the accept probability equals

$$\frac{1}{2} - \frac{1}{2} \sum_{S \subseteq [n]} (1 - 2 \cdot (9/10))^{|S|} \hat{f}^2(S). \tag{4.2}$$

If one, as before, attempts to produce a bound on the second term in terms of the largest coefficient of f, one gets an expression of the form

$$- \sum_{S \subseteq [n]} (-4/5)^{|S|} \hat{f}^2(S) \leq \max_S |\hat{f}(S)| \cdot \sum_S (5/4)^{-|S|} |\hat{f}(S)|. \tag{4.3}$$

In general, this quantity can be huge. Consider the following example, due to Samorodnitsky and Trevisan [93]. Suppose f is defined as

$$f(x) = g(x_1, x_2) g(x_3, x_4) \dots g(x_{n-1}, x_n),$$

where $g : \{-1, 1\}^2 \to \{-1, 1\}$ is the "AND" function on $\{-1, 1\}$, defined as $g(x_1, x_2) = -1$ if and only if $x_1 = x_2 = -1$. Then it is easily verified that $f(x)$ can be written

as follows

$$f(x) = \prod_{i=1}^{n/2} \left(\frac{1 + x_{2i} + x_{2i+1} - x_{2i}x_{2i+1}}{2} \right).$$

This implies that, for every $S \subseteq [n]$, $|\hat{f}(S)| = 2^{-n/2}$. In particular, we get that

$$\sum_S (5/4)^{-|S|} |\hat{f}(S)| = \sum_{k=0}^{n} (5/4)^{-k} \binom{n}{k} 2^{-n/2} \geq \sum_{k=0}^{n/2} (5/2)^{-n/2} \binom{n}{k} \geq \frac{1}{2}(8/5)^{n/2}$$

Hence the bound in Equation (4.3) is completely useless. Of course, this does not prove that the test does not work—the bound in Equation (4.3) is a quite poor estimate. Still, it shows that if we want to prove that this test works, we will need a more sophisticated analysis of Equation (4.2)—we will return to this momentarily.

Let us now look at the influences of f. It is easily computed that for every i, the influence $\mathrm{Inf}_i(f) = \frac{1}{2}$ is very large. Hence, f is "similar" to a dictator in the sense of having high influences. Thus, one might be tempted to revise the notion of "distance" to say that f is far from a dictator if all influences of f are small. However, it turns out that this does not work, because this notion of distance is in general *too weak*, so that we would not be able to turn such dictatorship tests into hardness results. Recall that, as described in Section 4.3, the key property that is needed for a hardness result, is that no function can be close to more than a small number of dictators. In the case of influences, most notably for the function used above, *every* variable has large influence and hence f is "close" to every dictator.

A better notion is then, similarly to the change we made from linearity test to dictator test in Section 4.6, to look at low-degree influence instead, and only require that f is rejected with good probability, if for every i, $\mathrm{Inf}_i^{\leq d}(f) \leq \tau$, for some constants d and τ. Recall that by Proposition 2.3.12, the number of variables having a large low-degree influence is bounded, and hence this notion could be used to obtain hardness results. Note however that our "counterexample" function f has all its low-degree influences small, $\mathrm{Inf}_i^{\leq d}(f) \approx n^d 2^{-n}$, hence we still need that this function is rejected with good probability. Thus we have returned to the problem of more carefully analyzing quantities such as Equation (4.2), this time for functions f with small low-degree influences. This is where the very powerful Theorems 2.5.1 and 2.5.2 come into play, as these give exactly this type of bounds. Recall the special case of noise correlation $\langle f, f \rangle_{\mathcal{N}}$ called noise stability, $\mathbb{S}_\rho(f)$, mentioned in Section 2.4. For the particular test in question, Theorem 2.5.2 gives that the probability that the test accepts is bounded by (ignoring an additive ϵ)

$$\frac{1 - \mathbb{S}_\rho(f)}{2} \leq 1 - \mathbb{E}[f] - 2\Gamma_\rho(\mathbb{E}[f], \mathbb{E}[f]),$$

where $\rho = -4/5$ (this can be verified using Proposition 2.2.8). Furthermore, suppose we enforce f to be balanced by using folding, so that $\mathbb{E}[f] = 0$. Then, we get that the soundness of the test is $1 - 2\Gamma_{-4/5}(0, 0)$, which turns out to

be around 0.7952, meaning that we can get $(0.7952, 0.9000)$-UG-hardness for the Max 2-Xor problem, and in particular that Max 2-Xor is hard to approximate within $\approx 0.7952/0.9 \approx 0.8835$, which comes very close to matching the ratio of ≈ 0.8785 of the classic Goemans-Williamson algorithm [42]. This concluding discussion can be viewed as a teaser for Chapter 6, which consists of a more general hardness result in this flavor.

Chapter 5

Constraints on Many Variables

In this chapter, we study the approximability of constraint satisfaction problems on k-ary constraints, in particular the MAX k-CSP$_q$ and MAX CSP(P) problems for objective functions $P : [q]^k \to [0, 1]$.

Let us start with considering the case when $q = 2$, i.e., the case of boolean variables. A particularly simple approximation algorithm is the algorithm which simply picks a random assignment to the variables. For MAX k-CSP, this algorithm has a ratio of $1/2^k$. It was first improved by Trevisan [101] who gave an algorithm with ratio $2/2^k$. Recently, Hast [52] gave an algorithm with ratio $\Omega(k/(2^k \log k))$, which was subsequently improved by Charikar et al. [21] who gave an algorithm with approximation ratio $c \cdot k/2^k$, where $c > 0.44$ is an absolute constant.

The PCP Theorem implies that the MAX k-CSP problem is NP-hard to approximate within $1/c^k$ for some constant $c > 1$. Samorodnitsky and Trevisan [93] improved this hardness to $2^{2\sqrt{k}}/2^k$, and this was further improved to $2^{\sqrt{2k}}/2^k$ by Engebretsen and Holmerin [31]. Finally, Samorodnitsky and Trevisan [94] proved that if the Unique Games Conjecture is true, then the MAX k-CSP problem is hard to approximate within $2k/2^k$. To be more precise, the hardness they obtained was $2^{\lceil \log_2 k+1 \rceil}/2^k$ (note that the numerator is simply $k+1$ rounded up to the next power of two), which is $(k+1)/2^k$ for $k = 2^r - 1$, but can be as large as $2k/2^k$ for general k. Thus, the current gap between hardness and approximability is a small constant factor of $2/0.44$.

For MAX k-CSP$_q$, the random assignment gives a $1/q^k$-approximation. The algorithm of Charikar et al. for $q = 2$ can be used to obtain a $0.44k\lfloor \log_2 q \rfloor/q^k$-approximation for general q. The best previous inapproximability for the MAX k-CSP$_q$ problem is due to Engebretsen [30], who showed that the problem is NP-hard to approximate within $q^{\mathcal{O}(\sqrt{k})}/q^k$.

For a predicate $P : [q]^k \to \{0, 1\}$, the random assignment algorithm achieves a ratio of m/q^k for the MAX CSP(P) problem, where m is the number of satisfying assignments of P. Surprisingly, it turns out that for certain choices of P, this is the best possible algorithm. As described in Section 4.6, Håstad [54] proved (among

other things) that the MAX 3-XOR problem is hard to approximate within $1/2+\epsilon$, whereas a random assignment gives a factor $1/2$.

In general, predicates P for which it is hard to approximate the MAX CSP(P) problem better than a random assignment are called *approximation resistant*. A very natural and important question in hardness of approximation is to understand the structure of approximation resistance. For $k = 2$, it is known that no predicates are resistant [42, 55]. For $k = 3$ and $q = 2$, it is known that a predicate is approximation resistant if and only if it is implied by an XOR of the three variables [54, 106]. For $k = 4$ and $q = 2$, Hast [53] managed to classify most of the predicates with respect to approximation resistance, but already for this case there does not appear to be as nice a characterization as in the case $k = 3$.

A slightly stronger notion is that of *hereditary* approximation resistance—a predicate P is hereditarily approximation resistant if all predicates implied by P are approximation resistant. It turns out that, assuming the Unique Games Conjecture, most predicates are in fact hereditarily approximation resistant. In particular, Håstad [56] showed that a random predicate on k boolean variables, with $\approx 2^k/\sqrt{k}$ satisfying assignments is hereditarily resistant with probability $1 - o(1)$ (hence a random predicate with more than this many satisfying assignments is also resistant with high probability). Thus, instead of attempting to understand the structure of approximation resistant predicates (which by the work of [53] seems quite complicated), one might try to understand the possibly easier structure of hereditary approximation resistant predicates, as the non-hereditarily resistant predicates constitute a negligible fraction of *all* predicates.

Additionally, approximation resistance is useful for proving inapproximability results for MAX k-CSP$_q$ in general—a natural approach for obtaining such inapproximability is to search for approximation resistant predicates with very few accepting inputs. This is indeed how all mentioned hardness results for MAX k-CSP$_q$ come about (except the one implied by the PCP Theorem).

In this chapter, we will describe a generic hardness result for objective functions $P : [q]^k \to [0, 1]$ on k variables over some domain $[q]$. Loosely speaking, the strength of the result will be closely related to whether there exists a pairwise independent distribution $\mu : [q]^k \to [0, 1]$ which "coincides" with the objective function. Here, "coincides" means that most of the support of μ is on assignments x where $P(x)$ is large.

Our result implies a very general sufficient condition for a predicate $P : [q]^k \to \{0, 1\}$ to be hereditarily approximation resistant, and gives sharper hardness for MAX k-CSP$_q$.

In addition, it can be used to obtain very strong bounds on the number of approximation resistant predicates, significantly improving the previous bounds [56]. This is discussed in Section 5.3.

5.1 Hardness from Pairwise Independence

The main theorem of this chapter is the following.

Theorem 5.1.1. *Let $P : [q]^k \to [0,1]$ be an objective function, and let μ be a balanced pairwise independent distribution over $[q]^k$. Then, for every $\epsilon > 0$, the* Max CSP*(P) problem is $(s+\epsilon, c-\epsilon)$-UG-hard, with*

$$\begin{aligned} s &= \mathop{\mathbb{E}}_{x \in ([q]^k, \mu_U)}[P(x)] \\ c &= \mathop{\mathbb{E}}_{x \in ([q]^k, \mu)}[P(x)], \end{aligned}$$

where μ_U denotes the uniform distribution over $[q]^k$.

Before we prove this theorem, let us make some remarks.

Note that, for any instance of Max CSP(P), a random assignment has expected value s, and hence we can not hope to improve the $s+\epsilon$ part of the $(s+\epsilon, c-\epsilon)$-hardness. Furthermore, since μ only shows up in the value of c in the conclusion of the theorem, the theorem is at its strongest when we choose μ so as to maximize c, i.e., so as to be as "contained" in P as possible. In other words, the theorem can be equivalently restated as saying that Max CSP(P) is $(s+\epsilon, c-\epsilon)$-hard where

$$c = \sup_{\substack{\mu \text{ balanced} \\ \text{pairwise independent}}} \mathop{\mathbb{E}}_{x \in ([q]^k, \mu)}[P(x)].$$

The main application of this theorem is in the case where $P : [q]^k \to \{0,1\}$ is a predicate such that $P^{-1}(1) \supseteq \mathrm{Supp}(\mu)$, i.e., when the set of satisfying assignments completely contains the support of the pairwise independent distribution μ. In this setting, we have $c = 1$, and conclude that Max CSP(P) is $(s+\epsilon, 1-\epsilon)$-UG-hard for every $\epsilon > 0$. In other words, P is approximation resistant.

Proof of Theorem 5.1.1. As described in Chapter 4, we will, for every $\epsilon > 0$ construct a PCP verifier for L-Unique Label Cover which uses P as its acceptance condition, has completeness $c - \epsilon$ and soundness $s + \epsilon$.

Let $\gamma := \gamma(\epsilon, k, q) > 0$ be a parameter, which will be chosen as a small enough function of ϵ, k, and q.

Given an L-Unique Label Cover instance $\Psi = (X, Y, E, \Pi)$, a proof that $\mathrm{Opt}(\Psi) \geq 1 - \gamma$ consists of functions $f_w : [q]^L \to [q]$ for every $w \in Y$. In a proper proof, $f_w = \mathrm{Dict}_{\ell(w)}$ is simply the dictator function corresponding to the label of w, for some labeling ℓ such that $\mathrm{Val}(\ell) \geq 1 - \gamma$. Furthermore, the verifier will assume that each f_w is *balanced*, i.e., that for each $a \in [q]$, $f_w(x) = a$ for exactly q^{n-1} inputs x. As described in Section 4.4, it can be enforced by *folding*.

Define the probability distribution μ' on $[q]^k$ by $\mu' = (1 - \frac{\epsilon}{2})\mu(x) + \frac{\epsilon}{2}\mu_U(x)$. A supposed proof $\Sigma = \{f_w\}_{w \in Y}$ is verified as follows.

Algorithm 1: The verifier $\mathcal{V}$
$\mathcal{V}(\Psi, \Sigma = \{f_w\}_{w \in Y})$
(1) Pick a vertex $v \in X$ uniformly at random
(2) Let X be a random $k \times L$ matrix s.t. the distribution of the ith column of X is μ', independently of the other columns
(3) **foreach** $i \in [k]$
(4) Pick $e_i = \{v_i, w_i\}$ with permutation π_i uniformly at random from $E(v)$
(5) Let $a_i = f_{w_i}(X_i \circ \pi_i^{-1})$
(6) Accept with probability $P(a_1, \ldots, a_k)$

We now need to prove that the completeness and soundness of this PCP verifier is at least $c - \epsilon$ and at most $s + \epsilon$, respectively. Let us begin with the completeness.

Lemma 5.1.2 (Completeness)**.** *If* $\mathrm{Opt}(\Psi) \geq 1 - \gamma$*, then there exists a proof* Σ *such that*

$$\Pr[\mathcal{V}(\Psi, \Sigma) \text{ accepts}] \geq c - \epsilon,$$

provided $\gamma \leq \epsilon/(2k)$*.*

Proof. Let $\ell : X \cup Y \to [L]$ be a labeling of Ψ such that $\mathrm{Val}(\ell) \geq 1 - \gamma$, and define a proof Σ by letting $f_w = \textsc{Dict}_{\ell(v)}$ for every $w \in Y$.

First, assume that all the k edges $e_1, \ldots, e_k$ chosen by $\mathcal{V}$ are satisfied by ℓ. For $i \in [k]$, define $h_i(x) = f_{w_i}(x \circ \pi_i^{-1})$. Then, for each i,

$$h_i = \textsc{Dict}_{\pi_i^{-1}(\ell(w_i))} = \textsc{Dict}_{\ell(v)},$$

the second inequality using the assumption that e_i is satisfied by ℓ. This implies that since $a_i = h_i(X_i)$, $(a_1, \ldots, a_k)$ is simply the $\ell(v)$th column of X, which in turn has distribution μ'. Hence the probability that $\mathcal{V}$ accepts in this case is exactly

$$\Pr_{x \in ([q]^k, \mu')}[P(x)] \geq (1 - \epsilon/2) \Pr_{x \in ([q]^k, \mu)}[P(x)] \geq c - \epsilon/2.$$

Finally, since $\mathrm{Val}(\ell) \geq 1 - \gamma$, each e_i has probability at most γ of not being satisfied by ℓ and so by a union bound, the probability that all the k edges $e_1, \ldots, e_k$ are satisfied by ℓ is at least $1 - k\gamma \geq 1 - \epsilon/2$. Hence, the overall accept probability of $\mathcal{V}(\Psi, \Sigma)$ can be lower-bounded by

$$(1 - \epsilon/2)(c - \epsilon/2) \geq c - \epsilon.$$

□

Let us then move to the soundness of $\mathcal{V}$.

Lemma 5.1.3 (Soundness)**.** *There is a function $\gamma := \gamma(\epsilon, k, q)$, such that if $\mathrm{Opt}(\Psi) \leq \gamma$, then for every proof Σ,*

$$\Pr[\mathcal{V}(\Psi, \Sigma) \text{ accepts}] \leq s + \epsilon,$$

Proof. For an arbitrary $v \in X$, let p_v denote the probability of $\mathcal{V}(\Psi, \Sigma)$ accepting conditioned on v being the vertex chosen in line (1) of $\mathcal{V}$.

Assume for contradiction that $\mathcal{V}(\Psi, \Sigma)$ accepts with probability larger than $s+\epsilon$. Then, there is a set $V \subseteq X$ of relative size at least $\epsilon/2$ such that for every $v \in V$, $p_v \geq s + \epsilon/2$.

Consider now an arbitrary $v \in V$. Since

$$\begin{aligned} p_v &= \sum_{y \in [q]^k} P(y) \Pr[a_1 = y_1 \wedge a_2 = y_2 \ldots \wedge a_k = y_k | v] \\ &> s + \epsilon/2 = \sum_{y \in [q]^k} \frac{P(y)}{q^k} + \epsilon/2, \end{aligned}$$

and $0 \leq P(y) \leq 1$ for every y, there must be some $y^* \in [q]^k$ such that

$$\Pr[a_1 = y_1^* \wedge a_2 = y_2^* \ldots \wedge a_k = y_k^* | v] > \frac{1 + \epsilon/2}{q^k}.$$

For an edge $e = (v, w) \in E(v)$ and $z \in [q]$, define the indicator function $g_{e,z} : [q]^L \to \{0, 1\}$ by

$$g_{e,z}(x) = \begin{cases} 1 & \text{if } f_w(x \circ \pi_e^{-1}) = z \\ 0 & \text{otherwise} \end{cases}.$$

Then,

$$\begin{aligned} \Pr[a_1 = y_1^* \wedge a_2 = y_2^* \ldots \wedge a_k = y_k^* | v] &= \mathop{\mathbb{E}}_{X, e_1, \ldots, e_k} \left[\prod_{i=1}^{k} g_{e_i, y_i^*} \right] \\ &= \mathop{\mathbb{E}}_{X} \left[\prod_{i=1}^{k} \mathop{\mathbb{E}}_{e \in E(v)} [g_{e, y_i^*}(X_i)] \right], \end{aligned}$$

where we used the independence of $e_1, \ldots, e_k$. Defining $g_z : [q]^L \to [0, 1]$ by $g_z(x) = \mathbb{E}_{e \in E(v)}[g_{e,z}(x)]$, we thus have

$$\left\langle g_{y_1^*}, \ldots, g_{y_k^*} \right\rangle_{\mathcal{N}} = \mathop{\mathbb{E}}_{X} \left[\prod_{i=1}^{k} g_{y_i^*}(X_i) \right] > \frac{1 + \epsilon/2}{q^k}.$$

Now, note that, since μ' is balanced, and f_w is balanced for every $w \in Y$, we have $\mathbb{E}[g_z] = \mathbb{E}[g_{e,z}] = 1/q$ for every $e \in E(v)$ and $z \in [q]$. Hence, the equation above can be reformulated as

$$\left\langle g_{y_1^*}, \ldots, g_{y_k^*} \right\rangle_{\mathcal{N}} > \prod_{i=1}^{k} \mathbb{E}[g_{y_i^*}] + \frac{\epsilon}{2q^k}.$$

It is easily checked that μ' satisfies the conditions of Theorem 2.5.1, and hence there exist τ and d depending only on ϵ, q and k such that, for some $i \in [k]$ we have

$$\mathrm{Inf}_i^{\leq d}(g_{y_i^*}) \geq \tau. \tag{5.1}$$

Furthermore, since

$$\begin{aligned}\mathrm{Inf}_i^{\leq d}(g_{y_i^*}) &= \sum_{\substack{i\in\sigma\\|\sigma|\leq d}} \widehat{g_{y_i^*}}(\sigma)^2 = \sum_{\substack{i\in\sigma\\|\sigma|\leq d}} \mathop{\mathbb{E}}_{e\in E(v)}[\widehat{g_{e,y_i^*}}(\sigma)]^2 \\ &\leq \sum_{\substack{i\in\sigma\\|\sigma|\leq d}} \mathop{\mathbb{E}}_{e\in E(v)}[\widehat{g_{e,y_i^*}}(\sigma)^2] = \mathop{\mathbb{E}}_{e\in E(v)}[\mathrm{Inf}_i^{\leq d}(g_{e,y_i^*})], \end{aligned} \tag{5.2}$$

we must have that for at least a fraction $\tau/2$ of all $e \in E(v)$, $\mathrm{Inf}_i^{\leq d}(g_{e,y_i^*}) \geq \tau/2$. Recall that for $e = (v, w)$, the function g_{e,y_i^*} is the indicator function $\mathbf{1}_{f_w=y_i^*}$ except that the inputs are permuted according to π^{-1},

$$g_{e,y_i^*}(x) = \mathbf{1}_{f_w=a}(x \circ \pi_i^{-1}).$$

Hence for at least a $\tau/2$ fraction of all edges $e = (v, w) \in E(v)$, we have

$$\mathrm{Inf}_{\pi(i)}^{\leq d}(\mathbf{1}_{f_w=y_i^*}) = \mathrm{Inf}_i^{\leq d}(g_{e,y_i^*}) \geq \tau/2. \tag{5.3}$$

Now we can define small sets of candidate labels for each vertex $v \in X \cup Y$. For $v \in X$, set

$$C(v) = \{\, i \in [L] \,:\, \mathrm{Inf}_i^{\leq d}(g_{v,a}) \geq \tau \text{ for some } a \in [q] \,\},$$

and for $w \in Y$, set

$$C(w) = \{\, i \in [L] \,:\, \mathrm{Inf}_i^{\leq d}(\mathbf{1}_{f_w=a}) \geq \tau/2 \text{ for some } a \in [q] \,\}.$$

Note that by Proposition 2.3.12 we have that $|C(v)| \leq \frac{dq}{\tau}$, and that $|C(w)| \leq \frac{2dq}{\tau}$. Now define a random labeling $\ell : X \cup Y \to [L]$ of Ψ by letting, for each $v \in X \cup Y$, $\ell(v)$ be a uniformly random element of $C(v)$ (or an arbitrary element of $[L]$ in case $C(v)$ is empty).

Then, for every $v \in V$, by Equation (5.1), $C(v)$ is non-empty and hence by Equation (5.3) there is some $i \in C(v)$ such that, for at least a $\tau/2$ fraction of all $e = (v, w) \in E(v)$, we also have $\pi(i) \in C(w)$. Any such edge has a probability at least

$$(|C(v)| \cdot |C(w)|)^{-1} \geq \frac{\tau^2}{2d^2q^2}$$

of being satisfied by ℓ, and these edges constitute a total fraction of $\frac{\epsilon\tau}{4}$ of all edges of Ψ. Thus

$$\mathbb{E}[\mathrm{Val}(\ell)] \geq \frac{\epsilon\tau^3}{8d^2q^2} =: \gamma(\epsilon, k, q)$$

(recall that τ and d depend only on ϵ, k, and q). This concludes the proof of the soundness lemma. □

Finally, the proof of Theorem 5.1.1 follows immediately from Lemma 5.1.2 and Lemma 5.1.3. □

5.2 Implications for Max k-CSP$_q$

In light of Theorem 5.1.1, a natural way of trying to obtain strong inapproximability results for MAX k-CSP$_q$ is to construct pairwise independent distributions over $[q]^k$ with small support. We have the following easy corollary of Theorem 5.1.1.

Corollary 5.2.1. *Let μ be a balanced pairwise independent distribution over $[q]^k$, and let $t = |\operatorname{Supp}(\mu)|$. Then for every $\epsilon > 0$, the* MAX k-CSP$_q$ *problem is $(t/q^k + \epsilon, 1 - \epsilon)$-UG-hard.*

In particular the predicate P defined by $P^{-1}(1) = \operatorname{Supp}(\mu)$ is $(t/q^k + \epsilon, 1 - \epsilon)$-UG-hard. Using the constructions of pairwise independent distributions from Section 3.5, this gives the following theorems.

Theorem 5.2.2. *For every $\epsilon > 0$, the* MAX k-CSP *problem is UG-hard to approximate within a factor*

$$\frac{k + \mathcal{O}(k^{0.525})}{2^k} + \epsilon.$$

If the Hadamard Conjecture is true, then for every $\epsilon > 0$ the MAX k-CSP *problem is UG-hard to approximate within a factor*

$$\frac{4\lceil (k+1)/4 \rceil}{2^k} + \epsilon \leq \frac{k+4}{2^k} + \epsilon.$$

We remark that the predicates of Samorodnitsky and Trevisan [94] can be obtained as a special case of this result when one constructs a pairwise independent distribution from a Hadamard matrix of dimension $2^l \times 2^l$. We would also like to stress that, while the factor 2 improvement from $2^{\lceil \log_2 k+1 \rceil}/2^k$ to $4\lceil (k+1)/4 \rceil/2^k$ may not seem huge, it is interesting because of the fact that $4\lceil (k+1)/4 \rceil$ is almost the optimal possible value. In particular, Hast [53] showed that any predicate with fewer than $2\lceil (k+1)/2 \rceil$ accepting assignments is *not* approximation resistant. Hence, for $k \equiv 2, 3 \pmod 4$, our result is exactly tight, and for $k \equiv 0, 1 \pmod 4$, the number of accepting assignments in our predicates is off by an *additive* error of 2.

For large q, we first have the following result.

Theorem 5.2.3. *For every prime power q and $\epsilon > 0$, the* MAX k-CSP$_q$ *problem is UG-hard to approximate within a factor*

$$\frac{kq(q-1)}{q^k} + \epsilon.$$

If $k = (q^r - 1)/(q - 1)$ for some $r > 1$, then the MAX k-CSP$_q$ *problem is UG-hard to approximate within a factor*

$$\frac{k(q-1)+1}{q^k} + \epsilon.$$

By a nice observation due to Yury Makarychev, Theorem 5.2.3 also extends to the case when q is not a prime power with a very small additional loss.

Theorem 5.2.4. *For every $q \geq 2$ and $\epsilon > 0$, the* MAX k-CSP$_q$ *problem is UG-hard to approximate within a factor*

$$\frac{kq^2(1+o(1))}{q^k} + \epsilon.$$

See Appendix B of [9] for further details.

5.3 Sharper Bounds for Random Predicates

The characterization of approximation resistance in terms of pairwise independence is useful not only for deriving stronger hardness for MAX k-CSP$_q$—it also gives very strong estimates on the number of resistant predicates.

In particular, fix a number $0 \leq t \leq q^k$, and consider the set of all predicates $P : [q]^k \to \{0, 1\}$ with exactly t satisfying assignments. How many of these are approximation resistant? Let us normalize, and instead ask: what is the probability $p(t)$ that a random predicate with t accepting assignments is UG-approximation resistant?

In light of approximation resistance almost always being hereditary, it seems natural to believe that $p(t)$ should be an increasing function: if a large part of the $p(t)$ fraction of resistant predicates with t accepting assignments are hereditarily resistant, then these will contribute a lot to $p(t+1)$.

For q a prime power, Theorem 5.2.3 shows that there is a constant $c := q^2$ such that if $t \geq c \cdot k$, we have $p(t) > 0$. For $q = 2$, Theorem 5.2.4 shows that assuming the Hadamard Conjecture, if $t \geq 4\lceil (k+1)/4 \rceil$, we have $p(t) > 0$. As mentioned in the previous section, this is close to optimal, as Hast proved that for $t < 2\lceil (k+1)/2 \rceil$, $p(t) = 0$ (strictly speaking, with $p(t)$ defined as above, this assumes that the Unique Games Conjecture is true, since if it is false it vacuously implies that every predicate is UG-approximation resistant, i.e., that $p(t) = 1$).

Let us now ask: how large does t have to be in order to have $p(t) > 1 - o(1)$? The first answer to this question was given by Håstad [56], who proved that for $q = 2$, it suffices to take $t > 2^k/k^c$, where $c \in [1/2, 1]$ is a constant the exact value of which depends on how close k is to a power of 2. Håstad proves that with high probability, such a predicate is implied by the predicates constructed by Samorodnitsky and Trevisan [94]. Furthermore, he proves that this bound on t is essentially optimal in the sense that predicates having fewer satisfying assignments

are very unlikely to be implied by the predicates of Samorodnitsky and Trevisan. As our results for boolean predicates are only up to a factor 2 better than those of Samorodnitsky and Trevisan, it is very unclear whether they can be used to improve this bound. However, an immediate corollary of the results of Chapter 8, in particular of Theorem 8.4.1, we have the following dramatic improvement.

Theorem 5.3.1. *For every q there exists a c such that if $t > ck^2$, then*

$$p(t) > 1 - \exp(-\Theta(\sqrt{k})),$$

where $p(t)$ is the fraction of predicates $P : [q]^k \to \{0,1\}$ with t accepting inputs which are approximation resistant under the UGC.

Finally, let us ask how large t has to be in order to actually have $p(t) = 1$. As far as we are aware, the first such result was given by Hast [53] for the case $q = 2$, who proved that it suffices to take $t > 2^k(1-2^{-\sqrt{k}})$ (this result is NP-hardness, not UG-hardness). This was improved by Håstad [56], who proved that one can take $t > 2^k\left(1 - \frac{1}{2k}\right)$. Again using the results of Chapter 8, this time Corollary 8.3.2, we improve this and show that a sufficiently large *constant* fraction of satisfying assignments suffices.

Theorem 5.3.2. *For every q there exists a δ such that if $t > (1-\delta)q^k$, then $p(t) = 1$, where $p(t)$ is the fraction of predicates $P : [q]^k \to \{0,1\}$ with t accepting inputs which are approximation resistant under the UGC.*

We remark that the constants c and δ in Theorem 5.3.1 and Theorem 5.3.2 can both be taken as polynomials in q. In the case of Theorem 5.3.1 this is particularly interesting, because for q not a prime power we do not know of explicit constructions achieving this bound.

Chapter 6

Constraints on Two Variables

In this chapter, we describe algorithms and hardness results for Max CSP(P) problems where $P : \{-1, 1\}^2 \to [0, 1]$ is an objective function on two boolean variables. This class of CSPs contains such fundamental problems as Max Cut and Max 2-Sat. As opposed to the CSP problems considered in the previous chapter, these are problems in which there are algorithms achieving approximation ratios better than the random assignment, and the amount of work needed to obtain sharp results is significantly greater than in the previous chapter.

For the general Max 2-CSP problem, the random assignment algorithm achieves an approximation ratio of 1/4. For the special cases of Max Cut and Max 2-Sat, it achieves ratios of 1/2 and 3/4, respectively. For several decades, no substantial improvements were made over these results until a seminal paper by Goemans and Williamson [42], where they constructed a 0.7960-approximation algorithm for Max 2-CSP, and 0.87856-approximation algorithms for Max Cut and Max 2-Sat. To do so, they relaxed the combinatorial problem at hand to a semidefinite programming problem, to which an optimal solution can be found with high precision, and then used a very clever technique to "round" the solution of the semidefinite programming back to a discrete solution for the original problem. This approach has since been successfully applied to several other hard combinatorial optimization problems, yielding significant improvements over existing approximation algorithms. Examples include coloring graphs using as few colors as possible [61, 15, 51, 3], Max Bisection [38] and quadratic programming over the boolean hypercube [22].

Some of the results by Goemans and Williamson were subsequently improved by Feige and Goemans [33], who strengthened the semidefinite relaxation using certain triangle inequalities [42]. They obtained a 0.931-approximation for Max 2-Sat, and a 0.859-approximation for Max 2-CSP. These results were further improved by Matuura and Matsui [76, 77], who obtained a 0.935-approximation for Max 2-Sat and a 0.863-approximation for Max 2-CSP. Shortly thereafter, Lewin et al. [71] obtained further improvements, getting a 0.94017-approximation

algorithm for MAX 2-SAT and a 0.87401-approximation algorithm for MAX 2-CSP, and these stand as the current best algorithms. It should be pointed out that these last two ratios arise as the minima of two complex numeric optimization problems, and it has not yet been proved formally that these are the actual ratios, though there seems to be very little doubt that this is indeed the case.

When it comes to *inapproximability*, the best NP-hardness results for these problems are obtained using Håstad's famous results for 3-CSPs [54]. The best NP-hardness results for MAX 2-CSP, MAX 2-SAT, and MAX CUT are $9/10+\epsilon \approx 0.900$, $21/22+\epsilon \approx 0.955$, and $16/17+\epsilon \approx 0.941$, respectively [103, 54]. Unfortunately, getting improved NP-hardness results for these problems seems to be beyond the reach of current techniques.

However, it again turns out that under the Unique Games Conjecture, one can prove remarkably strong results. Khot et al. [63] showed that the UGC implies $\alpha_{GW}+\epsilon$ hardness for MAX CUT, where $\alpha_{GW} \approx 0.87856$ is the performance ratio of the original Goemans-Williamson algorithm. Since then, many hardness results have been obtained with hardness ratio matching, or almost matching, the approximation ratio given by the best algorithms based on semidefinite programming. This chapter gives such hardness results for 2-CSP problems. Subsequent to our work, Raghavendra [87] obtained a much more general result applying to k-CSPs over any domain. We discuss this result in Section 6.8.

6.1 Our Contribution

We investigate the approximability of the MAX CSP(P) problem and explore the tight connection between semidefinite programming relaxations and the UGC. Following the paradigm introduced by Goemans and Williamson, we relax the MAX CSP(P) problem to a semidefinite programming problem. We then consider the following approach for rounding the relaxed solution to a boolean solution: given the SDP solution, we pick the "best" rounding from a certain class of randomized rounding methods (based on skewed random hyperplanes), where "best" is in the sense of giving a boolean assignment with maximum possible expected value. Informally, let $\alpha(P)$ denote the approximation ratio yielded by such an approach. We then have the following theorem.

Theorem 6.1.1. *For any objective function $P : \{-1,1\}^2 \to [0,1]$ and $\epsilon > 0$,* MAX CSP(P) *problem can be approximated within $\alpha(P) - \epsilon$ in polynomial time.*

The reason that we lose an additive ϵ is that we are not, in general, able to find the *best* rounding function, but we can come arbitrarily close.

Then, we turn our attention to hardness of approximation. Here, we are able to take instances which are hard to round, in the sense that the best rounding (as described above) is not very good, and translate them into a Unique Games-based hardness result. There is, however, a caveat: in order for the analysis to work, the instance needs to satisfy a certain "positivity" condition. Again, informally, let

$\beta(P)$ denote the approximation ratio when restricted to instances satisfying this condition. We then have:

Theorem 6.1.2. *For any objective function $P : \{-1,1\}^2 \to [0,1]$ and $\epsilon > 0$, the* Max CSP(P) *problem is UG-hard to approximate within $\beta(P) + \epsilon$.*

Both $\alpha(P)$ and $\beta(P)$ are the solutions to a certain numeric minimization problem. The function being minimized is the same function in both cases, the only difference is that in $\alpha(P)$, the minimization is over a larger domain, and thus, we could potentially have $\alpha(P) < \beta(P)$. However, there are strong indications that the minimum for $\alpha(P)$ is in fact obtained within the domain of $\beta(P)$, in which case they would be equal and Theorems 6.1.1 and 6.1.2 would be tight.

Conjecture 6.1.3. *For any objective function $P : \{-1,1\}^2 \to [0,1]$, we have $\alpha(P) = \beta(P)$.*

Because of the difficulty of actually computing the approximation ratios $\alpha(P)$ and $\beta(P)$, it may seem somewhat difficult to compare these results to previous results. However, previous algorithms and hardness results for Max Cut, Max 2-Sat, and Max 2-CSP can all be obtained as special cases of Theorems 6.1.1 and 6.1.2. In particular, for $P(x_1, x_2) = x_1 \oplus x_2$, the Xor predicate, it can be shown that $\alpha(P) = \beta(P) = \alpha_{GW}$.

We are also able to use Theorem 6.1.2 to obtain new inapproximability results for specific problems. In particular, we obtain improved hardness for the Max 2-Sat and Max 2-And problems. For Max 2-Sat, Khot et al. [63] proved UG-hardness of ≈ 0.9439, almost matching the ratio $\alpha_{LLZ} \approx 0.94017$ of the algorithm of Lewin et al. Furthermore, their hardness was for *balanced* Max 2-Sat, which is the special case when every variable occurs positively and negatively equally often, and they also proved that for balanced Max 2-Sat their result was tight, by giving a matching algorithm. It is natural to conjecture, especially considering these results, that balanced instances should be the hardest (and indeed, Khot et al. [63] do that). However, as we prove, the algorithm of Lewin et al. is in fact optimal if the UGC is true.

Theorem 6.1.4. *For the predicate $P(x_1, x_2) = x_1 \vee x_2$, we have $\beta(P) \leq \alpha_{LLZ} \approx 0.94017$.*

For Max 2-And, there was a similar state of affairs. Khot et al. [63] proved hardness of $\alpha_{GW} + \epsilon$, where $\alpha_{GW} \approx 0.87865$ is the approximation ratio of the Goemans-Williamson Max Cut algorithm, again for the special case of *balanced* Max 2-And and again with a matching algorithm for balanced Max 2-And. This comes quite close to matching the approximation ratio of 0.87401 of Lewin et al. We improve the inapproximability for Max 2-And and demonstrate that, assume the UGC, the algorithm of Lewin et al. is very close to being optimal.

Theorem 6.1.5. *For the predicate $P(x_1, x_2) = x_1 \wedge x_2$, we have $\beta(P) \leq 0.87435$.*

This comes very close to matching the 0.87401-approximation algorithm of Lewin et al. It also again demonstrates that balanced instances are not the hardest to approximate.

It also implies improved hardness for the MAX 2-CSP problem—as is well-known, the MAX k-CSP problem and the MAX k-AND problem are equally hard to approximate for every k (folklore, or see e.g. [101]). This demonstrates that the MAX 2-CSP problem is harder to approximate than MAX CUT, i.e., that MAX CUT is not the hardest 2-CSP, which was not known prior to our work. The best approximation algorithm for MAX 2-CSP is the 0.87401-approximation algorithm for MAX 2-AND, and the previous best hardness for MAX 2-CSP was the hardness of $\alpha_{GW} \approx 0.87865$ of Khot et al. [63].

Finally, as a by-product of our results, we obtain some insight regarding the possibilities of obtaining improved results by strengthening the semidefinite program with more constraints. Traditionally, the only constraints which have been useful in the design of MAX 2-CSP algorithms are triangle inequalities of a certain form (namely, those involving the vector v_0, coding the value false). It turns out that, for very natural reasons, these are exactly the inequalities that need to be satisfied in order for the hardness result to carry through. In other words, assuming that Conjecture 6.1.3 is true, it is UG-hard to do better than what can be achieved by adding only these triangle inequalities, and thus, it is unlikely that improvements can be made by adding additional inequalities (while still using polynomial time).

6.2 Semidefinite Relaxation

Any $P : \{-1,1\}^2 \to [0,1]$ can be arithmetized as

$$P(x_1, x_2) = \hat{P}_0 + \hat{P}_1 x_1 + \hat{P}_2 x_2 + \hat{P}_3 x_1 x_2,$$

for some coefficients $\hat{P}_0$, $\hat{P}_1$, $\hat{P}_2$ and $\hat{P}_3$. Throughout the remaining part of this chapter, we fix some arbitrary objective function P and its corresponding coefficients $\hat{P}_0 \dots \hat{P}_3$. Throughout this section and the next, there will be many inner products $\langle \cdot, \cdot \rangle_{\mathbb{R}}$, and we drop the subscript $\mathbb{R}$ for ease of presentation.

Hence, the MAX CSP(P) problem can be reformulated as the following integer quadratic programming problem

$$\begin{array}{l} \text{Maximize } \mathrm{Val}(x) = \sum_{\psi=(s_1 x_i, s_2 x_j)} \mathrm{wt}(\psi) \left(\hat{P}_0 + \hat{P}_1 s_1 x_i + \hat{P}_2 s_2 x_j + \hat{P}_3 s_1 s_2 x_i x_j \right) \\ \text{Subject to } x_i \in \{-1,1\} \quad \forall i. \end{array}$$

Here the sum over $\psi = (s_1 x_i, s_2 x_j)$ is the summation over all constraints, and the signs $s_1, s_2 \in \{-1,1\}$ indicates whether the variables x_i or x_j are negated.

One approach to solving integer quadratic programming problems which has turned out to be remarkably successful over the years is to relax the original problem to a semidefinite programming problem. This approach was first used in the

seminal paper by Goemans and Williamson [42] where they gave the first approximation algorithms for Max Cut, Max 2-Sat, and Max Di-Cut with a non-trivial approximation ratio (ignoring lower order terms).

For solving integer quadratic programming over the hypercube, where each variable is restricted to ± 1, the standard approach is to first homogenize the program by introducing a variable x_0 which is supposed to represent the value false and then replace each term x_i by $x_0 x_i$. We then relax each variable $x_i \in \{-1, 1\} = S^0$ to a vector $v_i \in S^n$ (i.e. a unit vector in $\mathbb{R}^{n+1}$), so that each term $x_i x_j$ becomes the scalar product $\langle v_i, v_j \rangle$.

In addition, we add the following inequality constraints to the program for all triples of vectors v_i, v_j, v_k.

$$\langle v_i, v_j \rangle + \langle v_j, v_k \rangle + \langle v_i, v_k \rangle \geq -1 \qquad -\langle v_i, v_j \rangle + \langle v_j, v_k \rangle - \langle v_i, v_k \rangle \geq -1 \tag{6.1}$$

$$\langle v_i, v_j \rangle - \langle v_j, v_k \rangle - \langle v_i, v_k \rangle \geq -1 \qquad -\langle v_i, v_j \rangle - \langle v_j, v_k \rangle + \langle v_i, v_k \rangle \geq -1 \tag{6.2}$$

These are equivalent to triangle inequalities of the form $||v_i - v_j||^2 + ||v_j - v_k||^2 \geq ||v_i - v_k||^2$, which clearly hold for the case that all vectors lie in a one-dimensional subspace of $\mathbb{R}^n$ (so this is still a relaxation of the original integer program), but is not necessarily true otherwise. There are of course many other valid inequalities which could also be added, considering k-tuples of variables rather than just triples. In particular, adding *all* valid constraints makes the optimum for the semidefinite program equal the discrete optimum (but there are an exponential number of constraints to consider).

The process of adding new constraints to LP or SDP relaxations of an integer programming problem is systematized by so-called hierarchies. The three most well-known such hierarchies are the Lovász-Schrijver hierarchy [73], the Sherali-Adams hierarchy [98], and the Lasserre hierarchy [70]. In general, these share the following features: the first level of the hierarchy is the "basic" SDP relaxation, and the rth level of the hierarchy is constructed from the $(r-1)$th by adding new constraints which have to be satisfied by any integral solution, in a certain systematic way. The SDP at the rth level of the hierarchy can be solved in time $n^{\mathcal{O}(r)}$, and any feasible solution at the nth level of the hierarchy is a convex combination of integral solutions.

While initially seeming like a powerful method for obtaining better approximation algorithms, results which make use of the higher levels of these hierarchies have been scarce, whereas there have been several results exhibiting cases where they *do not* help, e.g. [96, 41, 95]. In fact, the only result we are aware of which goes beyond the third level of any hierarchy is a very recent result by Chlamtac and Singh [24] for finding independent sets in hypergraphs, using the Lasserre hierarchy.

In particular, the only inequalities which have been used when analyzing the performance of approximation algorithms for 2-CSP problems are those of the triangle inequalities which involve the vector v_0. The results of this chapter shed some light on why this is the case—these are exactly the inequalities we need in order for the hardness of approximation to work out. Thus, assuming Conjecture 6.1.3 and

the Unique Games Conjecture, it is unlikely that adding other valid inequalities (while still being able to solve the SDP in polynomial time) will help achieve a better approximation ratio, as that would imply $\mathsf{P} = \mathsf{NP}$. This is supported by the subsequent work of Raghavendra [87]. See Section 6.8 for details.

In general, we cannot find the exact optimum of a semidefinite program. It is however possible to find the optimum to within an additive error of ϵ in time polynomial in $\log 1/\epsilon$ [1]. As is standard (see e.g. [42, 83]), we ignore this small point for notational convenience and assume that we can solve the semidefinite program exactly.

Given a vector solution $\{v_i\}_{i=0}^n$, the relaxed value of a constraint $\psi \in \Psi$ depends only on the three (possibly negated) scalar products $\langle v_0, v_i\rangle$, $\langle v_0, v_j\rangle$, and $v_i \cdot v_j$, where x_i and x_j are the two variables occurring in ψ. Most of the time, we do not care about the actual vectors, but are only interested in these triples of scalar products.

Definition 6.2.1. A *scalar product configuration* θ, or just a *configuration* for short, is a triple of real numbers (ξ_1, ξ_2, ρ) satisfying

$$\begin{array}{llll} \xi_1 + \xi_2 + \rho & \geq -1 & -\xi_1 + \xi_2 - \rho & \geq -1 \\ \xi_1 - \xi_2 - \rho & \geq -1 & -\xi_1 - \xi_2 + \rho & \geq -1. \end{array} \tag{6.3}$$

A *family of configurations* Θ is a probability space (X, η), where $X = \{\theta_1, \ldots, \theta_k\}$ is a set of configurations and η is a probability distribution over X. We routinely abuse notation by identifying Θ both with the set X and the probability space (X, η).

A configuration can be viewed as representing three vectors v_0, v_1, v_2, where $\langle v_0, v_i\rangle = \xi_i$, and $\langle v_1, v_2\rangle = \rho$. Note that the inequalities in Equation (6.3) then correspond exactly to those of the triangle inequalities (6.1) which involve v_0. Jumping ahead of ourselves, the important feature of these inequalities is that they precisely guarantee that Table 2.2 gives a valid probability distribution, which will be exactly what is needed in the hardness result in Section 6.4. It can also be shown that these inequalities ensure the existence of vectors v_0, v_1, v_2 with the corresponding scalar products.

Definition 6.2.2. The relaxed value of a configuration $\theta = (\xi_1, \xi_2, \rho)$ is given by

$$P_{\text{relax}}(\theta) = P_{\text{relax}}(\xi_1, \xi_2, \rho) = \hat{P}_0 + \hat{P}_1\xi_1 + \hat{P}_2\xi_2 + \hat{P}_3\rho.$$

We denote by

$$v|_\psi = (s_1 \langle v_0, v_i\rangle, s_2 \langle v_0, v_j\rangle, s_1 s_2 \langle v_i, v_j\rangle)$$

the configuration arising from the clause $\psi = (s_1 x_i, s_2 x_j)$ for the vector solution $v = \{v_i\}_{i=0}^n$. The relaxed value of the clause ψ is then simply given by $P_{\text{relax}}(v|_\psi)$.

Often we view a solution $\{v_i\}_{i=0}^n$ to the SDP as just the family of configurations $\Theta = \{ v|_\psi : \psi \in \Psi \}$ with the probability distribution where $\Pr_{\theta\in\Theta}[\theta = v|_\psi] = \mathrm{wt}(\psi)$. The relaxed value of an assignment of vectors $\{v_i\}_{i=0}^n$ is then given by

$$\mathrm{SDPVal}_\Psi(\{v_i\}) = \sum_{\psi\in\Psi} \mathrm{wt}(\psi) P_{\mathrm{relax}}(v|_\psi) = \mathop{\mathbb{E}}_{\theta\in\Theta}[P_{\mathrm{relax}}(\theta)].$$

Given a vector solution $\{v_i\}$, one natural attempt at an approximation algorithm is to set x_i to be true with probability $\frac{1-\xi_i}{2}$ (where $\xi_i = \langle v_0, v_i\rangle$), independently—the intuition being that the linear term ξ_i gives an indication of "how true" x_i should be. This assignment has the same expected value on the linear terms as the vector solution, and the expected value of a quadratic term $x_i x_j$ is $\xi_i\xi_j$. However, typically there is some correlation between the vectors v_i and v_j, so that the scalar product $\langle v_i, v_j\rangle$ contributes more than $\xi_i\xi_j$ to the objective function. To quantify this, write the vector v_i as

$$v_i = \xi_i v_0 + \sqrt{1-\xi_i^2}\tilde{v}_i,$$

where $\xi_i = \langle v_0, v_i\rangle$, and $\tilde{v}_i$ is the part of v_i orthogonal to v_0, normalized to a unit vector (if $\xi_i = \pm 1$, we define $\tilde{v}_i$ to be a unit vector orthogonal to all other vectors v_j). Then, we can rewrite the quadratic term $\langle v_i, v_j\rangle$ as

$$\langle v_i, v_j\rangle = \xi_i\xi_j + \sqrt{1-\xi_i^2}\sqrt{1-\xi_j^2}\,\langle \tilde{v}_i, \tilde{v}_j\rangle\,.$$

As it turns out, the relevant parameter when analyzing the quadratic terms is the scalar product $\langle \tilde{v}_i, \tilde{v}_j\rangle$, i.e. the difference between the value corresponding to $x_i x_j$ in the SDP compared to the expected value of $x_i x_j$ in the independent rounding (scaled by an appropriate factor). Motivated by this, we make the following definition.

Definition 6.2.3. The *inner angle* $\tilde{\rho}(\theta)$ of a configuration $\theta = (\xi_1, \xi_2, \rho)$ is

$$\tilde{\rho}(\theta) = \frac{\rho - \xi_1\xi_2}{\sqrt{1-\xi_1^2}\sqrt{1-\xi_2^2}}.$$

In the case that $\xi_1 = \pm 1$ or $\xi_2 = \pm 1$, we define $\tilde{\rho}(\theta) = 0$.

Note that, in the notation above, the advantage is exactly the scalar product $\langle \tilde{v}_i, \tilde{v}_j\rangle$. We are now ready to define the "positivity condition", alluded to in Section 6.1.

Definition 6.2.4. A configuration $\theta = (\xi_1, \xi_2, \rho)$ is *positive* if $\hat{P}_3 \cdot \tilde{\rho}(\theta) \geq 0$.

Intuitively, positive configurations should be more difficult to handle, since they are the configurations where we need to do something better than just setting the variables independently in order to get a good approximation ratio.

What Goemans and Williamson [42] do to round the vectors back to boolean variables is to pick a random hyperplane through the origin, and decide the value of the variables based on whether their vectors are on the same side of the hyperplane as v_0 or not. Feige and Goemans [33] suggested several generalizations of this approach, using preprocessing (e.g. first rotating the vectors) and/or more elaborate choices of hyperplanes. In particular, consider a rounding scheme where we pick a random vector $r \in \mathbb{R}^{n+1}$ and then set the variable x_i to true if

$$\langle r, \tilde{v}_i \rangle \leq T(\langle v_0, v_i \rangle) \tag{6.4}$$

for some threshold function $T : [-1, 1] \to \mathbb{R}$. This particular scheme (and more general ones) was first analyzed by Lewin et al. [71]. A very similar family of schemes, called RPR^2 roundings (short for Random Projection Randomized Rounding), was earlier analyzed by Feige and Langberg [35]. In an RPR^2 rounding, a variable x_i is set to true with probability $f(\langle r, v_i \rangle)$ for some function $f : \mathbb{R} \to [0, 1]$. In [83], RPR^2 roundings were shown to give optimal results for Max Cut (see Section 6.8 for further details). The crucial difference between RPR^2 and the rounding in Equation (6.4) is that Equation (6.4) gives the direction v_0 a special treatment, quite different from how the other directions are handled, which in turn means that the linear terms $\langle v_0, v_i \rangle$ are handled very differently from the quadratic terms $\langle v_i, v_j \rangle$. In Max Cut, this is not relevant, as there are no linear terms that need to be handled, but for a general 2-CSP, the scheme in Equation (6.4) appears more useful than RPR^2. On the flip side, the special treatment of v_0 makes it more cumbersome to recover the Goemans-Williamson Max Cut algorithm (the rounding of which can be viewed as a special case of RPR^2)—see Section 6.5.1 for details.

To describe the performance ratio yielded by this scheme, we begin by setting up some notation.

Definition 6.2.5. A *rounding* is a continuous function $R : [-1, 1] \to [-1, 1]$ which is odd, i.e. satisfies $R(\xi) = -R(-\xi)$. We denote by $\mathcal{R}$ the set of all such functions.

A rounding R is in one-to-one correspondence with a threshold function T as described above by the simple relation $R(x) = 1 - 2\Phi(T(x))$, where Φ is the normal distribution function (it will turn out to be more convenient to describe the rounding in terms of R rather than in terms of T). The reason that we require a rounding function to be odd is that a negated literal $-x_i$ should be treated the opposite way as x_i.

Definition 6.2.6. The *rounded value* of a configuration θ with respect to a rounding function $R \in \mathcal{R}$ is

$$P_{\text{round}}(\theta, R) = P_{\text{relax}}\left(R(\xi_1), R(\xi_2), 4\Gamma_{\tilde{\rho}(\theta)}(R(\xi_1), R(\xi_2)) + R(\xi_1) + R(\xi_2) - 1\right),$$

This seemingly arbitrary definition is motivated by the following lemma (which essentially traces back to Lewin et al. [71], though they never made it explicit).

Lemma 6.2.7. *There is a polynomial-time algorithm which, given a* MAX CSP(P) *instance Ψ, a semidefinite solution $\{v_i\}_{i=0}^n$ to Ψ, and a (polynomial-time computable) rounding function $R \in \mathcal{R}$, finds an assignment to Ψ with expected value*

$$\mathop{\mathbb{E}}_{\theta \in \Theta} \left[P_{\text{round}}(\theta, R) \right],$$

where Θ is the family of configurations corresponding to $\{v_i\}$.

Proof. The algorithm works as described above: First, we pick a random normal vector $r \in \mathbb{R}^{n+1}$ (i.e. each coordinate of r is a standard normal random variable). Then, we set the variable x_i to true if

$$\langle r, \tilde{v}_i \rangle \leq T(\langle v_0, v_i \rangle),$$

where we define the threshold function T as

$$T(x) = \Phi^{-1}\left(\frac{1 - R(x)}{2}\right).$$

To analyze the performance of this algorithm, we need to analyze the expected values $\mathbb{E}[x_i]$ and $\mathbb{E}[x_i x_j]$ (where the expectation is over the choice of random vector r). Note that, by Fact 2.2.10, $\langle \tilde{v}_i, r \rangle$ and $\langle \tilde{v}_j, r \rangle$ are jointly normal variables with variance 1 and covariance $\langle \tilde{v}_i, \tilde{v}_j \rangle$.

This means that x_i is set to true with probability $\frac{1-R(\xi_i)}{2}$. Thus, we have that the expected value $\mathbb{E}[x_i] = R(\xi_i)$.

For the quadratic terms, we analyze the probability that two variables x_i and x_j are rounded to the same value. The probability that both $\langle r, \tilde{v}_i \rangle \leq T(\langle v_0, v_i \rangle)$ and $\langle r, \tilde{v}_j \rangle \leq T(\langle v_0, v_j \rangle)$ is given by $\Gamma_{\tilde{\rho}}(R(\xi_i), R(\xi_j))$ where $\tilde{\rho} = \tilde{v}_i \tilde{v}_j$. By symmetry, the probability that both x_i and x_j are set to false is $\Gamma_{\tilde{\rho}}(-R(\xi_i), -R(\xi_j))$. Using Proposition 3.6.1, the expected value of $x_i x_j$ is then given by

$$\begin{aligned} \mathbb{E}[x_i x_j] &= 2\left(\Gamma_{\tilde{\rho}}\left(R(\xi_i), R(\xi_j)\right) + \Gamma_{\tilde{\rho}}\left(-R(\xi_i), -R(\xi_j)\right)\right) - 1 \\ &= 4\Gamma_{\tilde{\rho}}\left(R(\xi_i), R(\xi_j)\right) + R(\xi_i) + R(\xi_j) - 1, \end{aligned}$$

Thus, the expected value of the solution found (over the random choice of r) is given by

$$\begin{aligned} \mathop{\mathbb{E}}_{(\xi_1, \xi_2, \rho) \in \Theta} \left[\hat{P}_0 + \hat{P}_1 R(\xi_1) + \hat{P}_2 R(\xi_2) + \hat{P}_3 (4\Gamma_{\tilde{\rho}}(R(\xi_1), R(\xi_2)) + R(\xi_1) + R(\xi_2) - 1) \right] \\ = \mathop{\mathbb{E}}_{\theta \in \Theta} \left[P_{\text{round}}(\theta, R) \right], \end{aligned}$$

and we are done. □

We remark that the rounding procedure used in the proof of Lemma 6.2.7 is from the class of roundings Lewin et al. [71] called $\mathcal{THRESH}^-$. The rounding function R specifies an arbitrary rounding procedure from $\mathcal{THRESH}^-$.[1]

[1] In the notation of [71], we have $S(x) = T(x)\sqrt{1-x^2}$, or equivalently, $R(x) = 1 - 2\Phi(S(x)/\sqrt{1-x^2})$.

A statement similar to Lemma 6.2.7 holds for MAX CSP$^+(P)$, the difference being that, since there are no longer any negated literals, we can change the definition of a rounding function slightly and not require it to be odd (which could potentially give us a better algorithm). Motivated by Lemma 6.2.7, we make the following sequence of definitions

Definition 6.2.8. The *approximation ratio of a rounding* R for a family of configurations Θ is given by

$$\alpha_P(\Theta, R) = \frac{\mathbb{E}_{\theta\in\Theta}\left[P_{\text{round}}(\theta, R)\right]}{\mathbb{E}_{\theta\in\Theta}\left[P_{\text{relax}}(\theta)\right]}.$$

If $\mathbb{E}[P_{\text{relax}}(\theta)] = 0$, we let $\alpha_P(\Theta, R) = \infty$.

Definition 6.2.9. The *approximation ratio of a family of configurations* Θ is given by

$$\alpha_P(\Theta) = \max_{R\in\mathcal{R}} \alpha_P(\Theta, R).$$

It is not too hard to check that the max is attained by some R, so that the use of max instead of sup is valid. For a fixed Θ, $\alpha_P(\Theta, R)$ depends only on the value of $R(\xi)$ for at most $d = 2|\Theta|$ different ξ and we can view $\sup_R \alpha_P(\Theta, R)$ as being a supremum over a subset of $\mathbb{R}^d$ which is easily verified to be compact and convex. Furthermore, one can check that $\alpha_P(\Theta, R)$ is continuous in R and hence the supremum is attained.

Definition 6.2.10. Recall the definition of positive configurations, Definition 6.2.4. The *approximation ratios* of P for *families of k configurations* and *families of k positive configurations*, respectively, are given by

$$\alpha_P(k) = \min_{|\Theta|=k} \alpha_P(\Theta), \qquad \beta_P(k) = \min_{\substack{|\Theta|=k \\ \text{every } \theta\in\Theta \text{ is positive}}} \alpha_P(\Theta). \tag{6.5}$$

As in Definition 6.2.9, it can be seen that the min is attained so that the use of min instead of inf is valid: the set of all families of k configurations can be viewed as a compact convex subset of $[-1, 1]^{4k}$.

We would like to point out that we do not require that the family of configurations Θ can be derived from an SDP solution to some MAX CSP(P) instance Ψ—we only require that each configuration in Θ satisfies the inequalities in Equation (6.3). In other words, we have a lot more freedom when searching for a Θ which makes $\alpha_P(k)$ or $\beta_P(k)$ small, than we would have when searching for MAX CSP(P) instances and corresponding vector solutions. This is in fact the main strength of our result compared to the in almost all other respects superior subsequent result of Raghavendra [87]. We elaborate on this in Section 6.8.

Finally, we define

Definition 6.2.11. The *α and β ratios* of P are

$$\alpha(P) = \lim_{k\to\infty} \alpha_P(k), \qquad \beta(P) = \lim_{k\to\infty} \beta_P(k). \tag{6.6}$$

It is not hard to see that the limits are indeed well-defined, since $\alpha_P(k)$ and $\beta_P(k)$ for increasing k form decreasing sequences in $[0,1]$. The inequality $\alpha_P(k+1) \leq \alpha_P(k)$ holds since any family on k configurations can be viewed as a family on $k+1$ configurations in which we add an additional configuration which is given probability 0, and similarly for $\beta_P(k)$.

These are the approximation ratios arising in Theorems 6.1.1 and 6.1.2. Ideally, of course, we would like to prove hardness of approximating Max CSP(P) within $\alpha(P)$ rather than $\beta(P)$, getting rid of the requirement that every $\theta \in \Theta$ must be positive. The reason that we need it shows up when we do the proof of soundness for the PCP constructed in Section 6.4, and we have not been able to get around this. However, as we state in Conjecture 6.1.3, we do not *believe* that this restriction affects the approximation ratio achieved: by the intuition above, positive configurations seem to be the ones that are hard to round, so restricting our attention to such configurations ought not be a problem. And indeed, the configurations we use to obtain our results for Max 2-Sat and Max 2-And are all positive, as are all configurations which have appeared in previous proofs of hardness for 2-CSPs (e.g. for Max Cut and the balanced versions of Max 2-Sat and Max 2-And).

6.3 A Generic Algorithm

The approximation algorithm for Max CSP(P) (Theorem 6.1.1) is based on the following theorem.

Theorem 6.3.1. *For any $\epsilon > 0$, the value of a* Max CSP(P) *instance on k clauses can be approximated within $\alpha_P(k) - \epsilon$ in time polynomial in k.*

Note that this theorem immediately implies Theorem 6.1.1 since $\alpha_P(k) \geq \alpha(P)$. We remark that the exact value of $\alpha_P(k)$ is virtually impossible to compute for large k, making it somewhat hard to compare Theorem 6.3.1 with existing results. However, for Max Cut, Max 2-Sat and Max 2-And, it is not hard to prove that $\alpha(P)$ is at least the performance ratio of existing algorithms. See Section 6.5 for details.

Proof. Let Ψ be a Max CSP(P) instance and $\{v_i\}_{i=0}^n$ be an optimal solution to the semidefinite relaxation of Ψ. Note that, if we could find an optimal rounding function R for Ψ, the theorem would follow immediately from Lemma 6.2.7 (and we wouldn't need the ϵ). However, since we can not in general hope to find an optimal R, we'll discretize the set of possible angles and find the best rounding for the modified problem (for which there will be only a constant number of possible solutions).

We will use the simple facts that we always have

$$\mathrm{Opt}(\Psi) \geq \hat{P}_0 \geq \max(|\hat{P}_1|, |\hat{P}_2|, |\hat{P}_3|)$$

(to see that the second inequality holds, note that otherwise there would be x_1, x_2 such that $P(x_1, x_2) < 0$).

Construct a new SDP solution $\{u_i\}_{i=0}^n$ by letting $u_0 = v_0$, and, for each $1 \leq i \leq n$, letting u_i be the vector v_i rotated towards or away from v_0 so that $\langle u_0, u_i \rangle$ is an integer multiple of ϵ' (where ϵ' will be chosen small enough). In particular, we have $|\langle u_0, u_i \rangle - \langle v_0, v_i \rangle| \leq \epsilon'/2$. For the quadratic terms, Feige and Goemans [33] proved that for $i, j \geq 1$, we have

$$\langle u_i, u_j \rangle = \zeta_i \zeta_j + \tilde{\rho}_{ij} \cdot \sqrt{1-\zeta_i^2}\sqrt{1-\zeta_j^2},$$

where we define $\zeta_i := u_0 \cdot u_i$ and $\tilde{\rho}_{ij} := \frac{\langle v_i, v_j \rangle - \xi_i \xi_j}{\sqrt{1-\xi_i^2}\sqrt{1-\xi_j^2}}$. In other words, the rotation does not affect the value of $\tilde{\rho}_{ij}$. Thus, we have

$$\langle v_i, v_j \rangle - \langle u_i, u_j \rangle = \xi_i \xi_j - \zeta_i \zeta_j + \tilde{\rho}_{ij} \left(\sqrt{1-\xi_i^2}\sqrt{1-\xi_j^2} - \sqrt{1-\zeta_i^2}\sqrt{1-\zeta_j^2} \right).$$

Let us then estimate this difference. First, we have

$$|\xi_i \xi_j - \zeta_i \zeta_j| = |(\xi_i - \zeta_i)\xi_j + \zeta_i(\xi_j - \zeta_j)| \leq |(\xi_i - \zeta_i)\xi_j| + |\zeta_i(\xi_j - \zeta_j)| \leq \epsilon'. \quad (6.7)$$

For the $\sqrt{\cdot}$ terms, note that for every $\delta \in [0,1]$, the difference $\sqrt{1-x+\delta} - \sqrt{1-x}$ (for $x \in [\delta, 1]$) is maximized by $x = 1$ and hence bounded by $\sqrt{\delta}$. Thus,

$$\left| \sqrt{1-\xi_i^2} - \sqrt{1-\zeta_i^2} \right| \leq \sqrt{|\xi_i^2 - \zeta_i^2|} \leq \sqrt{\epsilon'}$$

and hence by the same argument as in Equation (6.7), we have

$$\left| \tilde{\rho}_{ij} \left(\sqrt{1-\xi_i^2}\sqrt{1-\xi_j^2} - \sqrt{1-\zeta_i^2}\sqrt{1-\zeta_j^2} \right) \right| \leq 2|\tilde{\rho}_{ij}|\sqrt{\epsilon'} \leq 2\sqrt{\epsilon'}.$$

Thus, we get that

$$|\langle v_i, v_j \rangle - \langle u_i, u_j \rangle| \leq \epsilon' + 2\sqrt{\epsilon'}.$$

However, the vectors $\{u_i\}_{i=0}^n$ could possibly violate some of the triangle inequalities. To remedy this, we adjust it slightly, by again defining a new SDP solution $\{v_i'\}_{i=0}^n$ as follows (ϵ'' will be chosen momentarily)

$$v_i' = \sqrt{1-\epsilon''}u_i + \sqrt{\epsilon''}w_i,$$

for $i \in \{0, \ldots, n\}$. Here, each w_i is a unit vector which is orthogonal to every other w_j, and to all the u_i vectors (such a set of w_i vectors is trivial to construct by embedding all vectors in $\mathbb{R}^{2(n+1)}$). These new vectors satisfy $\langle v_i', v_j' \rangle = (1 - \epsilon'')\langle u_i, u_j \rangle$ for all $i \neq j$. And since the original SDP solution $\{v_i\}_{i=0}^n$ satisfies the triangle inequalities, we have that

$$\langle u_i, u_j \rangle + \langle u_j, u_k \rangle + \langle u_k, u_i \rangle \quad \geq \quad -1 - 3\epsilon' - 6\sqrt{\epsilon'}$$

and hence

$$\left\langle v_i', v_j' \right\rangle + \left\langle v_j', v_k' \right\rangle + \left\langle v_k', v_i' \right\rangle \quad \geq \quad -(1 + 3\epsilon' + 6\sqrt{\epsilon'})(1 - \epsilon'').$$

Letting $\epsilon'' = 3\epsilon' + 6\sqrt{\epsilon'}$, the right hand side is at least -1, and this triangle inequality is satisfied. The other three sign combinations are handled identically. In other words, $\{v_i'\}_{i=0}^n$ is a feasible SDP solution. Its value can be lower-bounded by

$$\begin{aligned} &\mathrm{SDPVal}(\{v_i\}) - \mathrm{SDPVal}(\{v_i'\}) \\ &\qquad \leq \quad |\hat{P}_1|(\epsilon'/2 + \epsilon'') + |\hat{P}_2|(\epsilon'/2 + \epsilon'') + |\hat{P}_3|(\epsilon' + 2\sqrt{\epsilon'} + \epsilon'') \\ &\qquad \leq \quad |\hat{P}_0|(11\epsilon' + 20\sqrt{\epsilon''}). \end{aligned}$$

Choosing ϵ' small enough (e.g. $\epsilon' = (\epsilon/62)^2$), this is bounded by $\frac{\epsilon}{2}\operatorname{Opt}(\Psi)$.

Now, consider an optimal rounding function R for $\{v_i'\}$, and construct a new rounding function R' by letting $R'(\xi)$ be the nearest integer multiple of $\epsilon/8$ to $R(\xi)$ (so that $|R(\xi) - R'(\xi)| \leq \epsilon/16$ for all ξ). We then have for any configuration $\theta' = (\xi_1', \xi_2', \rho')$

$$\begin{aligned} P_{\mathrm{round}}(\theta', R) - P_{\mathrm{round}}(\theta', R') \quad &\leq \quad |\hat{P}_1|\epsilon/16 + |\hat{P}_2|\epsilon/16 + |\hat{P}_3|(4\epsilon/16 + \epsilon/16 + \epsilon/16) \\ &\leq \quad \frac{\epsilon}{2}\operatorname{Opt}(\Psi). \end{aligned}$$

To see this, we refer to Corollary 3.6.4, which implies that

$$|\Gamma_{\tilde{\rho}}(R(\xi_1'), R(\xi_2')) - \Gamma_{\tilde{\rho}}(R'(\xi_1'), R'(\xi_2'))| \leq \epsilon/16.$$

Note that we only need to define R' for values of ξ which are integer multiples of ϵ'. Since, for each of the $\approx 2/\epsilon'$ such values of ξ, there are only $\approx 16/\epsilon$ possible values for $R'(\xi)$, the number of possible R' is constant, $(1/\epsilon)^{\Theta(1/\epsilon')}$. Thus, we can find a rounding which is at least as good as R' in polynomial time by simply trying all possible choices of R', evaluating each one, and picking the best function found. Using Lemma 6.2.7, this means that we can find a solution to Ψ with expected value at least

$$\begin{aligned} \mathop{\mathbb{E}}_{\theta' \in \Theta'}[P_{\mathrm{round}}(\theta', R')] \quad &\geq \quad \mathop{\mathbb{E}}_{\theta' \in \Theta'}[P_{\mathrm{round}}(\theta', R)] - \frac{\epsilon}{2}\operatorname{Opt}(\Psi) \\ &= \quad \alpha_P(\Theta')\mathrm{SDPVal}(\{v_i'\}) - \frac{\epsilon}{2}\operatorname{Opt}(\Psi) \\ &\geq \quad \alpha_P(\Theta')\mathrm{SDPVal}(\{v_i\}) - \epsilon\operatorname{Opt}(\Psi) \\ &\geq \quad (\alpha_P(k) - \epsilon)\operatorname{Opt}(\Psi), \end{aligned}$$

where Θ' denotes the set of configurations arising from the SDP solution $\{v_i'\}_{i=0}^n$. □

6.4 A Generic Hardness Result

Theorem 6.1.2 immediately follows from the following Theorem 6.4.1 below. Taking k large enough so that $\beta_P(k) \leq \beta(P)+\epsilon$ and invoking Theorem 6.4.1 gives hardness of approximating MAX CSP(P) within $\beta(P) + 2\epsilon$.

Theorem 6.4.1. *It is UG-hard to approximate* MAX CSP(P) *within* $\beta_P(k)+\epsilon$ *for any* $\epsilon > 0$ *and* $k \in \mathbb{N}$.

As in previous chapters, we prove Theorem 6.4.1 by constructing a PCP verifier which checks a supposed dictatorship encoding of a good assignment to a UNIQUE LABEL COVER instance, and decides whether to accept or reject based on the evaluation of the objective function P. The verifier is parameterized by a family of k positive configurations $\Theta = \{\theta_1, \ldots, \theta_k\}$ and a probability distribution on Θ. Again, we point out that the requirement that the configurations of Θ are positive is by necessity rather than by choice, and if we could get rid of it, the hardness of approximation yielded would exactly match the approximation ratio from Theorem 6.1.1. The set Θ corresponds to a set of vector configurations for the semidefinite relaxation of MAX CSP(P). When proving soundness, i.e., in the case that there is no good assignment to the UNIQUE LABEL COVER instance, we prove that the best strategy for the prover corresponds to choosing a good rounding function R for the family of configurations Θ. Choosing a set of configurations which are hard to round, we obtain the desired result.

Since we can negate variables freely, we will use folding and assume that the purported dictatorships are balanced, as described in Section 4.4. This is what is going to ensure that the prover's rounding function is odd, i.e. that $R(\xi) = -R(-\xi)$. The verifier is given in Algorithm 2, below. Note that, because θ is a configuration, Equation (6.3) guarantees that we can choose x_1 and x_2 with the desired distribution in step (4).

Algorithm 2: The verifier $\mathcal{V}$
$\mathcal{V}(\Psi, \Sigma = \{f_v\}_{v \in Y})$
(1) Pick a random configuration $\theta = (\xi_1, \xi_2, \rho) \in \Theta$ according to the distribution on Θ.
(2) Pick a random $v \in X$.
(3) Pick $e_1 = \{v, w_1\}$ and $e_2 = \{v, w_2\}$ randomly from $E(v)$.
(4) Pick $x_1, x_2 \in \{-1, 1\}^L$ such that each bit of x_i is picked independently with expected value ξ_i and that the j:th bits of x_1 and x_2 are ρ-correlated for $j = 1, \ldots, L$.
(5) For $i = 1, 2$, let $b_i = f_{w_i}(x_i \circ \pi_{e_i}^{-1})$ (folded over true).
(6) Accept with probability $P(b_1, b_2)$.

Let us now study the completeness of $\mathcal{V}$.

Lemma 6.4.2 (Completeness)**.** *If* $\mathrm{Opt}(\Psi) \geq 1 - \eta$*, then there is a proof* Σ *such that*

$$\Pr[\mathcal{V}(\Psi, \Sigma) \text{ accepts}] \geq (1 - 2\eta) \mathop{\mathbb{E}}_{\theta \in \Theta}[P_{\mathrm{relax}}(\theta)].$$

Proof. Fix a labelling ℓ of the vertices of Ψ such that the fraction of satisfied edges is at least $1 - \eta$, and let $f_v : \{-1, 1\}^L \to \{-1, 1\}$ be $f_v = \textsc{Dict}_{\ell(v)}$. Note that for a satisfied edge $\{v, w\}$ and an arbitrary string $x \in \{-1, 1\}^L$, $f_w(x \circ \pi_e^{-1})$ equals the value of the $\ell(v)$:th bit of x.

Fix a choice of $\theta = (\xi_1, \xi_2, \rho)$. By the union bound, the probability that either of the two edges e_1, e_2 chosen by $\mathcal{V}$ are not satisfied is at most 2η. For a choice of edges that *are* satisfied, the expected value of $f_{w_i}(x_i \circ \pi_{e_i}^{-1})$ is the expected value of the $\ell(v)$:th bit of x_i, i.e. ξ_i, and the expected value of $f_{w_1}(x_1 \circ \pi_{e_1}^{-1}) f_{w_2}(x_2 \circ \pi_{e_2}^{-1})$ is the expected value of the $\ell(v)$:th bit of $x_1 x_2$, i.e. ρ.

Thus, the probability that $\mathcal{V}$ accepts is at least

$$\mathop{\mathbb{E}}_{\theta \in \Theta}\left[(1 - 2\eta)(\hat{P}_0 + \hat{P}_1 \xi_1 + \hat{P}_2 \xi_2 + \hat{P}_3 \rho)\right] = (1 - 2\eta) \mathop{\mathbb{E}}_{\theta \in \Theta}[P_{\mathrm{relax}}(\theta)],$$

and the proof is complete. □

Next, we turn to the soundness of $\mathcal{V}$.

Lemma 6.4.3 (Soundness)**.** *For every* $\epsilon > 0$ *there is a* $\gamma > 0$ *such that if* $\mathrm{Val}(X) \leq \gamma$*, then for any proof* Σ*, we have*

$$\Pr[\mathcal{V}(X, \Sigma) \text{ accepts}] \leq \max_{R \in \mathcal{R}} \mathop{\mathbb{E}}_{\theta \in \Theta}[P_{\mathrm{round}}(\theta, R)] + \epsilon.$$

Proof. Arithmetizing the acceptance predicate, we find that the acceptance probability of $\mathcal{V}$ can be written as

$$\mathop{\mathbb{E}}_{\theta \in \Theta}\Big[\mathop{\mathbb{E}}_{v, e_1, e_2, x_1, x_2}\big[\hat{P}_0 + \hat{P}_1 f_{w_1}(x_1 \circ \pi_{e_1}^{-1}) + \hat{P}_2 f_{w_2}(x_2 \circ \pi_{e_2}^{-1}) + \\ + \hat{P}_3 f_{w_1}(x_1 \circ \pi_{e_1}^{-1}) f_{w_2}(x_2 \circ \pi_{e_1}^{-1}) \mid \theta\big]\Big].$$

For $\xi \in [-1, 1]$ and $v \in V$, define $g_v^\xi \in L^2(\{-1, 1\}^n, \mu_\xi^{\otimes n})$ by

$$g_v^\xi(x) = \mathop{\mathbb{E}}_{e = \{v, w\} \in E(v)}\left[f_w(x \circ \pi_e^{-1})\right],$$

for $x \in \{-1, 1\}^n$, where μ_ξ is the distribution on $\{-1, 1\}$ which assigns probability $\frac{1+\xi}{2}$ to 1, and probability $\frac{1-\xi}{2}$ to -1. Define the function $R_v(\xi) := \mathbb{E}\left[g_v^\xi\right]$. Note that since the purported dictatorships are balanced, we have that both g_v^ξ and R_v are odd functions, and in particular that $R_v \in \mathcal{R}$. We remark that for a fixed v and different values of ξ, the functions g_v^ξ are the same function, but as random variables they are different.

We can now write $\mathcal{V}$'s acceptance probability as

$$\begin{aligned}\Pr[\mathcal{V} \text{ accepts}] &= \mathop{\mathbb{E}}_{\theta}\left[\mathop{\mathbb{E}}_{v,x_1,x_2}\left[\hat{P}_0 + \hat{P}_1 g_v^{\xi_1}(x_1) + \hat{P}_2 g_v^{\xi_2}(x_2) + \hat{P}_3 g_v^{\xi_1}(x_1) g_v^{\xi_2}(x_2) \mid \theta\right]\right] \\ &= \mathop{\mathbb{E}}_{\theta,v}\left[\hat{P}_0 + \hat{P}_1 R_v(\xi_1) + \hat{P}_2 R_v(\xi_2) + \hat{P}_3 \left\langle g_v^{\xi_1}, g_v^{\xi_2}\right\rangle_{\mathcal{N}}\right]. \qquad (6.8)\end{aligned}$$

Assume (for contradiction) that

$$\begin{aligned}\Pr[\mathcal{V} \text{ accepts}] &\geq \mathop{\mathbb{E}}_{\theta,v}\left[P_{\text{round}}(\theta, R_v)\right] + \epsilon \\ &= \mathop{\mathbb{E}}_{\theta,v}\Big[\hat{P}_0 + \hat{P}_1 R_v(\xi_1) + \hat{P}_2 R_v(\xi_2) + \\ &\qquad + \hat{P}_3(4\Gamma_{\tilde{\rho}}(R_v(\xi_1), R_v(\xi_2)) + R_v(\xi_1) + R_v(\xi_2) - 1)\Big] + \epsilon. \qquad (6.9)\end{aligned}$$

Combining this with Equation (6.8), this implies that there exists a $\theta = (\xi_1, \xi_2, \rho) \in \Theta$ such that

$$\mathop{\mathbb{E}}_{v}\left[\hat{P}_3 \cdot \left(\left\langle g_v^{\xi_1}, g_v^{\xi_2}\right\rangle_{\mathcal{N}} - 4\Gamma_{\tilde{\rho}(\theta)}(R_v(\xi_1), R_v(\xi_2)) - R_v(\xi_1) - R_v(\xi_2) + 1\right)\right] \geq \epsilon. \quad (6.10)$$

Using the fact that the absolute value of the expression inside the expectation is bounded by $2|\hat{P}_3|$, this implies that for at least a fraction $\epsilon' := \frac{\epsilon}{3|\hat{P}_3|}$ of all $v \in X$, we have

$$\hat{P}_3 \cdot \left\langle g_v^{\xi_1}, g_v^{\xi_2}\right\rangle_{\mathcal{N}} \geq \hat{P}_3\left(4\Gamma_{\tilde{\rho}(\theta)}(R_v(\xi_1), R_v(\xi_2)) + R_v(\xi_1) + R_v(\xi_2) - 1\right) + \epsilon'.$$

Let V be the set of all such v. Using that θ is a positive configuration (i.e. $\hat{P}_3\tilde{\rho}(\theta) \geq 0$), we then get that for $v \in V$,

$$\left\langle g_v^{\xi_1}, g_v^{\xi_2}\right\rangle_{\mathcal{N}} \geq 4\Gamma_{|\tilde{\rho}(\theta)|}(R_v(\xi_1), R_v(\xi_2)) + R_v(\xi_1) + R_v(\xi_2) - 1 + \epsilon'/|\hat{P}_3|$$

if $\hat{P}_3 > 0$, or

$$\left\langle g_v^{\xi_1}, g_v^{\xi_2}\right\rangle_{\mathcal{N}} \leq 4\Gamma_{-|\tilde{\rho}(\theta)|}(R_v(\xi_1), R_v(\xi_2)) + R_v(\xi_1) + R_v(\xi_2) - 1 - \epsilon'/|\hat{P}_3|$$

if $\hat{P}_3 < 0$ (note that if $\hat{P}_3 = 0$, we would have a contradiction in Equation (6.10) and so could jump directly to Equation (6.11) below). In either case, Theorem 2.5.2 (combined with Proposition 2.2.8) implies that there are constants τ and d (depending only on ϵ, θ, and P) such that for any $v \in V_{\text{good}}$ we have $\text{Inf}_i^{\leq d}(g_v^{\xi_1}) \geq \tau$ (and also that $\text{Inf}_i^{\leq d}(g_v^{\xi_2}) \geq \tau$, though we will not use that). Fixing θ and dropping the bias parameter ξ_1 for the remainder of the proof, we have that for any $v \in V$,

$$\tau \leq \text{Inf}_i^{\leq d}(g_v) \leq \mathop{\mathbb{E}}_{e=\{v,w\}}\left[\text{Inf}_{\pi_e(i)}^{\leq d}(f_w)\right],$$

where the second inequality is the same argument as Equation (5.2). The rest of the proof now follows the proof of Lemma 5.1.3 from Equation (5.2) and onwards: we can construct small sets of candidate labels $C(v)$ for every $v \in X$ and $w \in Y$ based on the influential coordinates of g_v and f_w, and use these to define a random labeling ℓ, which in expectation satisfies a fraction $\frac{\epsilon'\tau^3}{4d^2}$ of all edges. Making sure that $\gamma < \frac{\epsilon'\tau^3}{4d^2}$, we get a contradiction to the assumption of the acceptance probability (Equation (6.9)), implying that the soundness is at most

$$\begin{aligned} \Pr[\mathcal{V} \text{ accepts } \Sigma] &\leq \mathop{\mathbb{E}}_{\theta,v}[P_{\text{round}}(\theta, R_v)] + \epsilon \qquad (6.11) \\ &\leq \max_{R \in \mathcal{R}} \mathop{\mathbb{E}}_{\theta \in \Theta}[P_{\text{round}}(\theta, R)] + \epsilon, \end{aligned}$$

and we are done. □

Combining Lemma 6.4.2 and Lemma 6.4.3, and picking γ small enough, we get that, for every $\epsilon > 0$, Max CSP(P) is $(s + \epsilon, c - \epsilon)$-UG-hard, where

$$\begin{aligned} s &= \max_{R \in \mathcal{R}} \mathop{\mathbb{E}}_{\theta \in \Theta}[P_{\text{round}}(\theta, R)] \\ c &= \mathop{\mathbb{E}}_{\theta \in \Theta}[P_{\text{relax}}(\theta)] \end{aligned}$$

Picking a Θ with $|\Theta| = k$ that minimizes $\alpha_P(\Theta)$, we obtain Theorem 6.4.1.

6.5 Results for Specific Predicates

In the remaining part of this chapter, we study the implications of the generic results Theorem 6.1.1 and Theorem 6.1.2 on specific 2-CSP problems. In particular, we look at Max Cut, Max 2-Sat, and Max 2-And. On the algorithmic side, we show that the approximation guarantee of Theorem 6.1.1 matches the previous best algorithms for these problems. On the hardness side, we show that for Max Cut we recover the result of Khot et al. [63] that it is UG-hard to approximate Max Cut better than $\alpha_G W \approx 0.87856$. For both Max 2-Sat and Max 2-And, we improve upon existing hardness results. For Max 2-Sat, we obtain an inapproximability of $\alpha_{LLZ} \approx 0.94017$, exactly matching the believed approximation ratio of the algorithm by Lewin et. al [71]. For Max 2-And, we obtain an inapproximability of ≈ 0.87435, which almost matches the ratio of 0.87401 of the algorithm by Lewin et. al [71].

In general, given a family Θ, the very problem of computing $\alpha_P(\Theta)$ is a difficult numeric optimization problem. However, for the Θ we use, the number of distinct ξ-values used is small, so that computing $\alpha_P(\Theta)$ in this case is a numeric optimization problem in only 1 or 2 variables, which we are able to handle.

6.5.1 Max Cut

As a warm-up, let us show how our general results can be used to derive previous bounds for the Max Cut problem.

Rather than MAX CUT, we will start with the slightly more general case of MAX 2-XOR (of course, an algorithm for MAX 2-XOR is also an algorithm for MAX CUT). The classic Goemans-Williamson rounding algorithm works as follows: pick a standard normal random vector r, and then set $x_i = -1$ true if $r \cdot v_i \leq 0$. In our framework, using the representation $v_i = \xi_i v_0 + \sqrt{1-\xi_i^2}\tilde{v}_i$, this rounding can equivalently be formulated as follows: pick a standard normal variable $r_0 \in \mathbb{R}$, and a standard normal vector r, and then set $x_i = -1$ if

$$r \cdot \tilde{v}_i \leq \frac{-\xi_i r_0}{\sqrt{1-\xi_i^2}},$$

(if $|\xi_i| = 1$, the right hand side of the above expression is defined to be $+\infty$ or $-\infty$ according to the sign of $-\xi_i r_0$). Hence the Goemans-Williamson rounding algorithm can be interpreted as an algorithm in which we pick a threshold function T at random according to a certain distribution over threshold functions, and then apply threshold rounding using T. Clearly, the ratio obtained by such an approach is no better than the ratio obtained by picking the best threshold function, and hence the approximation ratio α_{GW} of the Goemans-Williamson algorithm is bounded by $\alpha(\oplus_2)$.

Let us then move to hardness, and again we will first consider the MAX 2-XOR problem. To prove hardness for MAX 2-XOR, it suffices to consider the single configuration $\theta = (0, 0, \rho)$, for some $\rho \in [-1, 1]$. A computation of $\alpha_{\oplus}(\{\theta\}, R)$ then gives

$$\alpha_{\oplus_2}(\{\theta\}, R) = \frac{2 - 2R(0) - 4\Gamma_\rho(R(0), R(0))}{1-\rho} = \frac{2 - 4\Gamma_\rho(0,0)}{1-\rho},$$

where the second equality uses that $R(0) = 0$ for any rounding R. Recall that $\Gamma_\rho(0,0)$ is the probability that two jointly gaussian random variables X and Y with covariance ρ are both smaller than $\Phi^{-1}(1/2) = 0$, which equals (see e.g. [79], Theorem B.1)

$$\Gamma_\rho(0,0) = \frac{1}{2} - \frac{1}{2\pi}\arccos\rho.$$

Hence

$$\alpha_{\oplus_2}(\{\theta\}, R) = \frac{2\arccos\rho}{\pi(1-\rho)}.$$

The minimum value of this expression over $\rho \in [-1, 1]$ is, by definition, exactly α_{GW}.

Now, MAX CUT, as opposed to MAX 2-XOR, is a MAX CSP$^+$ problem rather than a MAX CSP problem, so it does not quite make sense to talk about hardness for MAX CUT being a special case of Theorem 6.1.2. However, while we have only mentioned MAX CSP$^+$ problems in passing, analogues of Theorem 6.1.1 and Theorem 6.1.2 are true for MAX CSP$^+$ problems. The crucial difference is that one can no longer assume that a rounding is odd, an in particular one can not assume that $R(0) = 0$, which we used for the MAX 2-XOR hardness. This means that in

this case, the hardness we get for the configuration above is given by

$$\max_{R(0)\in[-1,1]} \frac{2 - 2R(0) - 4\Gamma_\rho(R(0), R(0))}{1-\rho}. \tag{6.12}$$

Fortunately, it is easy to prove that the expression $x+2\Gamma_\rho(x,x)$ is indeed minimized by $x = 0$ (see Corollary 3.6.3 for the derivative of Γ_ρ). Hence, Equation (6.12) is at most $\frac{2\arccos\rho}{\pi(1-\rho)}$ and by choosing an appropriate ρ we again get a hardness of α_{GW}.

6.6 Max 2-Sat

In this section, we look at Max 2-Sat. Throughout this section, let $P : \{-1,1\}^2 \to \{0,1\}$ be the Or predicate, i.e., the predicate which is 1 if at least one of its inputs equals -1.

The best existing algorithm for Max 2-Sat is the LLZ algorithm by Lewin, Livnat and Zwick [71]. This algorithm works by using a certain well-chosen fixed rounding. In their original paper, Lewin et al. used a somewhat complicated rounding function of the form

$$R_0(x) = 2\Phi(2\cot(f(\arccos x))) - 1$$

where $f(x) \approx 0.588314583\theta + 0.64667394$. However, as communicated to us by Zwick [108], a slightly better, and in particular much simpler choice of rounding function is

$$R(x) = \gamma x,$$

where $\gamma = \alpha_{LLZ} \approx 0.94016567$ is the approximation ratio of the algorithm. We will return to the difference between these two functions later in this section, but first we will analyze the performance ratio of the algorithm.

Already at this point, it is clear that this algorithm can be no better than the one in Theorem 6.1.1, as the latter always uses the best possible R which in particular is at least as good as the fixed R used by the LLZ algorithm.

6.6.1 The Definition of α_{LLZ}

In order to prove hardness for Max 2-Sat which exactly matches the approximation ratio of the algorithm, we will need to analyze the algorithm carefully. The approximation ratio of this algorithm is at least the worst approximation ratio of R on any configuration θ, i.e., at least

$$\min_\theta \alpha_P(\{\theta\}, R).$$

Let us then compute the rounded and relaxed values of a configuration θ. Arithmetizing P, we have

$$P(x_1, x_2) = \frac{3 - x_1 - x_2 - x_1x_2}{4}.$$

We then get for a configuration $\theta = (\xi_1, \xi_2, \rho)$,

$$\begin{aligned} P_{\text{round}}(\theta, R) &= \frac{3 - R(\xi_1) - R(\xi_2) - \left(4\Gamma_{\tilde{\rho}(\theta)}(R(\xi_1), R(\xi_2)) + R(\xi_1) + R(\xi_2) - 1\right)}{4} \\ &= \frac{4 - 2\gamma\xi_1 - 2\gamma\xi_2 - 4\Gamma_{\tilde{\rho}(\theta)}(\gamma\xi_1, \gamma\xi_2)}{4} \\ P_{\text{relax}}(\theta) &= \frac{3 - \xi_1 - \xi_2 - \rho}{4}. \end{aligned} \tag{6.13}$$

As R is completely determined by the scaling factor $\gamma \in [0, 1]$, we will from here on write $\alpha_P(\{\theta\}, \gamma)$ rather than $\alpha_P(\{\theta\}, R)$. We then get

$$\alpha_P(\{\theta\}, \gamma) = \frac{4 - 2\gamma\xi_1 - 2\gamma\xi_2 - 4\Gamma_{\tilde{\rho}(\theta)}(\gamma\xi_1, \gamma\xi_2)}{3 - \xi_1 - \xi_2 - \rho}.$$

There is no known way to analytically compute the minimum of this expression over all configurations θ. Hence, one has to resort to numerical computations. Extensive numerical computations, both our own and those of Lewin et al., show that $\alpha_P(\{\theta\}, \gamma)$ is minimized at two different configurations θ_1, θ_2, where

$$\theta_1 = (\xi, \xi, -1 + 2\xi) \qquad \theta_2 = (-\xi, -\xi, -1 + 2\xi), \tag{6.14}$$

for some ξ.[2] For now, let us ignore the numeric value of ξ and instead consider the implications of the worst configurations having this very specific form. We refer to a configuration of the form $(\xi, \xi, -1+2|\xi|)$ as a *simple configuration* ξ. Let us denote by $\alpha(\xi, \gamma)$ the approximation ratio of the LLZ algorithm on a simple configuration ξ, i.e.,

$$\alpha(\xi, \gamma) = \alpha_P(\{(\xi, \xi, -1 + 2|\xi|)\}, \gamma) = \frac{2 - 2\gamma\xi - 2\Gamma_{\tilde{\rho}(\xi)}(\gamma\xi, \gamma\xi)}{2 - \xi - |\xi|},$$

where we define

$$\tilde{\rho}(\xi) = \tilde{\rho}(\xi, \xi, -1 + 2|\xi|) = \frac{-1 + 2|\xi| - \xi^2}{1 - \xi^2} = \frac{|\xi| - 1}{|\xi| + 1}.$$

Finally, we define α_{LLZ} as the minimum value of $\alpha(\xi, \gamma)$ over all ξ, assuming that γ is chosen so as to maximize this quantity. I.e.,

$$\alpha_{LLZ} = \max_{\gamma \in [0,1]} \min_{\xi \in [-1,1]} \alpha(\xi, \gamma). \tag{6.15}$$

This is the quantity which we refer to as the approximation ratio of the LLZ algorithm. Its value is approximately 0.94017 (we discuss this further in Section 6.6.3). Note however that there is no formal proof that this is indeed the approximation

[2] Of course, numeric computations can not formally prove that the worst configurations have this specific form, only that they approximately have this form, up to some tiny error.

ratio of the algorithm. In particular, the only evidence that the worst case configurations for the algorithm are indeed simple configurations is by numerical computations. If this part could be proved formally, then one would have a formal proof that α_{LLZ} is indeed a lower bound on the approximation ratio of the algorithm, but as of now, there is only numerical evidence (albeit strong numerical evidence).

In the following sections, we will continue to analyze this quantity, in order to prove that $\beta(P) \leq \alpha_{LLZ}$ (which we can indeed prove formally without having to resort to numerical computations!). The first step towards proving this is to prove that $\gamma = \alpha_{LLZ}$ is indeed a maximizer of Equation (6.15) (i.e., that the optimal choice of γ is $\gamma = \alpha_{LLZ}$), and that if ξ is such that $\alpha(\xi, \alpha_{LLZ})$ is minimum, then so is $-\xi$ (i.e., if ξ is a worst-case configuration, then so is $-\xi$). Using this we will then be able to construct a probability distribution on the two worst-case configurations in Equation (6.14) such that the resulting family of configurations Θ satisfies $\alpha_P(\Theta) = \alpha_{LLZ}$, implying that $\beta(P) \leq \alpha_{LLZ}$.

Following this, we will briefly discuss the actual numeric value of α_{LLZ}, and finally, as promised in the beginning of this section, we will discuss the difference between the rounding function used here compared to the one originally used by Lewin et al.

6.6.2 Analyzing α_{LLZ}

Throughout this section, let $\Gamma_{\tilde{\rho}}(x) := \Gamma_{\tilde{\rho}}(x, x)$, and let $\Gamma'_{\tilde{\rho}}(x)$ be the derivative of $\Gamma_{\tilde{\rho}}(x)$ with respect to x.

First, we show that:

Proposition 6.6.1. *The function $\alpha(\xi, \gamma)$ satisfies the following two properties:*

1. $\min_{\xi \in [-1,1]} \alpha(\xi, \alpha_{LLZ}) = \alpha_{LLZ}$. *In other words, Equation (6.15) is maximized by setting $\gamma = \alpha_{LLZ}$.*
2. *If ξ satisfies $\alpha(\xi, \alpha_{LLZ}) = \alpha_{LLZ}$ then so does $-\xi$. In other words, if ξ is a worst-case configuration for $\gamma = \alpha_{LLZ}$, then so is $-\xi$.*

Proof. Define

$$\begin{aligned}
\text{gain}(\xi, \gamma) &= P_{\text{round}}(\xi, \xi, -1 + 2|\xi|), \gamma) - \alpha_{LLZ} P_{\text{relax}}(\xi, \xi, -1 + 2|\xi|) \\
&= (1 - \gamma\xi - \Gamma_{\tilde{\rho}}(\gamma\xi)) - \alpha_{LLZ} \cdot \left(1 - \frac{\xi}{2} - \frac{|\xi|}{2}\right) \qquad (6.16) \\
&= \frac{\alpha_{LLZ} - \gamma}{2}\xi + \frac{2 - \Gamma_{\tilde{\rho}}(\gamma\xi) - \Gamma_{\tilde{\rho}}(-\gamma\xi)}{2} - \alpha_{LLZ}\frac{2 - |\xi|}{2} \qquad (6.17)
\end{aligned}$$

to be the "advantage" over α_{LLZ} when rounding the configuration $(\xi, \xi, -1 + 2|\xi|)$ using a particular value of γ (where we used Proposition 3.6.1 to get Equation (6.17)). The first part of the proposition amounts to showing that

$$\text{gain}(\xi, \alpha_{LLZ}) \geq 0$$

for all $\xi \in [-1,1]$. Let γ^* be such that $\min_{\xi\in[-1,1]} \alpha(\xi,\gamma^*) = \alpha_{LLZ}$, i.e. an optimal choice of γ. By definition, we have that $\text{gain}(\xi,\gamma^*) \geq 0$ for all simple configurations ξ.

The intuition behind the remainder of the proof is as follows: it is not hard to see that $\text{gain}(\xi,\alpha_{LLZ}) = \text{gain}(-\xi,\alpha_{LLZ})$. Furthermore, we will see that the sign of the derivative of $\text{gain}(\xi,\gamma)$ with respect to γ depends only on the sign of ξ. Hence, to prove that $\text{gain}(\xi,\alpha_{LLZ}) \geq 0$, we make sure that the derivative of gain is negative from α_{LLZ} to γ^*, by potentially switching the sign of ξ. This then implies that $\text{gain}(\xi,\alpha_{LLZ}) \geq \text{gain}(\pm\xi,\gamma^*) \geq 0$.

Formally, let

$$\text{gain}_\gamma(\xi,\gamma) = \frac{\partial \text{gain}}{\partial\gamma}(\xi,\gamma) = -\xi\left(1+\Gamma'_{\tilde{\rho}}(\gamma\xi)\right)$$

be the derivative of gain (in the form of Equation (6.16)) with respect to γ. Note that by Corollary 3.6.3 we have $1+\Gamma'_{\tilde{\rho}}(\gamma\xi) \in [0,1]$. In particular, the sign of the derivative depends only on the sign of ξ.

Consider an arbitrary configuration ξ. Define a new configuration ξ' by

$$\xi' = \begin{cases} \xi & \text{if } (\gamma^* - \alpha_{LLZ})\xi \geq 0 \\ -\xi & \text{otherwise.} \end{cases}$$

It might help to think of ξ' the following way: if ξ has the right sign for the function $\text{gain}(\xi,\cdot)$ to be increasing from γ^* to α_{LLZ}, we are happy, but otherwise we flip ξ, thereby also flipping the sign of gain_γ.

By the Mean Value Theorem, there is then a γ' between γ^* and α_{LLZ} such that

$$\begin{aligned} \text{gain}(\xi',\alpha_{LLZ}) &= \text{gain}(\xi',\gamma^*) + (\alpha_{LLZ}-\gamma^*)\text{gain}_\gamma(\xi',\gamma') \\ &\geq (\gamma^*-\alpha_{LLZ})\xi'\left(1+\Gamma'_{\tilde{\rho}}(\gamma'\xi')\right) \geq 0. \end{aligned}$$

Now, from Equation (6.17) we see that $\text{gain}(-\xi,\gamma) - \text{gain}(\xi,\gamma) = (\gamma-\alpha_{LLZ})\xi$ for every γ and ξ. In particular, since $\xi' = \pm\xi$, this implies that $\text{gain}(\xi,\alpha_{LLZ}) = \text{gain}(\xi',\alpha_{LLZ}) \geq 0$, which proves the first part of the proposition.

The second part of the proposition follows from $\text{gain}(\xi,\alpha_{LLZ}) = \text{gain}(-\xi,\alpha_{LLZ})$. In particular, if ξ is a worst case configuration, they both equal 0. □

Analyzing this a bit further will (unsurprisingly) show that $\gamma = \alpha_{LLZ}$ is indeed the *only* maximum of the Equation (6.15), though we will not need this fact. In order to show this, it suffices to realize that $\xi = 0$ can never be a worst-case configuration.

We are now finally ready to prove the matching hardness $\beta(P) \leq \alpha_{LLZ}$ claimed in Theorem 6.1.4.

Theorem 6.6.2. *For $P(x_1,x_2) = \frac{3-x_1-x_2-x_1x_2}{4}$, we have*

$$\beta(P) \leq \alpha_{LLZ}.$$

Proof. Let $\xi, \Delta \in [-1,1]$ be parameters to be determined later. Consider the family of configurations $\Theta = \{\theta_1, \theta_2\}$, where

$$\theta_1 = (\xi, \xi, -1+2\xi) \qquad \theta_2 = (-\xi, -\xi, -1+2\xi), \tag{6.18}$$

the probability of θ_1 is $(1+\Delta)/2$, and the probability of θ_2 is $(1-\Delta)/2$. Then both θ_1 and θ_2 are positive configurations (since $\tilde{\rho}(\theta_1) = \tilde{\rho}(\theta_2) = \frac{|\xi|-1}{|\xi|+1} < 0$ has the same sign as $\hat{P}_3$). Hence $\beta(P) \leq \alpha_P(\Theta) = \max_{R \in \mathcal{R}} \alpha_P(\Theta, R)$, so it remains to bound this quantity. By definition,

$$\alpha_P(\Theta, R) = \frac{\mathbb{E}_{\theta \in \Theta}[P_{\text{round}}(\theta, R)]}{\mathbb{E}_{\theta \in \Theta}[P_{\text{relax}}(\theta, R)]}.$$

Similarly to the proof of Proposition 6.6.1, let

$$\begin{aligned} \text{gain}(R) &= \mathop{\mathbb{E}}_{\theta \in \Theta}[P_{\text{round}}(\theta, R)] - \alpha_{LLZ} \mathop{\mathbb{E}}_{\theta \in \Theta}[P_{\text{relax}}(\theta)] \\ &= \frac{2 - (1+\Delta)R(\xi) - 2\Gamma_{\tilde{\rho}}(R(\xi))}{2} - \alpha_{LLZ} \cdot \frac{2 - \Delta\xi - |\xi|}{2} \end{aligned}$$

be the advantage over α_{LLZ} when rounding Θ using a particular rounding R (where the second equality used Equation (6.13)).

We need to prove that, with an appropriate choice of ξ and Δ, we have $\text{gain}(R) \leq 0$ for every rounding R. Note that $\text{gain}(R)$ depends only on the rounded value $r = R(\xi)$ of ξ, hence we think of gain as a function $\text{gain}(r)$ from $[-1,1]$ to $\mathbb{R}$.

Now, let $\gamma^* = \alpha_{LLZ}$, and let ξ be a worst simple configuration, i.e. such that $\alpha(\xi, \gamma^*) = \alpha_{LLZ}$. By Proposition 6.6.1, we also have $\alpha(-\xi, \gamma^*) = \alpha_{LLZ}$. This implies that $\text{gain}(\gamma^*\xi) = 0$. Computing the derivative of $\text{gain}(r)$, we have

$$\text{gain}'(r) = -\frac{1+\Delta}{2} - \Gamma'_{\tilde{\rho}}(r).$$

Solving $\text{gain}'(\gamma^*\xi) = 0$ for Δ gives

$$\Delta = -2\Gamma'_{\tilde{\rho}}(\gamma^*\xi) - 1 = 2\Phi\left(\sqrt{\frac{1-\tilde{\rho}}{1+\tilde{\rho}}} t(\gamma^*\xi)\right) = 2\Phi\left(\Phi^{-1}\left(\frac{1-\gamma^*\xi}{2}\right) / \sqrt{|\xi|}\right) - 1.$$

It is clear that $\Delta \in [-1,1]$. Furthermore, $\text{gain}''(r) = -\Gamma''_{\tilde{\rho}}(r) < 0$, implying that gain is concave, so that when Δ is chosen as above, gain attains a global maximum at $r = \gamma^*\xi$, where it takes the value 0. Hence $\text{gain}(r) \leq 0$ for every $r \in [-1,1]$. □

6.6.3 The Numeric Value of α_{LLZ}

We will now (briefly) discuss the actual numeric value of α_{LLZ}. Let $B = 0.9401656724$. To give a feel for $\alpha(\xi, B)$, Figure 6.1 gives a plot of this function in the interval

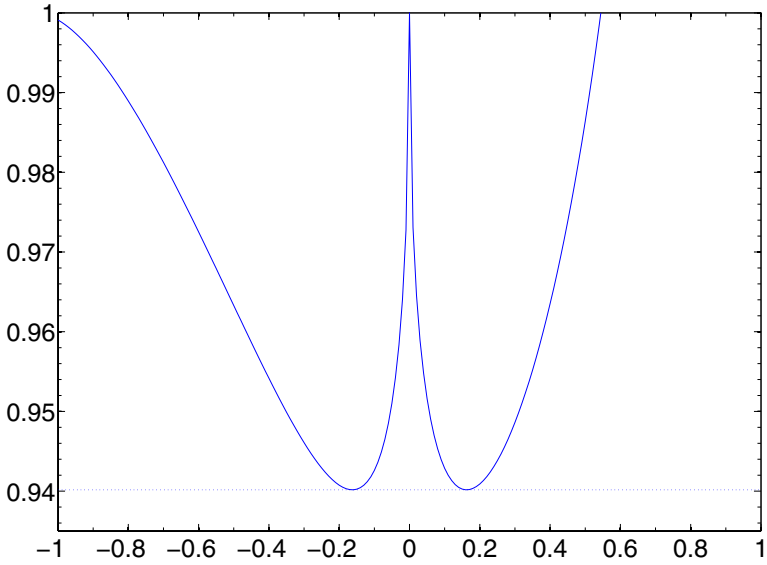


Figure 6.1: $\alpha(\xi, 0.94016567248)$

$\xi \in [-1, 1]$, along with the line $y = B$ (dashed). The one-dimensional optimization problem

$$\min_{\xi \in [-1,1]} \alpha(\xi, B)$$

can be solved numerically to a high level of precision. This gives a lower bound $\alpha_{LLZ} \geq 0.9401656724$. The two minima seen in Figure 6.1 turn out to be roughly $\xi_1 = -0.1624783294$ and $\xi_2 = 0.1624783251$ (had we plugged in $B = \alpha_{LLZ}$, Proposition 6.6.1 implies that we would have had $\xi_1 = \xi_2$, but since B differs slightly from α_{LLZ}, we get slightly different values). In order to obtain an upper bound on α_{LLZ}, we can then solve the one-dimensional optimization problem

$$\max_{\gamma \in [-1,1]} \min\left(\alpha(\xi_1, \gamma), \alpha(\xi_2, \gamma)\right)$$

numerically to a high level of precision. This results in an upper bound of $\alpha_{LLZ} \leq 0.9401656725$. In conclusion, we have $|\alpha_{LLZ} - 0.94016567245| \leq 5 \cdot 10^{-11}$.

Our worst configurations $\xi \approx \pm 0.1625$ differ slightly from the worst configurations $\xi \approx \pm 0.169$ found by Lewin et al. This is because of the small difference in behavior of the two rounding functions (see Section 6.6.4); the approximation ratio is marginally worse when using the original function of [71] rather than the one used in this paper [108].

We can also compute the amount of imbalance in the worst-case instances, by using the formula for Δ given in the end of the proof of Theorem 6.6.2. Plugging $\xi = 0.1625$ into this expression gives $\Delta = 0.3673$. This implies that in the instances produced by applying Theorem 6.1.2 to the configurations constructed in Theorem 6.6.2, the total weight on positive (resp. negative) occurrences of a variable is roughly 0.68 (resp. 0.32). We find it remarkable that so greatly imbalanced instances should be the hardest to approximate.

6.6.4 The Tale of the Two Rounding Functions

The rounding function of the LLZ algorithm used in this thesis, communicated to us by Zwick [108], differs from the rounding function used by Lewin et al. [71]. The rounding function used in this thesis is $R(x) = \gamma \cdot x$, where $\gamma = \alpha_{LLZ} \approx 0.94016567$. The rounding function used in [71] is $R_0(x) = 1 - 2\Phi(S(x)/\sqrt{1-x^2})$. Here, $S(x) = -2\cot(f(\arccos x))\sqrt{1-x^2}$ where f is the linear rotation function given by

$$f(\theta) \approx 0.588314580\theta + 0.64667394.$$

$R_0(x)$ can be simplified to

$$R_0(x) = 1 - 2\Phi(-2\cot(f(\arccos x))) = 2\Phi(2\cot(f(\arccos x))) - 1.$$

Figure 6.2 gives plots of the functions $R(x)$ and $R_0(x)$ for the interval $x \in [0, 1]$ (since both functions are odd we restrict our attention to positive x). As can be seen, the functions are fairly close to each other. Most importantly, the functions behave almost the same in the critical interval $x \in [0.1, 0.2]$. Nevertheless, there is a small difference between the functions in this interval as well, and as noted in Section 6.6.3, this causes the worst configuration when using $R_0(x)$ to be slightly different from the worst configuration when using $R(x)$. This small difference in fact causes the (apparent) approximation ratio when using $R(x)$ to be marginally better than when using $R_0(x)$.

For large x, the functions $R(x)$ and $R_0(x)$ differ noticeably, but using *the* best rounding does not matter there; these are configurations that are in some sense easy to round, and any function with a reasonable behavior suffices to get a decently good approximation ratio.

6.7 Max 2-And

In this section, we obtain an upper bound of $\beta(P) \leq 0.87435$ for the case when $P(x_1, x_2) = x_1 \wedge x_2$, i.e., the Max 2-And problem, establishing Theorem 6.1.5. We do this by exhibiting a set Θ of $k = 4$ (positive) configurations on 2 distinct non-zero ξ-values (and a probability distribution on the elements of Θ), such that $\alpha_P(\Theta) < 0.87435$.

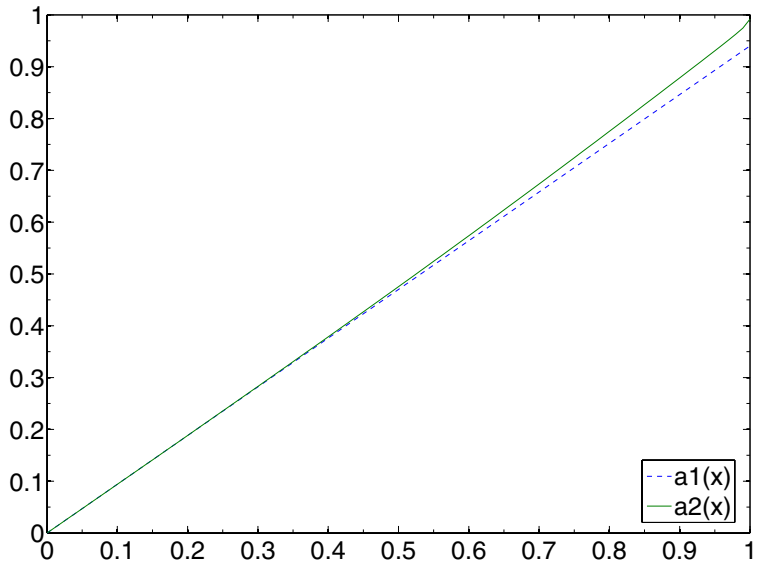


Figure 6.2: $R_0(x)$ vs. $R(x)$

6.7.1 An Easier Bound

Before giving our strongest bound, let us start with an even smaller set of configurations, sufficient to give an inapproximability of 0.87451, only marginally worse than 0.87435. In particular, this bound is strong enough to demonstrate that Max 2-And is harder to approximate than Max Cut and balanced Max 2-And. This set of configurations $\Theta = \{\theta_1, \theta_2\}$ contains only one non-zero ξ-value, and is given by

$$\begin{array}{rcll} \theta_1 & = & (0, -\xi, 1-\xi) & \text{with probability } 0.64612 \\ \theta_2 & = & (0, \xi, 1-\xi) & \text{with probability } 0.35388, \end{array}$$

where $\xi = 0.33633$.

To compute the hardness factor given by this set of configurations, we must compute

$$\alpha_P(\Theta) = \max_{R \in \mathcal{R}} \frac{\mathbb{E}_{\theta \in \Theta}[P_{\text{round}}(\theta, R)]}{\mathbb{E}_{\theta \in \Theta}[P_{\text{relax}}(\theta)]}. \tag{6.19}$$

Since $P(x_1, x_2) = \frac{1 - x_1 - x_2 + x_1 x_2}{4}$ we have that for an arbitrary configuration $\theta =$

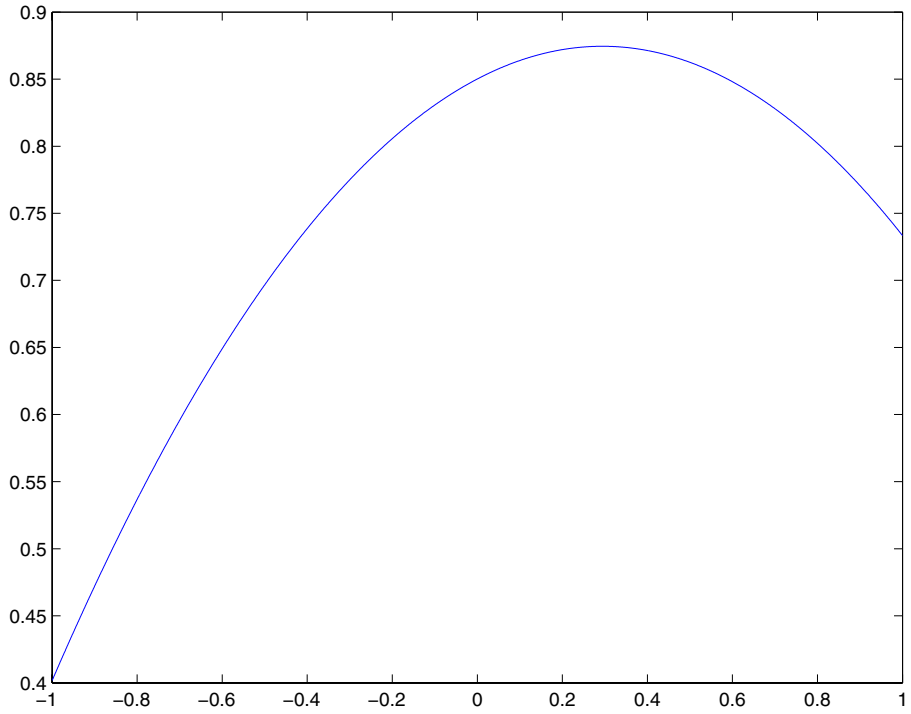


Figure 6.3: Approximation ratio as a function of R

(ξ_1, ξ_2, ρ),

$$\begin{aligned} P_{\text{relax}}(\theta) &= \frac{1 - \xi_1 - \xi_2 + \rho}{4} \\ P_{\text{round}}(\theta, R) &= \frac{1 - R(\xi_1) - R(\xi_2) + 4\Gamma_{\tilde{\rho}(\theta)}(R(\xi_1), R(\xi_2)) + R(\xi_1) + R(\xi_2) - 1}{4} \\ &= \Gamma_{\tilde{\rho}(\theta)}(R(\xi_1), R(\xi_2)). \end{aligned}$$

In our case, using the two configurations given above, R is completely specified by its value on the angle ξ (since $R(0) = 0$ and $R(-\xi) = -R(\xi)$). Figure 6.3 gives a plot of the right-hand side of Equation (6.19), as a function of the value of $R(\xi)$. The maximum turns out to occur at $R(\xi) \approx 0.29412$, and gives a ratio of approximately 0.87450517. Thus, we see that $\alpha_P(\Theta) \leq 0.87451$. We remark that it is not very difficult to make this computation rigorous—it can be proven analytically that the curve of Figure 6.3 is indeed convex (as in the proof of Theorem 6.6.2), and so the only maximum can be computed to within high precision (using easy bounds on the derivative) using a simple golden section search.

6.7.2 The Stronger Bound

Let us now turn to the larger set of configurations, based on four configurations, mentioned earlier. This set of configurations $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ is as follows:

$$\begin{array}{rcll} \theta_1 & = & (0, -\xi_A, 1-\xi_A) & \text{with probability } 0.52850 \\ \theta_2 & = & (0, \xi_A, 1-\xi_A) & \text{with probability } 0.05928 \\ \theta_3 & = & (\xi_A, -\xi_B, 1-\xi_A-\xi_B) & \text{with probability } 0.29085 \\ \theta_4 & = & (-\xi_A, \xi_B, 1-\xi_A-\xi_B) & \text{with probability } 0.12137, \end{array}$$

where $\xi_A = 0.31988$ and $\xi_B = 0.04876$.

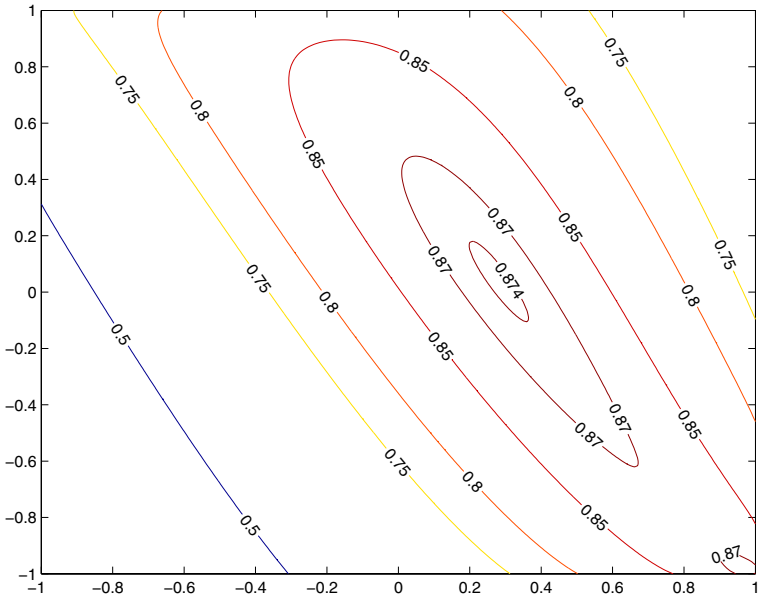
As before, to compute the approximation ratio given by Θ, we need to find the best R for Θ, and again, such an R is completely specified by its values on the non-zero ξ-values. In other words, we now need to specify the values of R on the two angles ξ_A and ξ_B. Figure 6.4(a) gives a contour plot of approximation ratio, as a function of the values of $R(\xi_A)$ and $R(\xi_B)$. There are now two local maxima, one around the point $(R(\xi_A), R(\xi_B)) \approx (0.27846, 0.044376)$, and one around the point $(1, -1)$. Figure 6.4(b) gives a contour plot of the area around the first point. This maximum turns out to be approximately 0.87434075. At the point $(1, -1)$ (which is indeed the other maximum), the approximation ratio is approximately 0.87434007. Thus, we have $\alpha_P(\Theta) \leq 0.87435$.

It seems likely that additional improvements can be made by using more and more ξ-values, though these improvements will be quite small. Indeed, using larger Θ we are able to improve upon Theorem 6.1.5, but the improvements we have been able to make are minute (of order 10^{-5}), and it becomes a lot more difficult to verify them. Note that θ_1 and θ_2 used in the larger set of configurations are very similar to the first set of configurations—they are of the same form, and the ξ-value used is only slightly different. It appears that it is useful to follow this pattern when adding even more configurations: the values of ξ_A and ξ_B are adjusted slightly, and we add two configurations of the form $(\pm\xi_B, \mp\xi_C, 1 - \xi_B - \xi_C)$. Essentially this type of sequence of configurations has appeared before, see e.g. the analysis of lower bounds for certain Max Di-Cut algorithms in [107].

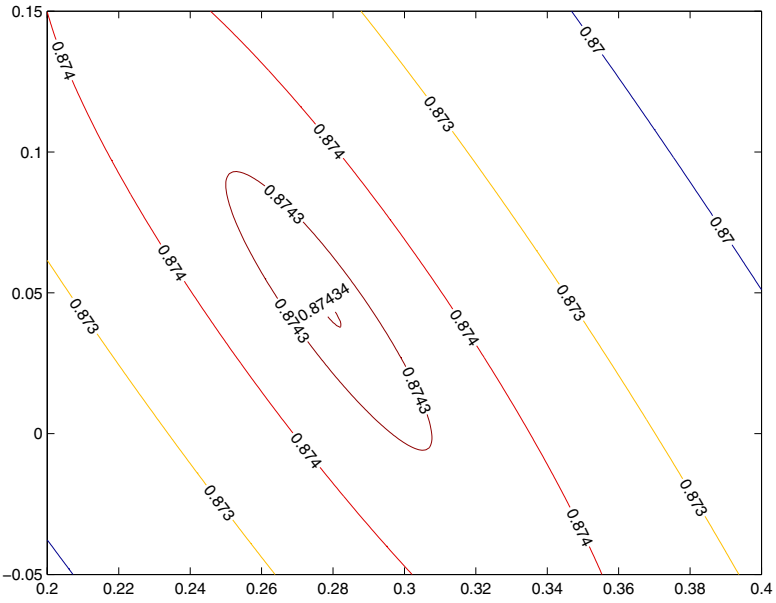
6.8 Subsequent Work

Subsequent to our result [8], there have been two very closely related results. In order to discuss them, it will be useful to define the notion of an *integrality gap*. An (s, c)-integrality gap for an SDP relaxation of a constraint satisfaction problem $\mathcal{P}$ is an instance Ψ of $\mathcal{P}$ such that $\text{Opt}(\Psi) \leq s$ and $\text{SDPOpt}(\Psi) \geq c$.

An integrality gap can be viewed as a type of unconditional hardness results, in that they state that a very specific computational model (semidefinite programming) can not solve a certain problem. In particular, they indicate that the SDP relaxation for which the integrality gap was proved can not distinguish between $\text{Opt} \leq s$ and $\text{Opt} \geq c$.



(a) The entire range of R



(b) Restricted to the critical area

Figure 6.4: Approximation ratio as a function of R

6.8.1 The Approximability Curve for Max Cut

In an "orthogonal" work to the results of this chapter, O'Donnell and Wu [83] analyzed the entire "approximability curve" of the Max Cut problem. In particular, they considered a certain function $s : [1/2, 1] \to [1/2, 1]$, and constructed:

- A polynomial-time $(s - \epsilon)$-approximation algorithm for Max Cut. To be more specific, they gave an algorithm $\mathcal{A}$ which, for every constant $\epsilon > 0$, on input a Max Cut instance Ψ, finds a cut of value at least $s(\mathrm{Opt}(\Psi)) - \epsilon$.
- A $(s(c) + \epsilon, c - \epsilon)$-UG-hardness result for Max Cut for every $\epsilon > 0$ and $c \in [1/2, 1]$.
- An $(s(c) + \epsilon, c - \epsilon)$-integrality gap for the standard SDP relaxation of Max Cut (with triangle inequalities), for every $\epsilon > 0$ and $c \in [1/2, 1]$.

The main difference which makes Max Cut more amenable to analysis than general CSPs is the absence of the linear coefficients $\hat{P}_1$ and $\hat{P}_2$. In particular, as we saw in Section 6.5.1, determining the approximability ratio for Max Cut is, because of the lack of linear terms, a fairly straight-forward task, whereas for general predicates, we can not even prove that $\alpha(P) = \beta(P)$. Determining the entire approximability curve for Max Cut is significantly more involved than just finding the worst ratio: in the language of this chapter, for certain ranges of c it does not suffice to look at a single configuration.

The rounding scheme used by the algorithm of [83] uses the RPR^2 rounding scheme of Feige and Langberg [35] mentioned in Section 6.2, in which every variable x_i is set to true with probability $f(\langle r, v_i \rangle)$ for some function f, where r is a standard normal random vector. By the same argument as in Section 6.5.1, it is not hard to see that the approximation ratio obtained by an RPR^2 rounding in which $f(x) \in \{0, 1\}$ is at most $\alpha(P)$. However, the general family of RPR^2 rounding schemes is not immediately comparable to the family of rounding schemes we use. Hence, it could be the case that our algorithm is not an $(s - \epsilon)$-approximation algorithm.

Finally, we note that it is straightforward to adapt our algorithm and hardness results to yield approximability curves rather than just ratios. Whether these curves match would depend on the truth of a conjecture analogous to Conjecture 6.1.3, that for *every* c, the quantities

$$\inf_{\mathbb{E}_{\theta \in \Theta}[P_{\mathrm{relax}}(\theta)] = c} \alpha_P(\Theta) \qquad \inf_{\substack{\mathbb{E}_{\theta \in \Theta}[P_{\mathrm{relax}}(\theta)] = c \\ \text{every } \theta \in \Theta \text{ positive}}} \alpha_P(\Theta)$$

are equal. It is quite possible that this is true, though we do not have as strong faith in it as in Conjecture 6.1.3.

6.8.2 UG-Hardness from Integrality Gaps

In a remarkable result, Raghavendra [87] essentially proved the following theorem: let $P : [q]^k \to [-1, 1]$, and suppose a certain natural SDP relaxation for

MAX CSP(P) has an (s, c)-integrality gap. Then, MAX CSP(P) is $(s + \epsilon, c - \epsilon)$-UG-hard. In other words, assuming the UGC, if semidefinite programming can not approximate MAX CSP(P) to within some factor α, then no polynomial time algorithm can.

For the special case of objective functions $P : \{-1, 1\}^2 \to [0, 1]$, i.e., the setting we have considered in this chapter, the SDP relaxation used is exactly the standard SDP relaxation used in this chapter, with those of the triangle inequalities that involve v_0. In other words, the results of [87] verify the indication given by the results of this chapter, that these inequalities are the only ones which help.

If our Conjecture 6.1.3 is true, the results of this chapter are as strong as the results of [87] for $P : \{-1, 1\}^2 \to [0, 1]$, since any integrality gap instance can be viewed as a family of configurations with a gap between P_{round} and P_{relax}.

The main advantage of our results is that [87] requires an actual integrality gap instance in order to be able to derive a hardness result. Integrality gaps which satisfy the triangle inequalities can be quite difficult to construct, and hence, for many problems we do not know the exact approximation ratio of the associated SDP relaxation. Our result, on the other hand, only needs to start with a family of configurations, which is a much simpler object to construct. One can view a family of configurations as a "recipe" for an integrality gap, in the sense that it specifies that the inner products involved should take certain values for a certain fraction of constraints. In particular, if one wants to compute explicit inapproximability ratios for different problems, it can be much easier to find an appropriate family of configurations instead of a complete integrality gap instance. For instance, we do not know of any integrality gap instances for MAX 2-AND with gap larger than α_{GW}, the MAX CUT constant. On the other hand, it is not too complicated to find a family of configurations with a larger gap, as we did in Section 6.7.

Part III

Some Limited Independence Results

They called me mad, and I called them mad,
and damn them, they outvoted me.

Nathaniel Lee

Chapter 7

Preliminaries

In this section, we give some background material necessary for the results of Chapter 8 and Chapter 9. Most of this material is about properties of the Fourier representation of functions, or rather, about properties of functions whose Fourier representation is a low-degree polynomial.

7.1 Hypercontractivity

For a random variable f and $1 \leq p \leq q \leq \infty$, we always have $||f||_p \leq ||f||_q$. In short, hypercontractivity is the phenomenon when a weak form of the converse inequality is also true. Formally, we define

Definition 7.1.1. Let (Ω, μ) be a probability space and $0 < \eta < 1$. A random variable $f : \Omega^n \to \mathbb{R}$ is said to be (p, q, η)-*hypercontractive* if for every $a \in \mathbb{R}$

$$||a + f||_p \geq ||a + \eta f||_q.$$

It is also common to define hypercontractivity in terms of the norms of f under a certain "noise operator". However, as we will only use hypercontractivity as a tool to bound large norms in terms of small norms, Definition 7.1.1 is more suitable for our purposes.

Hypercontractivity is of great importance in many areas of analysis. In computer science, there are many important applications in the analysis of boolean functions, where it was first introduced by Kahn, Kalai and Linial [60], in a famous result which states that in every balanced boolean function $f : \{-1, 1\}^n \to \{-1, 1\}$, there is a variable with influence (w.r.t. the uniform distribution on $\{-1, 1\}^n$) at least $\Omega\left(\frac{\log n}{n}\right)$.

The perhaps most famous hypercontractivity result is a theorem which in the computer science literature is sometimes called *the* Hypercontractivity Theorem, but more often referred to as the Bonami-Beckner Theorem, referring to the work of Bonami [17, 18] and later work by Beckner [11]. However, many variations of it

in different settings were discovered independently at around the same time, e.g. in the work by Nelson [80] and Gross [49].

We will use the following recent result of Wolff [105], establishing essentially optimal hypercontractivity estimates for any finite probability space.

Theorem 7.1.2 ([105])**.** *Let (Ω, μ) be a finite probability space in which the minimum non-zero probability is $\alpha(\mu) \leq 1/2$. Then for $p \geq 2$, every random variable $f \in L^2(\Omega, \mu)$ with $\mathbb{E}[f] = 0$ is $(2, p, \eta_p(\alpha))$-hypercontractive with*

$$\eta_p(\alpha) = \sqrt{\frac{A^{1/p} - A^{-1/p}}{A^{1/p'} - A^{-1/p'}}}$$

where $A = (1-\alpha)/\alpha$ and $1/p + 1/p' = 1$. The value at $\alpha = 1/2$ is taken to be the limit of the above expression as $\alpha \to 1/2$, i.e., $\eta_p(1/2) = 1/\sqrt{p-1}$.

As the $\eta_q(\alpha)$ quantity in Theorem 7.1.2 is somewhat ungainly to work with, we will instead use the following bounds which are sufficient for our purposes.

Corollary 7.1.3. *Let (Ω, μ) be a finite probability space in which the minimum non-zero probability is $\alpha \leq 1/2$. Then every random variable $f \in L^2(\Omega, \mu)$ with $\mathbb{E}[f] = 0$ is $(2, 3, (\alpha/8)^{1/6})$-hypercontractive.*

Proof. We have

$$\eta_3(\alpha)^2 = \frac{A^{1/3} - A^{-1/3}}{A^{2/3} - A^{-2/3}} = \frac{1}{A^{1/3} - A^{-1/3}} \geq \frac{1}{2A^{1/3}} \geq \frac{\alpha^{1/3}}{2},$$

which gives the desired bound. □

Corollary 7.1.4. *Let (Ω, μ) be a finite probability space in which the minimum non-zero probability is $\alpha \leq 1/2$. Then for $p \geq 2$, every random variable $f \in L^2(\Omega, \mu)$ with $\mathbb{E}[f] = 0$ is $(2, p, \tau_p(\alpha))$-hypercontractive with*

$$\tau_p(\alpha) = \sqrt{2\alpha/p}.$$

We remark that a sharper bound for $\alpha \leq 1/e$ is $\sqrt{2\alpha \ln(1/\alpha)/(p-1)}$, and an even sharper bound can be found in e.g. [26]. As the log factor will not have any impact for our applications, we sacrifice it for the sake of the very simple bound of Corollary 7.1.4.

Proof. It suffices to prove that $\eta_p(\alpha)$ in Theorem 7.1.2 is lower-bounded by $\tau_p(\alpha)$. It will be more convenient to view τ_p and η_p as functions of $A = \frac{1-\alpha}{\alpha}$ rather than of α. At $\alpha = 1/2$, the statement is clearly true, so let $\alpha < 1/2$, implying $A > 1$.

We need to show that for every $A > 1$ and $p \geq 2$,

$$\frac{A^{1/p} - A^{-1/p}}{A^{1/p'} - A^{-1/p'}} \geq \frac{2}{(1+A)p}.$$

Multiplying numerator and denominator by $A^{-1/p}$, the left hand side is clearly equal to

$$\frac{A^{1/p} - A^{-1/p}}{A^{1/p'} - A^{-1/p'}} = \frac{A^{2/p} - 1}{A - A^{-1+2/p}}.$$

We will now bound the numerator and denominator of this expression. For the numerator, we have

$$A^{2/p} - 1 = e^{\frac{2}{p}\ln A} - 1 \geq \frac{2}{p}\ln A,$$

since $e^x > 1 + x$ for every $x > 0$. For the denominator, we claim that

$$A - A^{-1+2/p} \leq (1 + A)\ln A. \tag{7.1}$$

To see this, first note that the inequality holds at $A = 1$. Furthermore, the derivatives of the left and right hand sides of the inequality with respect to A are

$$\begin{aligned} \frac{\partial}{\partial A}(A - A^{-1+2/p}) &= 1 + (1 - 2/p)A^{-2+2/p} \\ \frac{\partial}{\partial A}((1 + A)\ln A) &= 1 + A^{-1} + \ln(A). \end{aligned}$$

It is easy to see that for $A \geq 1$, $1 + A^{-1} + \ln(A) \geq 1 + (1 - 2/p)A^{-2+2/p}$ and hence Equation (7.1) holds for every $A \geq 1$.

Combining the two bounds, we obtain

$$\eta_q^2 = \frac{A^{1/p} - A^{-1/p}}{A^{1/p'} - A^{-1/p'}} \geq \frac{\frac{2}{p}\ln A}{(1 + A)\ln A} = \frac{2}{p(1 + A)} = \tau_q^2,$$

for every $A > 1$. □

It is well-known that if (Ω, μ) is hypercontractive, then low-degree polynomials over $(\Omega^n, \mu^{\otimes n})$ are hypercontractive, see e.g. [79] and [58].

Theorem 7.1.5. *If every mean-zero function $f \in L^2(\Omega, \mu)$ is $(2, p, \eta)$-hypercontractive, then every mean-zero degree-d polynomial $g \in L^2(\Omega^n, \mu^{\otimes n})$ is $(2, p, \eta^d)$-hypercontractive.*

As a simple corollary of Corollary 7.1.4 and Theorem 7.1.5, we have

Theorem 7.1.6. *Every degree-d polynomial $f \in L^2(\Omega^n, \mu^{\otimes n})$ satisfies*

$$||f||_2 \geq \left(\frac{2\alpha}{p}\right)^{d/2} ||f||_p.$$

The proof follows by writing $f = f^{=0} + f^{>0}$, noting that $||f||_2 \geq ||\eta f^{=0} + f^{>0}||_2$ (for every $0 \leq \eta \leq 1$), and then appealing to the hypercontractivity of $f^{>0}$. See also [58], Theorem 5.10.

For the application in Chapter 8, we will need that the ℓ_1 and ℓ_2 norms are comparable, rather than ℓ_2 and ℓ_p for $p > 2$. The former follows from the latter by a classic application of Hölder's inequality. Let us just state the special case when $p = 3$, as this will be sufficient for our purposes.

Theorem 7.1.7. *If f satisfies $||f||_2 \geq \eta||f||_3$, then $||f||_1 \geq \eta^3||f||_2$.*

Proof. Applying Cauchy-Schwarz to the functions $g(x) = |f(x)|^{1/2}$ and $h(x) = |f(x)|^{3/2}$, we have

$$||f||_2^2 = \langle g, h\rangle \leq ||g||_2||h||_2 = \mathbb{E}[|f(x)|]^{1/2}\,\mathbb{E}[|f(x)|^3]^{1/2} = ||f||_1^{1/2}||f||_3^{3/2}.$$

By the relation between the ℓ_2 and ℓ_3 norms of f we can bound $||f||_3^{3/2} \leq \eta^{-3/2}||f||_2^{3/2}$. Dividing both sides by $||f||_2^{3/2}$ and squaring, we conclude that

$$||f||_2 \leq \eta^{-3}||f||_1,$$

as desired. □

For future reference, we note that combining Corollary 7.1.3, Theorem 7.1.5, and Theorem 7.1.7 gives:

Theorem 7.1.8. *Let $(\Omega^n, \mu^{\otimes n})$ be a finite product space. Then any degree-d polynomial $f \in L^2(\Omega^n, \mu^{\otimes n})$ satisfies*

$$||f||_1 \geq \left(\frac{\alpha(\mu)}{8}\right)^{d/2} ||f||_2.$$

When we use Theorem 7.1.6 and Theorem 7.1.8, we sometimes refer to them as "hypercontractivity", even though they are formally not hypercontractivity estimates. For Theorem 7.1.6, the only reason that we do not state it as a hypercontractivity estimate is that this would require adding the condition $\mathbb{E}[f] = 0$ (a necessary condition for f to be hypercontractive is that $\mathbb{E}[f] = 0$). For Theorem 7.1.8, adding the condition $\mathbb{E}[f] = 0$ is not sufficient to obtain hypercontractivity (it is an easy exercise to check that if $\mathbb{E}[f] \neq 0$, then for all $0 < \eta < 1$ there is an a such that $||a + f||_1 < ||a + \eta f||_2$).

7.2 Concentration Bounds

It is known that hypercontractivity implies good concentration bounds for low-degree polynomials (see e.g. [26]).

Theorem 7.2.1. *Let $(\Omega^n, \mu^{\otimes n})$ be a finite product space. Then, for any degree-d polynomial $f \in L^2(\Omega^n, \mu^{\otimes n})$ with $||f||_2 = 1$ and any $t > e^{d/2}$,*

$$\Pr[|f| > t] \leq \exp(-ct^{2/d}),$$

where $c := \frac{d \cdot \alpha(\mu)}{e}$.

Proof. Set $p = t^{2/d} \cdot \frac{2\alpha}{e}$. By Markov's inequality, we have

$$\Pr[|f| > t] = \Pr[|f|^p > t^p] \leq \frac{||f||_p^p}{t^p}. \tag{7.2}$$

Now, since $t > e^{d/2}$, p is at least 2α. This implies that

$$||f||_p \leq \sqrt{p/(2\alpha)}^d ||f||_2 = t/e^{d/2}.$$

If $p > 2$, this follows from Theorem 7.1.6, whereas if $2\alpha \leq p \leq 2$, it follows from the monotonicity of ℓ_p norms.

Plugging this into Equation (7.2) we get

$$\Pr[|f| > t] \leq \left(\frac{t/e^{d/2}}{t}\right)^p = \exp\left(-\frac{d\alpha}{e} t^{2/d}\right). \qquad \square$$

For quadratic polynomials, hypercontractivity combined with the standard "Chernoff method" for concentration inequalities give the following strong bound (it also follows from Theorem 7.2.1). The fact that sums of independent random variables with sufficiently "nice" moments are concentrated is sometimes also called Bernstein's inequality.

Theorem 7.2.2. *Let $(\Omega^n, \mu^{\otimes n})$ be a finite product space. Then for every $\epsilon > 0$ there is a $\delta > 0$ depending only on ϵ and $\alpha(\mu)$ such that the following holds. Let $f \in L^2(\Omega^n, \mu^{\otimes n})$ be a degree-2 polynomial. Let $x_1, \ldots, x_m$ be i.i.d. samples from $(\Omega^n, \mu^{\otimes n})$. Then*

$$\Pr\left[\left|\sum_{i=1}^m |f(x_i)| - m\,\mathbb{E}[|f|]\right| > \epsilon m\right] \leq 2\exp(-\delta m)$$

Proof. Let $\lambda > 0$ be a parameter to be determined later, and let $\mu = \mathbb{E}[|f|]$ and $X = \sum_{i=1}^m |f(x_i)|$. By Markov's inequality and the independence of the x_i:s,

$$\Pr[X \geq (\mu+\epsilon)m] \leq \frac{\mathbb{E}[\exp(\lambda X)]}{\exp(\lambda(\mu+\epsilon)m)} = \left(\frac{\mathbb{E}[\exp(\lambda|f|)]}{\exp(\lambda(\mu+\epsilon))}\right)^m := \beta^m.$$

Hence it suffices to prove that for an appropriately chosen λ, we have $\beta < 1$.

Let c be such that $||f||_p \leq cp||f||_2$ for every $p \geq 2$ (by Theorem 7.1.6 we can take $c = \frac{1}{2\alpha(\mu)}$). By the Taylor expansion $\exp(x) = \sum_{k=0}^\infty x^k/k!$, we have

$$\begin{aligned}
\mathbb{E}[\exp(\lambda|f|)] &= \sum_{k=0}^\infty \frac{\mathbb{E}\left[(\lambda|f|)^k\right]}{k!} \leq 1 + \lambda\mu + \sum_{k=2}^\infty \frac{(\lambda ck)^k}{k!} \\
&\leq 1 + \lambda\mu + \sum_{k=2}^\infty \frac{(\lambda ck)^k}{(k/e)^k} = 1 + \lambda\mu + \sum_{k=2}^\infty (\lambda ce)^k \\
&= 1 + \lambda\mu + \frac{(\lambda ce)^2}{1-\lambda ce} \leq 1 + \lambda(\mu + \epsilon/2) \leq \exp(\lambda(\mu+\epsilon/2)),
\end{aligned}$$

where the last line assumes that λ is small enough so that $\lambda ce < 1$, and that $\lambda\frac{(ce)^2}{1-\lambda ce} \leq \epsilon/2$. Hence, $\beta \leq \exp(-\lambda\epsilon/2)$, which with $\delta := \lambda\epsilon/2 = \Theta(\epsilon^2/c^2)$ proves that

$$\Pr\left[X - m\mu > \epsilon m\right] \leq \exp(-\delta m).$$

The lower bound on X follows by either applying the same argument to the random variable $-X$, or by noting that since $-X$ is upper bounded by 0 one can do an even easier argument. □

7.3 Nets

For the Theorem 8.4.1 in Chapter 8, we will use the following result on the existence of small ϵ-nets of the unit sphere in $\mathbb{R}^n$.

Theorem 7.3.1. *For every n and $0 < \epsilon < 1/3$, there exists a set S of at most $(5/\epsilon)^n$ unit vectors in $\mathbb{R}^n$, such that, for any unit vector $u \in \mathbb{R}^n$, there is a $v \in S$ satisfying*

$$\langle u, v\rangle_{\mathbb{R}} \geq 1 - \epsilon.$$

The following construction is due to Kochol [68]. The proof we sketch here is from Brieden et al. [19].

Proof sketch. Define $R = \sqrt{n}/\epsilon$, and let $W = \mathbb{Z}^n \cap B(R)$, where $B(R) \subseteq \mathbb{R}^n$ is ball of radius R (in ℓ_2 norm). Let the net be $V = \{w/||w||_2 : w \in W\}$, the vectors of W normalized to unit length.

For any unit vector v, define $z = \lfloor R\cdot v\rfloor$ (i.e., the vector u in which $u_i = \lfloor R\cdot v_i\rfloor$). Then $z/||z||_2 \in V$, and

$$\langle v, z/||z||_2\rangle_{\mathbb{R}} = \frac{1}{||z||_2}\sum_i v_i\lfloor Rv_i\rfloor \geq \frac{1}{R}\sum_i Rv_i^2 - v_i = 1 - ||v||_1/R \geq 1 - \epsilon.$$

For the size $|V|$, it is not hard to verify that $|V| \leq \mathrm{Vol}(B((1+\epsilon/2)R))$, by considering all unit cubes centered at the points of W, and noting that each such cube lies inside $B((1+\epsilon/2)R)$. Letting $\gamma = (1+\epsilon/2)/\epsilon$, we get $|V| \leq \mathrm{Vol}(B(\gamma\sqrt{n}))$. Assuming for convenience that n is even, we have

$$\mathrm{Vol}(B(\gamma\sqrt{n})) = \frac{\pi^{n/2}(\gamma\sqrt{n})^n}{(n/2)!} \leq \frac{\gamma^n\pi^{n/2}n^{n/2}}{\left(\frac{n}{2e}\right)^{n/2}} = \left(\gamma\sqrt{2\pi e}\right)^n.$$

Finally, for $\epsilon < 1/3$, the expression $\gamma\sqrt{2\pi e}$ is bounded by $5/\epsilon$. □

Chapter 8

Randomly Supported Independence

In this chapter, we study questions which, somewhat informally, can be described as follows: given a random set X of m points from Ω^n, for some finite set Ω, what is the probability p that there exists a k-wise independent distribution η over Ω^n such that $\mathrm{Supp}(\eta) \subseteq X$? In particular, we will be interested in how p increases as m increases.

If we ask how large m needs to be in order for $p > 0$, we are simply asking about the minimum size of the support of a k-wise independent distribution over Ω^n. This is a question which has been explored a lot in the past, and there is a rich literature of results (see Section 3.5).

Here, we will focus on the "high end" of the scale, and ask, how large does m have to be in order to have $p \geq 1 - o(1)$ (where we think of both m and p as functions of n and in particular, the $o(1)$ refers to something which tends to 0 as $n \to \infty$), or even $p = 1$?

Apart from being interesting mathematical questions by themselves, these questions are motivated by the connection between pairwise independence and approximation resistance, given in Chapter 5. Given that k-wise independence is of fundamental importance in many areas of computer science, we hope that these results may find further applications in the future.

The main theorem of this chapter is Theorem 8.4.1, which says that with high probability, $c(q)n^2$ random points in $[q]^n$ can support a pairwise independent distribution. For higher independence, we prove the somewhat weaker Theorem 8.5.1, which says that with high probability $(c(q) \cdot n)^k \log(n^k)$ random points in $[q]^n$ can support a k-wise independent distribution. We also give a lower bound, stating that with high probability, fewer than $\Omega\left(\frac{n^k}{q^{k^2}k^k}\right)$ random points do not support a *balanced* k-wise independent distribution. For $k = 2$ this matches Theorem 8.4.1 up to a constant, and for $k = \mathcal{O}(1)$ it matches Theorem 8.5.1 up to a factor $\Theta(\log n)$. It would be interesting to understand the dependency on k better.

For the question of when a subset of $[q]^n$ certainly supports a k-wise independent distribution, we prove that *every* subset of $[q]^n$ with size at least $q^n(1 - c(q)^{-k})$ support a k-wise independent distribution.

As we show in Theorem 8.2.2, the question of randomly supported independence can be viewed as a question about random polytopes, in particular, whether a certain random polytope contains the origin. In this setting, Füredi [39] proved that, for the special case of the uniform distribution on the boolean hypercube $\{-1, 1\}^n$, the threshold for a set of m random points to be likely to support a "1-wise independent" distribution, i.e., a distribution in which each bit is uniform, is exactly $m = 2n$. Unfortunately, Füredi's very elegant proof can not be adapted to k-wise independence for $k \geq 2$.

8.1 Definitions

Throughout this chapter, we fix some finite product space $(\Omega^n, \mu^{\otimes n})$, with a Fourier basis $\{\chi_\sigma\}_{\sigma \in \mathbb{Z}_q^n}$. As we will frequently use the set of multi-indices σ with cardinality $1 \leq |\sigma| \leq k$, we let

$$D_k = \{\, \sigma \in \mathbb{Z}_q^n \mid 1 \leq |\sigma| \leq k \,\}.$$

Note that $|D_k| = \sum_{i=1}^{k} (q-1)^i \binom{n}{i} \leq (qn)^k$.

Definition 8.1.1. Given a vector $x \in \Omega^n$, we define $x^{:\leq k:}$ as

$$x^{:\leq k:} := \bigoplus_{\sigma \in D_k} \chi_\sigma(x) \in \mathbb{R}^{D_k},$$

Here, $\oplus$ denotes the direct sum, e.g., $a \oplus b \oplus c = (a, b, c)$. In other words, $x^{:\leq k:}$ is the vector obtained by writing down the values of all non-constant monomials of degree at most k, evaluated at x. For a set $X \subseteq \Omega^n$, we use $X^{:\leq k:} \subseteq \mathbb{R}^{D_k}$ to denote the set $\{\, x^{:\leq k:} \mid x \in X \,\}$.

Note that every $v \in \mathbb{R}^{D_k}$ corresponds to a degree-k polynomial $f_v \in L^2(\Omega^n, \mu^{\otimes n})$ with $\mathbb{E}[f_v] = 0$, defined by $f_v(x) = \left\langle v, x^{:\leq k:} \right\rangle_{\mathbb{R}}$ for every $x \in \Omega^n$ (i.e., we simply interpret v as the Fourier coefficients of f_v).

Given a set $X \subseteq \mathbb{R}^n$, $\mathrm{conv}(X)$ denotes the *convex hull* of X, defined as the minimum convex set containing X. For $X = \{x_1, \ldots, x_m\}$ finite, $\mathrm{conv}(X)$ is simply the set of all points which are convex combinations of $x_1, \ldots, x_m$,

$$\mathrm{conv}(X) = \{\, \sum_{i=1}^{m} \alpha_i x_i \,:\, \alpha_i \geq 0, \sum_{i=1}^{m} \alpha_i = 1 \,\}.$$

8.2 Limited Independence and Low-Degree Polynomials

First, we characterize the sets $X \subseteq \Omega^n$ which support k-wise independent distributions, in terms of degree-k polynomials over Ω^n. We begin with the following easy proposition.

Proposition 8.2.1. *Let $(\Omega^n, \mu^{\otimes n})$ be a finite product space with Fourier basis $\{\chi_\sigma\}_{\sigma \in \mathbb{Z}_q^k}$, and let (Ω^n, η) be an arbitrary probability space. Then $\mu^{\otimes n} = \eta$ if and only if*

$$\mathop{\mathbb{E}}_{x' \in (\Omega^n, \eta)}[\chi_\sigma(x')] = 0$$

for every $\sigma \in \mathbb{Z}_q^k$ with $|\sigma| > 0$.

Proof. Define $f : \Omega^n \to \mathbb{R}$ by $f(x) = \eta(x)/\mu^{\otimes n}(x)$. Note that $\eta = \mu^{\otimes n}$ iff f is a constant, i.e., iff $\mathrm{Var}[f] = 0$, which happens iff $\hat{f}(\sigma) = 0$ for every $\sigma \neq \mathbf{0}$. Let us then compute $\hat{f}$. We have

$$\begin{aligned} \hat{f}(\sigma) &= \langle \chi_\sigma, f \rangle = \mathop{\mathbb{E}}_{x \in (\Omega^n, \mu^{\otimes n})}[\chi_\sigma(x)\eta(x)/\mu^{\otimes n}(x)] \\ &= \sum_{x \in \Omega^n} \mu^{\otimes n}(x)\chi_\sigma(x)\eta(x)/\mu^{\otimes n}(x) = \mathop{\mathbb{E}}_{x \in (\Omega^n, \eta)}[\chi_\sigma(x)]. \end{aligned}$$

Thus, $\eta = \mu^{\otimes n}$ if and only if

$$\mathop{\mathbb{E}}_{x \in (\Omega^n, \eta)}[\chi_\sigma(x)] = 0$$

for all $\sigma \neq \mathbf{0}$, as desired. □

We now state the characterization of the subsets of Ω^n that support k-wise independent distributions.

Theorem 8.2.2. *Let $X \subseteq \Omega^n$ be a set of vectors. Then, the following conditions are equivalent:*

(1) There exists a k-wise independent distribution η over Ω^n with marginals μ such that $\mathrm{Supp}(\eta) \subseteq X$

(2) $\mathbf{0} \in \mathrm{conv}(X^{:\leq k:})$

(3) There is no degree k polynomial $f \in L^2(\Omega^n, \mu^{\otimes n})$ such that $f(x) > \mathbb{E}[f]$ for every $x \in X$.

This characterization is likely already known, but as we have not been able to find it in the literature, we give a proof here.

Proof. **(1) ⇔ (2).** We view $\mathrm{conv}(X^{:\leq k:})$ as the set of probability distributions over Ω^n supported on X. Any convex combination $\sum_{x \in X} \alpha_x \cdot x^{:\leq k:} \in \mathrm{conv}(X^{:\leq k:})$ corresponds to the probability distribution η_α over Ω^n in which

$$\eta_\alpha(x) = \begin{cases} \alpha_x & \text{if } x \in X \\ 0 & \text{otherwise} \end{cases}.$$

Thus, it suffices to prove that, for every convex combination $\{\alpha_x\}_{x \in X}$, the corresponding distribution η_α has all k-dimensional marginals equal to $\mu^{\otimes k}$ if and only

if $\sum \alpha_x x^{:\leq k:} = \mathbf{0}$. This in turn is an immediate consequence of Proposition 8.2.1. Formally, fix an arbitrary convex combination $\{\alpha_x\}$. For a subset $S = \{i_1, \ldots, i_k\} \subseteq [n]$, we write η_S for the marginal distribution $\eta_\alpha|_S$.

If $\sum_{x \in X} \alpha_x x^{:\leq k:} = \mathbf{0}$, then in particular for every S with $|S| = k$ and every $\mathbf{0} \neq \sigma \subseteq S$, the σ:th coefficient of $\sum_{x \in X} \alpha_x x^{:\leq k:}$ is 0. In other words, for every such S and σ,

$$\mathop{\mathbb{E}}_{x_S \in (\Omega^S, \eta_S)} [\chi_\sigma(x)] = 0, \tag{8.1}$$

which by Proposition 8.2.1 implies that for every S it holds that $\eta_S = \mu^{\otimes k}$ (up to an appropriate identification of the indices).

Conversely, if $\eta_S = \mu^{\otimes k}$, then Proposition 8.2.1 implies that Equation (8.1) holds for every $\mathbf{0} \neq \sigma \subseteq S$ and hence the σ:th coordinate of $\sum_{x \in X} \alpha_x x^{:\leq k:}$ is 0 for all such σ. This implies that if $\eta_S = \mu^{\otimes k}$ for every S with $|S| = k$, every coordinate of $\sum \alpha_x x^{:\leq k:}$ is 0 and hence $\sum \alpha_x x^{:\leq k:} = \mathbf{0}$.

(2) $\Leftrightarrow$ (3). Without loss of generality, we can restrict our attention to f such that $\mathbb{E}[f] = 0$. Now, $\mathbf{0}$ is *not* in the convex hull of $X^{:\leq k:}$ if and only if there exists a separating hyperplane $v \in \mathbb{R}^{D_k}$ such that $\left\langle v, x^{:\leq k:} \right\rangle_{\mathbb{R}} > 0$ for every $x \in X$. The equivalence now follows by the correspondence between $v \in \mathbb{R}^{D_k}$ and degree-k polynomials f with $\mathbb{E}[f] = 0$.

□

8.3 Polynomials Are Balanced

In this section we prove that low-degree polynomials must exceed their expectation by a constant amount on a constant fraction of inputs.

Theorem 8.3.1. *For every probability space (Ω, μ) there is a $c = \mathrm{poly}(\alpha(\mu))$ such that for any degree-d polynomial $f \in L^2(\Omega^n, \mu^{\otimes n})$ with $\mathbb{E}[f] = 0$ and $\mathrm{Var}[f] = 1$,*

$$\Pr[f > c^d] > c^d.$$

A similar statement can be found in [26]. They lower bound $\Pr[f > 0]$ rather than $\Pr[f > c^d]$ and only consider $\Omega = \{-1, 1\}$, but these differences are superficial, and their proof (which is quite different from the one below) could be adapted to a proof of Theorem 8.3.1 as well.

Proof. Let $\delta = \mathrm{poly}(\alpha(\mu)) > 0$ be the constant from Theorem 7.1.8 (so that $||f||_1 \geq \delta ||f||_2$), and let $c = (\delta/4)^2$.

Define $g \in L^2(\Omega^n, \mu^{\otimes n})$ by

$$g(x) = \mathbf{1}_{[f > c^d]}(x) \cdot f(x) = \begin{cases} f(x) & \text{if } f(x) > c^d \\ 0 & \text{otherwise} \end{cases}.$$

We will lowerbound $\Pr[f > c^d] = \Pr[g > 0]$ by the second moment:

$$\Pr[g > 0] \geq \frac{\mathbb{E}[g]^2}{\mathbb{E}[g^2]} > ||g||_1^2,$$

where the last inequality follows from $\mathbb{E}[g^2] < \mathbb{E}[f^2] = 1$. For $||g||_1$, note that, since $\mathbb{E}[f] = 0$, we have $\mathbb{E}[\mathbf{1}_{[f>0]} \cdot f] = \frac{1}{2}||f||_1$, implying that

$$||g||_1 = \mathbb{E}[g] = \frac{1}{2}||f||_1 - \mathbb{E}[\mathbf{1}_{[0<f\leq c^d]}f] \geq \frac{1}{2}||f||_1 - c^d,$$

which, by hypercontractivity, is lower-bounded by

$$\frac{1}{2}\delta^d||f||_2 - c^d \geq \sqrt{c}^d$$

so that $\Pr[g > 0] > c^d$, as desired. □

As an easy corollary, we see that for every k, any set $X \subseteq \Omega^n$ of sufficiently large constant density supports a k-wise independent distribution.

Corollary 8.3.2. *For every probability space (Ω, μ) there exists a $c = \text{poly}(\alpha(\mu))$ such that every set $X \subseteq \Omega^n$ of total measure $\mu^{\otimes n}(X) \geq 1 - c^k$ supports a k-wise independent distribution with marginals μ.*

The proof is just a direct consequence of Theorems 8.2.2 and 8.3.1. While the corollary only needs the weaker bound $\Pr[f > 0] \geq c^k$ rather than $\Pr[f > c^k] \geq c^k$ as Theorem 8.3.1 gives, the stronger form will be necessary for the proof of Theorem 8.5.1 in Section 8.5.

We note that the exponential dependence on the degree/independence in both Theorem 8.3.1 and Corollary 8.3.2 is tight, which can be seen by looking at the function $f : \{-1, 1\}^n \to \mathbb{R}$ defined by

$$f(x) = \prod_{i=1}^{d}(1 - x_i) - 1,$$

which is a degree-d polynomial under the uniform distribution on $\{-1, 1\}^n$, which has value $2^d - 1$ with probability 2^{-d}, and value -1 with probability $1 - 2^{-d}$.

8.4 Pairwise Independence

In this section, we prove the following theorem.

Theorem 8.4.1. *For every (Ω, μ) there are constants $c, \delta > 0$ such that the following holds. Let $x_1, \ldots, x_m \in \Omega^n$ be a sequence of m independent samples from $(\Omega^n, \mu^{\otimes n})$. Then, if $m > cn^2$, the probability that $X = \{x_1, \ldots, x_m\}$ contains a pairwise independent distribution with marginals μ is at least $1 - \exp(-\delta\sqrt{n})$*

Before proceeding with the proof of Theorem 8.4.1, let us briefly describe the intuition behind it. The idea is to look at the convex hull K of the set of all ± 1 combinations of $x_1^{:\leq 2:}, \ldots, x_m^{:\leq 2:}$, and compare this to the sum $\overline{x} = x_1^{:\leq 2:} + \ldots + x_m^{:\leq 2:}$. By a simple application of Theorem 8.2.2, it suffices to prove that the latter sum

lies inside K with high probability. Intuitively, since $\overline{x}$ is a sum of m independent vectors with expected value $\mathbf{0}$ and length about $\sqrt{|D_2|}$, the total length of $\overline{x}$ should be around $\sqrt{m|D_2|}$. On the other hand, K consists of all $[-1,1]$-valued linear combinations of $x_1^{:\leq 2:}, \ldots, x_m^{:\leq 2:}$ and as an easy consequence of hypercontractivity it will turn out that, in every direction v, each $x_i^{:\leq 2:}$ contributes a constant to the expected width of K in direction v. Thus one can hope that the size of K grows linearly in m so that if m is a sufficiently large multiple of $|D_2|$, K is a lot larger than $||x|| \approx \sqrt{m|D_2|}$. It turns out that this is indeed the case, but in order to be able to show that the size of K grows linearly in *every* direction, we need to use the concentration inequality Theorem 7.2.2 for quadratic polynomials. It is this part which breaks down when one tries to repeat the same proof for k-wise independence in general—the analogue of Theorem 7.2.2 is simply not true any more. We feel that this limitation to pairwise independence is a limitation of our proof rather than an inherent limitation in the problem, and that the analogue of Theorem 8.4.1 (where we require $m > (cn)^k$) should be true also for higher independence.

Finally, we remark that it can be seen that the constant c in Theorem 8.4.1 can be chosen as $\mathrm{poly}(1/\alpha(\mu))$. Being careful, one can take c to be $\Theta(1/\alpha^4 \log 1/\alpha)$. It would be interesting to determine exactly how small c can be—a reasonable guess seems to be $c'/\alpha(\mu)^2$ for some universal constant c', though the present proof does not reach this bound. The main bottleneck in the current proof turns out to be an application of Theorem 7.2.2, and in particular the fact that the value of δ in Theorem 7.2.2 will be of order $\Theta(\epsilon^2\alpha^2)$.

Proof of Theorem 8.4.1. Let $m > c_0|D_2|$, where c_0 is a constant that will be chosen sufficiently large. We will prove that, with probability at least $1 - \exp(-\delta\sqrt{n})$, for some $\delta > 0$, we have $\mathbf{0} \in \mathrm{conv}(X^{:\leq 2:})$. By Theorem 8.2.2 this implies that X contains a pairwise independent distribution. This then implies Theorem 8.4.1 with $c := c_0 q^2$, since $|D_2| = (q-1)n + (q-1)^2\binom{n}{2} \leq q^2 n^2$.

Let

$$K = \left\{ \sum_{i=1}^{m} a_i x_i^{:\leq 2:} \,:\, |a_i| \leq 1 \right\},$$

and define

$$\overline{x} = \sum_{i=1}^{m} x_i^{:\leq 2:} \in \mathbb{R}^{D_2}.$$

Then, it suffices to prove that $\overline{x}$ lies in the interior of K, since if $\overline{x} = \sum_i a_i x^{:\leq 2:}$ with not all $a_i = 1$, we can rearrange and write $\mathbf{0}$ as the convex combination

$$\mathbf{0} = \sum_{i=1}^{m} \frac{1 - a_i}{\sum_j (1 - a_j)} x_i^{:\leq 2:} \in \mathrm{conv}(X^{:\leq 2:}).$$

For a unit vector $v \in \mathbb{R}^{D_k}$, let

$$\mathrm{width}(K, v) = \sup_{x \in K} \{\langle x, v \rangle_{\mathbb{R}}\}$$

be the *width* of K in the direction v.

We will prove that, with high probability, the minimum width of K is larger than $||\overline{x}||$ (where $||\cdot||$ denotes the standard Euclidean norm in $\mathbb{R}^{D_k}$). In particular, we have the following two lemmas.

Lemma 8.4.2. *There are constants $c_1 \in \mathbb{R}$, $c_2 > 0$ and $\delta_1 > 0$ such that, if $m > c_1|D_2|$, the probability that*

$$\inf_v \text{width}(K, v) < c_2 m \tag{8.2}$$

is at most $\exp(-\delta_1 m)$.

Lemma 8.4.3. *There is a constant $\delta_2 > 0$ such that if $m \geq |D_2|$, the probability that*

$$||\overline{x}|| > 2\sqrt{m|D_2|} \tag{8.3}$$

is at most $\exp(-\delta_2\sqrt{n})$.

Before proving the lemmas, let us see how they suffice to finish the proof of Theorem 8.4.1. Let $c_0 = \max(c_1, (2/c_2)^2)$, and $m > c_0|D_2|$. Then by a union bound there is a δ such that with probability at least $1 - \exp(-\delta\sqrt{n})$, neither Equation (8.2) nor Equation (8.3) hold, and we have

$$\inf_v \text{width}(K, v) \geq c_2 m > 2\sqrt{m|D|} \geq ||\overline{x}||.$$

This implies that $\overline{x}$ lies strictly inside K, as desired. Hence, if $m > cq^2n^2 \geq c_0|D_2|$, the probability that $\mathbf{0} \in \text{conv}(X^{:\leq 2:})$ is at least $1-\exp(-\delta\sqrt{n})$, and we are done. □

It remains to prove the two lemmas. We begin with Lemma 8.4.3 as this is the easier of the two.

Proof of Lemma 8.4.3. Let

$$l = ||\overline{x}||^2 = \sum_{\sigma \in D_2} \left(\sum_{i=1}^{m} \chi_\sigma(x_i) \right)^2$$

be the squared length of $\overline{x}$. We can then view l as a degree 4 polynomial over $L^2(\Omega^{nm}, \mu^{\otimes mn})$. Our goal is to apply the concentration bound Theorem 7.2.1 to l. To be successful in this, we need that the variance $\text{Var}[l]$ is of a lower order than $\mathbb{E}[l]^2$. The expectation of l is easily seen to be $\mathbb{E}[l] = |D_2|m$. To compute the variance of l, we compute

$$\begin{aligned} l^2 &= \sum_{\sigma_1, \sigma_2} \left(\sum_{i=1}^{m} \chi_{\sigma_1}(x_i) \right)^2 \left(\sum_{i=1}^{m} \chi_{\sigma_2}(x_i) \right)^2 \\ &= \sum_{\sigma_1, \sigma_2} \sum_{i_1, i_2, i_3, i_4 \in [m]} \chi_{\sigma_1}(x_{i_1}) \chi_{\sigma_1}(x_{i_2}) \chi_{\sigma_2}(x_{i_3}) \chi_{\sigma_2}(x_{i_4}). \end{aligned}$$

Define

$$S(\sigma_1, \sigma_2) = \sum_{i_1,i_2,i_3,i_4 \in [m]} \chi_{\sigma_1}(x_{i_1})\chi_{\sigma_1}(x_{i_2})\chi_{\sigma_2}(x_{i_3})\chi_{\sigma_2}(x_{i_4}),$$

and let us analyze $\mathbb{E}[S(\sigma_1, \sigma_2)]$. If $\sigma_1 \neq \sigma_2$, the expected value of

$$\chi_{\sigma_1}(x_{i_1})\chi_{\sigma_1}(x_{i_2})\chi_{\sigma_2}(x_{i_3})\chi_{\sigma_2}(x_{i_4})$$

is 0 unless $i_2 = i_1$ and $i_4 = i_3$. Hence for $\sigma_1 \neq \sigma_2$, we have

$$\mathbb{E}[S(\sigma_1, \sigma_2)] = \sum_{i_1,i_3} \mathbb{E}[\chi_{\sigma_1}(x_{i_1})^2\chi_{\sigma_2}(x_{i_3})^2].$$

The terms where $i_1 \neq i_3$ contribute 1 to this sum, and the terms where $i_1 = i_3$ contribute at most $1/\alpha^2$ by Fact 2.3.4. Hence we have for $\sigma_1 \neq \sigma_2$

$$\mathbb{E}[S(\sigma_1, \sigma_2)] \leq m^2 + m/\alpha^2.$$

Now let $\sigma_1 = \sigma_2 := \sigma$, and consider the expected value of

$$\chi_{\sigma}(x_{i_1})\chi_{\sigma}(x_{i_2})\chi_{\sigma}(x_{i_3})\chi_{\sigma}(x_{i_4}).$$

If for any $j \in [m]$ it is the case that only one of the i_k:s equal j, this expectation is 0. Thus the only tuples (i_1, i_2, i_3, i_4) for which the expectation is not 0 are those where the values are paired up in the sense that $i = j$ and $k = l$, or $i = k$ and $j = l$, or $i = l$ and $j = k$. There are exactly $3m(m-1) + m \leq 3m^2$ ways to choose i_1, i_2, i_3, i_4 in such a paired way and hence in this case

$$\mathbb{E}[S(\sigma, \sigma)] \leq 3m^2/\alpha^2,$$

where we again used Fact 2.3.4. After these lengthy computations we thus find that

$$\mathbb{E}[l^2] = \sum_{\sigma_1,\sigma_2} \mathbb{E}[S(\sigma_1, \sigma_2)] \leq |D_2|^2m^2 + |D_2|^2m/\alpha^2 + 3|D_2|m^2/\alpha^2,$$

so that

$$\mathrm{Var}[l] \leq |D_2|^2m/\alpha^2 + 3|D_2|m^2/\alpha^2 \leq 4|D_2|m^2/\alpha^2,$$

where the last inequality assumed that $m \geq |D_2|$. Applying Theorem 7.2.1 to the polynomial $(L - \mathbb{E}[L])/\sqrt{\mathrm{Var}[L]}$, we have

$$\begin{aligned} \Pr[||x|| > 2\sqrt{|D_2|m}] &= \Pr[L - \mathbb{E}[L] > 3|D_2|m] \\ &\leq \exp(-c(3|D_2|m/\sqrt{\mathrm{Var}[L]})^{1/2}) \leq \exp(-c'|D_2|^{1/4}), \end{aligned}$$

for $c' = c\sqrt{3\alpha/2}$. Since $|D_2| \geq q^2n^2$, the lemma follows. □

We now move on to the proof of Lemma 8.4.2, which is a bit more involved. We will lower bound the minimum width of K by first proving that the width of K in any direction is sharply concentrated around its expected value, then using this to prove that the *maximum* width of K is bounded, which together with the concentration result also gives that the minimum width is bounded.

We begin with stating and proving the concentration result.

Lemma 8.4.4. *There is a constant $c_3 := \text{poly}(\alpha(\mu))$ and $\tau > 0$ such that the following holds: for every $v \in \mathbb{R}^D$ with $||v|| = 1$, the probability that*

$$c_3 m \leq \text{width}(K, v) \leq (1 + c_3)m$$

is at least $1 - \exp(-\tau m)$.

Proof. Set $2c_3 := (\alpha(\mu)/8)$, the constant from Theorem 7.1.8 for $d = 2$.

For $v \in \mathbb{R}^D$ with $||v|| = 1$, let $f_v \in L^2(\Omega^n, \mu^{\otimes n})$ be the corresponding degree-2 polynomial such that $f_v(x) = \left\langle v, x^{\otimes \leq 2} \right\rangle$.

By definition, $\text{width}(K, v) = \max_{a \in [-1,1]^m} \sum_{i=1}^m a_i \left\langle v, x_i^{\otimes \leq 2} \right\rangle_{\mathbb{R}}$. The maximum is clearly attained by setting $a_i = \text{sgn}\left(\left\langle v, x_i^{\otimes \leq 2} \right\rangle\right)$ so that

$$\text{width}(K, v) = \sum_{i=1}^m \left|\left\langle v, x_i^{\otimes \leq 2} \right\rangle\right| = \sum_{i=1}^m |f_v(x_i)|.$$

By Theorem 7.2.2 the probability that $\sum_i |f_v(x_i)|$ deviates by more than $c_3 m$ from its expectation is at most $\exp(-\tau m)$ for some constant τ depending only on μ and c_3. But the expectation of $\sum_i |f_v(x_i)|$ equals $||f||_1 \cdot m$, which is trivially upper bounded by $||f||_2 \cdot m = m$, and by Theorem 7.1.8 lower bounded by $2c_3||f||_2 \cdot m = 2c_3 m$.

Hence, with probability at least $1 - \exp(-\tau m)$, we have

$$\begin{aligned} (||f||_1 - c_3)m &\leq \text{width}(K, v) \leq (||f||_1 + c_3)m \\ c_3 m &\leq \text{width}(K, v) \leq (1 + c_3)m. \end{aligned}$$

□

We now prove the lower bound on the minimum width of K.

Proof of Lemma 8.4.2. Let $V = \{v_1, \ldots, v_L\}$ be an ϵ-net of the unit sphere in $\mathbb{R}^{D_2}$, i.e., a set of vectors such that, for every $v \in \mathbb{R}^{D_2}$ with $||v|| = 1$, there is a vector $v_i \in V$ such that $\langle v, v_i \rangle_{\mathbb{R}} \geq 1 - \epsilon$. Such a set can be constructed of size at most $L = (5/\epsilon)^{|D_2|}$ (Theorem 7.3.1).

For any $v_i \in V$, Lemma 8.4.4 tells us that

$$c_3 m \leq \text{width}(K, v_i) \leq (1 + c_3)m$$

except with probability at most $\exp(-\tau m)$. By a union bound, these inequalities then hold for every $v_i \in V$ except with probability

$$L\exp(-\tau m) = \exp(-\tau m + \ln(5/\epsilon)|D_2|) = \exp(-\tau m/2),$$

provided m is a sufficiently large multiple of $|D_2|$.

Let $W_{\max} = \sup_{||v||=1} \text{width}(K, v)$. We now prove that $W_{\max}$ is small.

For any $w \in \mathbb{R}^D$ with $||w|| = 1$, we can write $w = (1-\epsilon)v_i + \sqrt{2\epsilon}w'$ for some $v_i \in V$ and vector w' with $||w'|| \leq 1$. We then have for any $u \in K$

$$\begin{aligned}
\langle u, w\rangle &= (1-\epsilon)\langle u, v_i\rangle + \sqrt{2\epsilon}\langle u, w'\rangle \\
&\leq (1-\epsilon)\,\text{width}(K, v_i) + \sqrt{2\epsilon}\,\text{width}(K, w') \\
&\leq (1-\epsilon)(1+c_3)m + \sqrt{2\epsilon}W_{\max}.
\end{aligned}$$

Taking the supremum over all $u \in K$ and unit vectors $w \in \mathbb{R}^D$, we obtain

$$\begin{aligned}
W_{\max} &\leq (1-\epsilon)(1+c_3)m + \sqrt{2\epsilon}W_{\max} \\
W_{\max} &\leq \frac{(1-\epsilon)(1+c_3)}{1-\sqrt{2\epsilon}} \leq (1+2c_3)m,
\end{aligned}$$

provided ϵ is chosen sufficiently small compared to c_3.

But then, we have, again for any $w = (1-\epsilon)v_i + \sqrt{2\epsilon}w'$ and $u \in K$,

$$\begin{aligned}
\langle u, w\rangle &= (1-\epsilon)\langle u, v_i\rangle + \sqrt{2\epsilon}\langle u, w'\rangle \\
&\geq (1-\epsilon)c_3 m - \sqrt{2\epsilon}\,\text{width}(K, w') \\
&\geq ((1-\epsilon)c_3 - \sqrt{2\epsilon}(1+c_3))m \geq c_3/2m,
\end{aligned}$$

again provided ϵ is sufficiently small compared to c_3.

Hence, with probability at least $1-\exp(-\delta m)$, we have $\inf_{||v||=1} \text{width}(K, v) \geq c_3/2m := c_2 m$, provided that m is a sufficiently large multiple $c_1|D_2|$ of $|D_2|$. □

8.5 k-wise Independence

In this section, we prove a result similar to Theorem 8.4.1 for k-wise independence. Unfortunately, the bounds we get are not as strong as for pairwise independence, in the sense that we get an extra factor $\log(n^k)$ in the number of random points needed.

Theorem 8.5.1. *For every (Ω, μ) there are constants $c, \delta > 0$ such that the following holds. Let $x_1, \ldots, x_m \in \Omega^n$ be a sequence of m independent samples from $(\Omega^n, \mu^{\otimes n})$. Then, if $m > (cn)^k k \log n$, the probability that $X = \{x_1, \ldots, x_m\}$ contains a pairwise independent distribution with marginals μ is at least $1-\exp(-\delta n^k)$*

Proof. By Theorem 8.2.2, $x_1, \ldots, x_m$ *does not* support a k-wise independent distribution with marginals μ if and only if, there is a degree-k polynomial $f \in L^2(\Omega^n, \mu^{\otimes n})$, such that $f(x_i) < 0$ for every x_i.

For any fixed f, Theorem 8.3.1 gives that the probability that $f(x_i) < \tau^k$ for every x_i is at most $(1-\tau^k)^m \leq \exp(-\tau^k m)$, where $\tau = \mathrm{poly}(\alpha(\mu))$. Thus, it is clear that any fixed f has a very small probability of witnessing that $x_1, \ldots, x_m$ does not support a k-wise independent distribution.

To bound the probability that any f witnesses that $x_1, \ldots, x_m$ supports a k-wise independent distribution, we construct a net of degree-k polynomials as follows: let $\mathcal{F}_\delta$ denote the set of degree-k polynomials $f \in L^2(\Omega^n, \mu^{\otimes n})$ such that $\mathbb{E}[f] = 0$, $\mathrm{Var}[f] \leq 2$ and every Fourier coefficient of f is an integer multiple of δ.

We then have that $|\mathcal{F}_\delta| \leq (1/\delta)^{\mathcal{O}(|D_k|)} = \exp(c_1 q^k n^k \log 1/\delta)$ for some universal constant c_1. Then Theorem 8.3.1 and a union bound gives that the probability that there exists an $f \in \mathcal{F}_\delta$ such that $f(x_i) < \tau^k$ for every x_i, is bounded by

$$|\mathcal{F}_\delta|(1-\tau^k)^m \leq \exp(c_1 q^k n^k \log(1/\delta) - \tau^k m) \leq \exp(-\tau^k m/2),$$

provided $m \geq 2c_1(nq/\tau)^k \log(1/\delta)$.

Now, given a degree-k polynomial f with $\mathbb{E}[f] = 0$, denote by $\tilde{f}$ the polynomial in $\mathcal{F}_\delta$ which is closes to f in ℓ_∞ norm. Then, if $||f - \tilde{f}||_\infty \leq \tau^k$ for every degree-k polynomial f, we would be done, since the existence of $f \in L^2(\Omega^n, \mu^{\otimes n})$ such that $f(x_i) < 0$ for every x_i then implies that $\tilde{f}(x_i) \leq f(x_i) + |\tilde{f}(x_i) - f(x_i)| < \tau^k$, which happens with probability at most $\exp(-\tau^k m/2) := \exp(-\delta m)$.

We have the following easy bound on the distance $||f - \tilde{f}||_\infty$.

Claim 8.5.2. *For every f with $||f||_2 = 1$,*

$$||f - \tilde{f}||_\infty \leq \delta \left(\frac{nq}{\sqrt{\alpha(\mu)}}\right)^k,$$

provided this quantity is smaller than 1*.*

Proof. Let f' be the result of rounding every Fourier coefficient of f to its nearest multiple of δ. Then, for any $x \in \Omega^n$,

$$|f(x) - f'(x)| = \left|\sum_{\sigma \in D_k} (\hat{f}(\sigma) - \hat{f}'(\sigma))\chi_\sigma(x)\right| \leq \delta \sum_{\sigma \in D_k} ||\chi_\sigma||_\infty \leq \delta \left(\frac{nq}{\sqrt{\alpha(\mu)}}\right)^k,$$

where the last step used Fact 2.3.4 and $|D_k| \leq (nq)^k$. It remains to show that $f' \in \mathcal{F}_\delta$, i.e., that $\mathrm{Var}[f'] \leq 2$. But this follows immediately since

$$\mathrm{Var}[f'] = ||f'||_2 \leq ||f||_2 + ||f - f'||_2 \leq 1 + ||f - f'||_\infty \leq 2$$

provided the bound on $||f - f'||_\infty \leq 1$. □

To finish the proof of Theorem 8.5.1, we thus conclude that in order to have $||f - \tilde{f}||_\infty \leq \tau^k$, it suffices to take

$$\delta = \left(\frac{\sqrt{\alpha}\tau}{nq}\right)^k,$$

giving the bound

$$m \geq 2c_1(nq/\tau)^k \log(1/\delta) = (cn)^k k \log n$$

for c depending only on α, q and τ, which in turn depend only on (Ω, μ). □

8.6 A Lower Bound

In this section we give give a lower bound on the number of random points of Ω^n needed to get a set supporting a *balanced* k-wise independent distribution.

Theorem 8.6.1. *Let μ_U be the uniform distribution over Ω, and let $x_1, \ldots, x_m$ be a sequence of m independent samples from $(\Omega^n, \mu_U^{\otimes n})$. Then, if $m < \frac{n^k}{2q^{k^2}k^{2k}}$, the probability that $x_1, \ldots, x_m$ can support a k-wise independent distribution with marginals μ_U (i.e., a balanced k-wise independent distribution) is at most*

$$\exp(-\Theta(n^k/q^k)).$$

Proof. Let $x_1, \ldots, x_m$ be a set of m independent samples of $(\Omega^n, \mu_U^{\otimes n})$. We will prove that, if $m < \frac{n^k}{2q^{k^2}k^{2k}}$, then with high probability $x_1^{:\leq k:}, \ldots, x_m^{:\leq k:}$ are linearly independent. In particular, this implies that any convex combination of $x_1^{:\leq k:}, \ldots, x_m^{:\leq k:}$ is non-zero, so that, by Theorem 8.2.2, $x_1^{:\leq k:}, \ldots, x_m^{:\leq k:}$ does not support a k-wise independent distribution.

The main component of the proof is the following lemma.

Lemma 8.6.2. *Let $m \leq \frac{n^k}{2q^{k^2}k^{2k}}$, and let $y_1, \ldots, y_m \in \mathbb{R}^{D_k}$ be m arbitrary points. Then, the probability that a uniformly random point $x \in \Omega^n$ has $x^{:\leq k:}$ lying in the space spanned by $y_1, \ldots, y_m$ is at most $\exp(-\Theta(n^k/q^k))$.*

Let $m = \frac{n^k}{2q^{k^2}k^{2k}}$, and let $x_1, \ldots, x_m$ be m uniformly random points of Ω^n. Using Lemma 8.6.2, we conclude that the probability that $x_1^{:\leq k:}, \ldots, x_m^{:\leq k:}$ are linearly independent is at least

$$1 - m \exp\left(-\frac{n^k}{kq^k}\right) = 1 - \exp(-\Theta(n^k/q^k)),$$

which proves Theorem 8.6.1. □

Next, we turn to the proof of the lemma.

Proof of Lemma 8.6.2. Let $S \subseteq \mathbb{R}^{D_k}$ be the space spanned by the vectors $y_1, \ldots, y_m$. Then S has dimension at most m and hence is determined by at least $|D_k| - m$ linearly independent equations $v_1, \ldots, v_{|D_k|-m} \in \mathbb{R}^{D_k}$ such that $y \in S$ iff $\langle v_i, y \rangle_{\mathbb{R}} = 0$ for every $i \in [|D_k| - m]$. Equivalently, for $x \in \Omega^n$, we have $x^{:\leq k:} \in S$ iff $v_i(x) = 0$ for every i, where we again interpret v_i as a degree-k polynomial. We will prove that only an exponentially small fraction of all points $x \in \Omega^n$ satisfy these conditions.

In what follows, we define

$$d(n) := \sum_{i=1}^{k} (q-1)^i \binom{n}{i} \geq \left(\frac{(q-1)n}{k} \right)^k,$$

i.e., the size of the set D_k of indices, for a given value of n. Let $T(n, m)$ be the maximum possible number of solutions $x \in \Omega^n$ to a system of at least $d(n) - m$ linearly independent degree-k polynomials $v_1, \ldots, v_{d(n)-m}$. We will prove that

$$T(n, m) \leq (q^k - 1)^{n/k} \cdot \exp(km^{1/k}). \tag{8.4}$$

If $d(n) \leq m$ so that $n \leq m^{1/k} k/(q-1)$, we have the trivial bound $T(n, m) \leq q^n \leq \exp(km^{1/k})$, so let $d(n) > m$ and assume inductively that Equation (8.4) holds for all $n' < n$. Assume that there is a v_i which has degree exactly k (if all v_i have degree at most $k-1$, we would get an even better bound). Without loss of generality, we can take v_1 to have degree exactly k, and having a non-zero coefficient σ with active set $S(\sigma) = [k]$.

Next, eliminate (by standard Gaussian elimination) all coordinates σ' with $\sigma' \subseteq [k]$. As there are exactly $d(n) - d(n-k)$ such values of σ', the resulting system has at least $(d(n) - m) - (d(n) - d(n-k)) = d(n-k) - m$ equations, and hence has at most $T(n-k, m)$ solutions. Let us, for each such solution $x^* \in \Omega^{[n]-[k]}$, consider the number of ways of extending it to a solution for the original system. Plugging in x^* in the equation $v_1(x) = 0$, this equation becomes an equation of the form

$$p(x_{[k]}) = 0,$$

for some function $p : \Omega^k \to \mathbb{R}$. Furthermore, the function p is not identically zero, since $\hat{p}(\sigma) \neq 0$. This implies that the number of ways of extending x^* is at most $q^k - 1$, and hence we have

$$T(n, m) \leq (q^k - 1) \cdot T(n-k, m) \leq (q^k - 1)^{n/k} \cdot \exp(km^{1/k}).$$

Thus, the probability that $x^{:\leq k:}$ lies inside S for a uniformly random point $x \in \Omega^n$ is at most

$$(q^k - 1)^{n/k} \exp(km^{1/k})/q^n = (1 - q^{-k})^{n/k} \exp(km^{1/k}) \leq \exp\left(-\frac{n}{kq^k} + km^{1/k} \right).$$

Plugging in $m \leq \frac{n^k}{2q^{k^2}k^{2k}}$, the lemma follows. □

Chapter 9

Noise Correlation Bounds for Uniform Functions

This chapter reports on work attempting to achieve good noise correlation bounds for more general classes of functions than those of Section 2.5. In particular, we are interested in obtaining correlation bounds under pairwise independent distributions for functions with no large Fourier coefficients. Functions in which all Fourier coefficients are bounded by δ are sometimes called δ-uniform, hence the title of this chapter. The search for such bounds is motivated by their potential applicability to hardness of approximation, derandomization, and additive combinatorics (see Section 9.4).

Unfortunately, we do not quite reach the goal of achieving such general bounds. In particular, the main result Theorem 9.1.1 gives good bounds on $\langle f_1, \ldots, f_k \rangle_{\mathcal{N}}$ only in the case when the degrees of the functions are small.

We remark that earlier noise correlation bounds such as Theorem 2.5.1 are proved this way, and then extended to arbitrary functions by (sometimes complicated) truncation arguments. Such truncation arguments appear a lot more difficult to achieve in our setting. Some discussion on this appears in Section 9.3.

9.1 Main Theorem

Theorem 9.1.1. *Let (Ω, μ) be a pairwise independent product space $\Omega = \Omega_1 \times \ldots \times \Omega_k$. There is a constant C depending only on μ such that the following holds.*

Let $f_1, \ldots, f_k$ be functions $f_i \in L^2(\Omega_i^n, (\mu|_i)^{\otimes n})$. Denote by $\delta := \max_{\sigma \in \mathbb{Z}_q^n} |\hat{f}_1(\sigma)|$ the size of the largest Fourier coefficient of f_1, and let $D := \deg_{-2}(f_1, \ldots, f_k)$ denote the sum of the $k-2$ smallest degrees of $f_1, \ldots, f_k$. Then,

$$\langle f_1, \ldots, f_k \rangle_{\mathcal{N}} \leq C^D \delta \prod_{i=2}^{k} ||f_i||_2.$$

Furthermore, one can always take $C = \left(k\sqrt{\frac{q-1}{\alpha(\mu)}}\right)^3$. *If* μ *is balanced, i.e., if all marginals* $\mu|_i$ *are uniform, then there is a choice of Fourier basis such that one can take* $C = (k\sqrt{q-1})^3$.

We remark that, while Theorem 9.1.1 is very limited because of its requirement on the degrees of the f_i:s, the lack of any other assumptions is nice. In particular, we do not need to assume that the f_i:s are bounded, nor do we need any assumptions on μ beyond the pairwise independence condition.

Proof. We prove this by induction over n. If $n = 0$, the statement is easily verified (either $D = -\infty$, or $D = 0$, depending on whether one of the functions is 0 or not).[1]

Write $f_i = g_i + h_i$, where

$$g_i = \sum_{1 \notin \sigma} \hat{f}(\sigma)\chi_\sigma$$
$$h_i = \sum_{1 \in \sigma} \hat{f}(\sigma)\chi_\sigma,$$

i.e., g_i is the part of f_i which does not depend on x_1, and h_i is the part which depends on x_1. Then

$$\langle f_1, \dots, f_k \rangle_{\mathscr{N}} = \mathbb{E}\left[\prod f_i(X_i)\right] = \sum_{T \subseteq [k]} \mathop{\mathbb{E}}_{X}\left[\prod_{i \notin T} g_i(X_i) \prod_{i \in T} h_i(X_i)\right].$$

For $T \subseteq [k]$, define

$$E(T) = \mathop{\mathbb{E}}_{X}\left[\prod_{i \notin T} g_i(X_i) \prod_{i \in T} h_i(X_i)\right].$$

The key ingredient will be the following Lemma, bounding $E(T)$.

Lemma 9.1.2. *Let* $\emptyset \subseteq T \subseteq [k]$. *Then:*

- *If* $T = \emptyset$, *we have*
$$E(T) \leq C^D \delta \prod_{i=2}^{k} ||g_i||_2.$$
- *If* $1 \leq |T| \leq 2$, *we have*
$$E(T) = 0.$$

[1] We point out that $f_i \in L^2(\Omega_i^0, (\mu|_i)^{\otimes 0})$ does not formally make sense. However in this case, the appropriate way to view f_i is as an element of $L^2(\Omega_i^N, (\mu|_i)^{\otimes N})$ which only depends on the n first coordinates. In particular, for the case $n = 0$ we have that f_i is a constant.

- *If $|T| \geq 3$, we have*

$$E(T) \leq C^{D+2} \left(\frac{\sqrt{(q-1)/\alpha}}{C} \right)^{|T|} \delta \prod_{\substack{i \notin T \\ i \neq 1}} ||g_i||_2 \prod_{\substack{i \in T \\ i \neq 1}} ||h_i||_2.$$

Before proving the Lemma, let us see how to use it to finish the proof of Theorem 9.1.1.

Write $||h_i||_2 = \tau_i ||f_i||_2$ for some $\tau_i \in [0,1]$, so that $||g_i||_2 = \sqrt{1-\tau_i^2}||f_i||_2$ (by orthogonality of the Fourier decomposition). By plugging in the different cases of Lemma 9.1.2, we can then bound $\langle f_1, \ldots, f_k \rangle_{\mathscr{N}}$ by

$$\begin{aligned}
&\langle f_1, \ldots, f_k \rangle_{\mathscr{N}} \\
&\quad \leq \quad C^D \delta \prod_{i=2}^{k} ||g_i||_2 + \sum_{|T| \geq 3} C^{D+2} \left(\frac{\sqrt{(q-1)/\alpha}}{C} \right)^{|T|} \delta \prod_{\substack{i \notin T \\ i \neq 1}} ||g_i||_2 \prod_{\substack{i \in T \\ i \neq 1}} ||h_i||_2 \\
&\quad = \quad C^D \delta \prod_{i=2}^{k} ||f_i||_2 \Bigg(\prod_{i=2}^{k} \sqrt{1-\tau_i^2} + \\
&\qquad\qquad + \sum_{|T| \geq 3} C^2 \left(\frac{\sqrt{(q-1)/\alpha}}{C} \right)^{|T|} \prod_{\substack{i \notin T \\ i \neq 1}} \sqrt{1-\tau_i^2} \prod_{\substack{i \in T \\ i \neq 1}} \tau_i \Bigg).
\end{aligned}$$

Hence, it suffices to bound the "error factor" inside the large parenthesis by 1 in order to complete the proof of Theorem 9.1.1.

Let $\tau = \max_{i \geq 2} \tau_i$. The error factor can then be bounded by

$$\sqrt{1-\tau^2} + \tau^2 \sum_{i=3}^{k} \binom{k}{i} \left(\frac{((q-1)/\alpha)^{3/2}}{C} \right)^{i/3},$$

where we assumed that $C > 1$ and then used that, for $i \geq 3$, $C^{2-i} \leq C^{-i/3}$. To bound this, we use the following simple lemma:

Lemma 9.1.3. *For every $k \geq 3$,*

$$\sum_{i=3}^{k} \binom{k}{i} \frac{1}{k^i} \leq 1/2.$$

Proof. Since $\binom{k}{i} \leq k^i/i!$ we have

$$\sum_{i=3}^{k} \binom{k}{i} \frac{1}{k^i} \leq \sum_{i=3}^{k} \frac{1}{i!} \leq e - 5/2 \leq 1/2,$$

where the second inequality is by the Taylor expansion $e = \sum_{i=0}^{\infty} \frac{1}{i!} \leq \sum_{i=0}^{k} \frac{1}{i!}$. □

Hence, if $C \geq \left(k\sqrt{\frac{q-1}{\alpha}}\right)^3$, the error factor is bounded by

$$\sqrt{1-\tau^2} + \tau^2/2 \leq 1.$$

This concludes the proof of Theorem 9.1.1. We have not yet addressed the claim that if the marginals $\mu|_i$ are uniform, there is a Fourier basis such that C can be chosen as $(k\sqrt{q-1})^3$. See the comment after the proof of Lemma 9.1.2. □

We now prove the lemma used in the previous proof.

Proof of Lemma 9.1.2. The case $T = \emptyset$ is a direct application of the induction hypothesis, since the functions g_i depend on at most $n-1$ variables (and have $\deg_{-2}(g_1, \ldots, g_k) \leq D$).

Write

$$h_i(x) = \sum_{j=1}^{q-1} \chi_j(x_1) h_{i,j}(x_2, \ldots, x_n)$$

for a Fourier basis $\chi_0 = 1, \chi_1, \ldots, \chi_{q-1}$ of $L^2(\Omega_i, \mu|_i)$. Denoting by X^j the jth column of X, we can write $E(T)$ as

$$\begin{aligned} E(T) &= \mathop{\mathbb{E}}_{X^2,\ldots,X^n}\left[\prod_{i\notin T} g_i(X_i) \mathop{\mathbb{E}}_{X^1}\left[\prod_{i\in T} h_i(X_i)\right]\right] \\ &= \mathop{\mathbb{E}}_{X^2,\ldots,X^n}\left[H(T,X)\cdot\prod_{i\notin T} g_i(X_i)\right], \end{aligned}$$

where

$$\begin{aligned} H(T,X) &= \mathop{\mathbb{E}}_{X^1}\left[\prod_{i\in T} h_i(X_i)\right] \\ &= \sum_{J\in[q-1]^T} \mathop{\mathbb{E}}_{X^1}\left[\prod_{i\in T}\chi_{J_i}(X_i^1)\right]\prod_{i\in T} h_{i,J_i}(X_i). \end{aligned}$$

Now for $1 \leq |T| \leq 2$, the pairwise independence of μ gives that for any $J \in [q-1]^T$,

$$\mathop{\mathbb{E}}_{X^1}\left[\prod_{i\in T}\chi_{J_i}(X_i^1)\right] = \prod_{i\in T}\mathbb{E}[\chi_{J_i}] = 0,$$

hence in this case $H(T,X) = 0$ and by extension $E(T) = 0$.

Thus, only the case $|T| \geq 3$ remains. By the repeated Hölder's inequality, we can bound

$$\mathop{\mathbb{E}}_{X^1}\left[\prod_{i\in T}\chi_{J_i}(X_i^1)\right] \leq \prod_{i\in T}||\chi_{J_i}||_{|T|}.$$

By Fact 2.3.4, this can be bounded by $(1/\alpha)^{|T|/2}$ Plugging this into $E(T)$ gives

$$E(T) \leq (1/\alpha)^{|T|/2}\mathop{\mathbb{E}}_{X^2,\ldots,X^n}\left[\sum_{\sigma\in[q-1]^T}\prod_{i\in T}h_{i,\sigma_i}(X_i)\prod_{i\notin T}g_i(X_i)\right].$$

For $J \in [q-1]^T$, let D_J be the sum of the $k-2$ smallest degrees of the polynomials $\{g_i : i \notin T\} \cup \{h_{i,J_i} : i \in T\}$. Since g_i and h_{i,J_i} are functions of $n-1$ variables, we can use the induction hypothesis to get a bound of

$$E(T) \leq (1/\alpha)^{|T|/2}\sum_{J\in[q-1]^T}C^{D_J}\delta\prod_{\substack{i\in T\\ i\neq 1}}||h_{i,J_i}||_2\prod_{\substack{i\notin T\\ i\neq 1}}||g_i||_2.$$

But since the h_{i,J_i}:s have strictly smaller degrees than the corresponding f_i:s, D_J is bounded by $D-|T|+2$, and hence we have that

$$\begin{aligned}E(T) &\leq C^{D-|T|+2}\sum_{J\in[q-1]^T}\delta\prod_{\substack{i\in T\\ i\neq 1}}||h_{i,J_i}||_2\prod_{\substack{i\notin T\\ i\neq 1}}||g_i||_2\\ &\leq C^{D+2}\left(\frac{\sqrt{(q-1)/\alpha}}{C}\right)^{|T|}\delta\prod_{\substack{i\in T\\ i\neq 1}}||h_i||_2\prod_{\substack{i\notin T\\ i\neq 1}}||g_i||_2,\end{aligned}$$

where we used the fact that $\sum_{j\in[q-1]}||h_{i,j}||_2 \leq \sqrt{q-1}||h_i||_2$ (by Cauchy-Schwarz and orthogonality of the functions $h_{i,j}$).

This concludes the proof of Lemma 9.1.2. □

In the case when the marginal distributions $\mu|_i$ are uniform, one can take as basis of (Ω,μ) the *complex* basis $\chi_j(x) = e^{2\pi i\frac{j\cdot x}{q}}$ (where we identify the elements x of Ω with $\mathbb{Z}_q$). For this basis, $||\chi_j||_\infty = 1$ and hence Equation (9.1) can be bounded by 1 rather than $1/\sqrt{\alpha}$, which propagates to give that, for this basis, we can choose $C = (k\sqrt{q-1})^3$.

9.2 Corollaries

Let us note some easy corollaries of Theorem 9.1.1. The first is a statement more in the spirit of Theorem 2.5.1, saying that, if all non-zero Fourier coefficients of f_1 are small, then the noisy inner product is close to the products of expectation.

Corollary 9.2.1. *Assume the setting of Theorem 9.1.1, but with*

$$\delta := \max_{i\in[k]} \max_{\sigma\neq\mathbf{0}} |\hat{f}_i(\sigma)|$$

being the largest non-zero *Fourier coefficient in any of the functions. Assume also that* $||f_i||_2 \leq F \leq 1$ *for all* i*. Then,*

$$\left| \langle f_1, \ldots, f_k \rangle_{\mathcal{N}} - \prod_{i=1}^{k} \mathbb{E}[f_i] \right| \leq \delta 2^k C^D F^2,$$

where C *is the constant from Theorem 9.1.1.*

Proof. Let $g_i(x) = f_i(x) - \mathbb{E}[f_i]$. Then,

$$\langle f_1, \ldots, f_k \rangle_{\mathcal{N}} = \sum_{S\subseteq[k]} \prod_{i\notin S} \mathbb{E}[f_i] \cdot \mathbb{E}\left[\prod_{i\in S} g_i\right] = \prod_{i=1}^{k} \mathbb{E}[f_i] + \sum_{|S|>0} \prod_{i\notin S} \mathbb{E}[f_i] \cdot \mathbb{E}\left[\prod_{i\in S} g_i\right].$$

By Theorem 9.1.1, for $|S| \geq 3$,

$$\mathbb{E}\left[\prod_{i\in S} g_i\right] \leq C^D \delta F^{|S|-1} \leq C^D \delta F^2,$$

whereas for $1 \leq |S| \leq 2$ the expected value vanishes because of the pairwise independence. Hence, we have

$$\langle f_1, \ldots, f_k \rangle_{\mathcal{N}} \leq \prod_{i=1}^{k} \mathbb{E}[f_i] + 2^k C^D \delta F^2.$$

The lower bound follows by replacing f_1 by $-f_1$ and applying the upper bound. □

The second corollary uses a standard "iteration argument" (more about these in Section 9.4) to show that, if the noisy inner product deviates from the product of expectations, not only can we find a large Fourier coefficient in each of the functions, we can actually find three functions with large intersecting coefficients. This type of result is often useful in applications to hardness of approximation, though the limitation of our current statement to low-degree functions seems to prevent such applications.

Corollary 9.2.2. *Assume the setting of Theorem 9.1.1, and further assume that* $\rho(\mu) < 1$*. Then for every* $\epsilon > 0$ *there exist* d *and* δ *(depending only on* ϵ *and* μ*) such that the following holds. Let* $\mathbb{E}[f_i] = 0$ *and* $||f_i||_2 \leq 1$ *for every* i*. Then if*

$$\left| \langle f_1, \ldots, f_k \rangle_{\mathcal{N}} - \prod_{i=1}^{k} \mathbb{E}[f_i] \right| > \epsilon,$$

there exist three distinct indices $i_1, i_2, i_3 \in [k]$*, and multi-indices* $\sigma_1, \sigma_2, \sigma_3$ *satisfying:*

- $|S(\sigma_j)| \leq d$ *for* $1 \leq j \leq 3$.
- $\sigma_1, \sigma_2, \sigma_3$ *intersect in the sense that* $S(\sigma_1) \cap S(\sigma_2) \cap S(\sigma_3) \neq \emptyset$.
- $|\hat{f}_{i_j}(\sigma_j)| \geq \frac{\delta}{C^D}$ *for* $1 \leq j \leq 3$, *where* C *is the constant from Theorem 9.1.1.*

Proof. Let τ, d be the constants given by Theorem 2.5.1 with parameter $\epsilon/2$, and define $\delta := \frac{\tau\epsilon}{2k}$. Then Theorem 2.5.1 implies that there exist three indices $i_1, i_2, i_3 \in [k]$ and $j \in [n]$ such that

$$\mathrm{Inf}_j^{\leq d}(f_{i_1}), \mathrm{Inf}_j^{\leq d}(f_{i_2}), \mathrm{Inf}_j^{\leq d}(f_{i_3}) \geq \tau.$$

Write $f_i = g_i + h_i$, where h_i is the *low-degree part* of the part of f_i that depends on the jth coordinate, i.e.,

$$h_i = \sum_{\substack{j \in \sigma \\ |\sigma| \leq d}} \hat{f}_i \chi_\sigma.$$

If h_{i_1}, h_{i_2}, and h_{i_3} each have some Fourier coefficient larger than δ/C^D, we are done, since every non-zero Fourier coefficient of h_i contains j, and has size at most d. Otherwise, if the largest Fourier coefficient of, say, h_{i_1} is smaller than δ/C^D, we can write

$$\langle f_1, \ldots, f_k \rangle_{\mathscr{N}} = \langle f_1, \ldots, g_{i_1}, \ldots, f_k \rangle_{\mathscr{N}} + \langle f_1, \ldots, h_{i_1}, \ldots, f_k \rangle_{\mathscr{N}}.$$

By Theorem 9.1.1, the second term is bounded by δ, and the functions in the first term have the same expected values as the f_i:s.

$$\left| \langle f_1, \ldots, g_{i_1}, \ldots, f_k \rangle_{\mathscr{N}} - \mathbb{E}[g_{i_1}] \cdot \prod_{i \neq i_1} \mathbb{E}[f_i] \right| \geq \epsilon - \delta.$$

We now repeat this argument many times. If at any point we find three large Fourier coefficients, we are done. Assume for contradiction that this does not happen. Then, the process will stop when we arrive at k functions $f_1', f_2', \ldots, f_k'$ such that

$$\left| \langle f_1', \ldots, f_k' \rangle_{\mathscr{N}} - \prod_{i=1}^{k} \mathbb{E}[f_i'] \right| < \epsilon/2.$$

Since the difference decreases by at most δ in every step, this means that it will take more than $\frac{\epsilon}{2\delta} = k/\tau$ iterations for this to happen. However, in each iteration, the ℓ_2^2 mass of one of the functions is decreased by at least τ (since there was that much influence in the removed part). Hence, after more than k/τ iterations, the total mass has decreased by more than k, which since the initial mass $\sum_{i=1}^{k} ||f_i||_2^2 \leq k$ is a contradiction. ∎

9.3 Is Low Degree Necessary?

Because of the degree restriction in Theorem 9.1.1, it is quite limited in its possible applications. Thus, one would like to extend Theorem 9.1.1 to arbitrary functions by some sort of "truncation" argument.

The standard approach to dealing with f which is not low-degree is to apply noise to it. Specifically, for $\rho \in [0,1]$, define an operator T_ρ on $L^2(\Omega^n, \mu^{\otimes n})$ by

$$T_\rho f(x) = \mathop{\mathbb{E}}_{y}[f(y)],$$

where y is distributed as follows: for each $j \in [n]$, independently, set $y_j = x_j$ with probability ρ, and y_j a uniformly random element of Ω with probability $1-\rho$. It is well known that the effect of T_ρ on the Fourier decomposition is to "dampen" the high-degree part of f. In particular, it is not hard to prove that

$$T_\rho f = \sum_{\sigma \in \mathbb{Z}_q^n} \rho^{|\sigma|} \hat{f}(\sigma) \chi_\sigma.$$

In other words, $T_\rho f$ is similar to being low-degree in the sense that the ℓ_2 norm of the high-degree part is small, $||T_\rho f^{\geq d}||_2 \leq \rho^d ||f||_2$.

It is known that if $\tilde{\rho}(\Omega_1, \ldots, \Omega_k, \mu) := \tilde{\rho} < 1$, then for small $\gamma > 0$, adding T_ρ does not change $\langle f_1, \ldots, f_k \rangle_{\mathscr{N}}$ by much. In particular, if each f_i is bounded by 1, then, by [78] Lemma 6.2, in order for

$$\left| \langle f_1, \ldots, f_k \rangle_{\mathscr{N}} - \langle T_{1-\gamma} f_1, \ldots, T_{1-\gamma} f_k \rangle_{\mathscr{N}} \right| \leq \epsilon \tag{9.1}$$

to hold it suffices to take $\gamma = \Omega((1-\tilde{\rho})\epsilon / \log 1/\epsilon)$. Hence, for μ where $\tilde{\rho} < 1$, if one can prove an analogue Theorem 9.1.1 for functions with exponentially decaying Fourier tails, one has a theorem for arbitrary bounded functions.

Unfortunately, as communicated to us by Hamed Hatami, Shachar Lovett, Alex Samorodnitsky and Julia Wolf, the direct analogue of Theorem 9.1.1 for arbitrary bounded functions is false. In particular, consider a pairwise independent distribution μ on $\{0,1\}^k$ in which the first $\approx \log k$ bits are chosen uniformly at random, and the remaining bits are sums of different subsets of the first $\log k$ bits. This distribution does not have $\tilde{\rho} < 1$, but that can be easily arranged by adding a small amount of noise to μ, which will not have any significant impact on the calculations which follow. Let $f : \{0,1\}^n \to \{0,1\}$ be the function which returns 1 on the all-zeros string, and 0 otherwise. Then, one has that

$$\langle f, \ldots, f \rangle_{\mathscr{N}} = \Pr[X_1 = \ldots = X_k = 0] \approx 2^{-n \log k},$$

whereas $\delta \leq ||f||_2 \approx 2^{-n/2}$ and hence the product $\delta \cdot \prod_{i=2}^k ||f||_2$ is bounded by

$$\delta \cdot \prod_{i=2}^{k} ||f||_2 \leq 2^{-nk/2} \ll \langle f, \ldots, f \rangle_{\mathscr{N}}.$$

One could argue that the reason that the analogue of Theorem 9.1.1 fails is that the ℓ_2 norms of $f_2, \ldots, f_k$ are the wrong things to look at. Note that the trivial bound on $\langle f_1, \ldots, f_k \rangle_{\mathscr{N}}$ obtained by simply using Hölder's inequality is $\langle f_1, \ldots, f_k \rangle_{\mathscr{N}} \leq \prod_{i=1}^{k} ||f_i||_k$. Hence, if one wants to get a statement to the effect that if f_1 has all Fourier coefficients smaller than δ then $\langle f_1, \ldots, f_k \rangle_{\mathscr{N}}$ is also small, a more reasonable type of bound would be a bound where we exchange the ℓ_k norm of f_1 by its largest Fourier coefficient, i.e.,

$$\langle f_1, \ldots, f_k \rangle_{\mathscr{N}} \leq \mathcal{O}\left(\delta \prod_{i=2}^{k} ||f_i||_k\right).$$

Note that the counterexample f used above has $||f||_k \approx 2^{-n/k}$ and thus does not constitute a counterexample for the possibility of obtaining this type of bound. We do not know whether this type of bound is possible, but let us be bold and state it as a conjecture.

Conjecture 9.3.1. *Let $(\Omega_1 \times \ldots \times \Omega_k, \mu)$ be a product space. Then for every $\gamma > 0$ and $\epsilon > 0$, there exists a constant $\delta := \delta(\gamma, \epsilon) > 0$ such that if $f_1, \ldots, f_k$ are functions $f_i \in L^2(\Omega_i^n, (\mu|_i)^{\otimes n})$ satisfying*

- *For every $i \in [k]$, $||f_i||_\infty \leq 1$.*
- *For every $d \in [n]$, $||f_i^{\geq d}||_2^2 \leq (1-\gamma)^d$.*
- *For every $\sigma \in \mathbb{Z}_q^n$, $|\hat{f}_1(\sigma)| \leq \delta$.*

Then

$$\langle f_1, \ldots, f_k \rangle_{\mathscr{N}} \leq \epsilon \prod_{i=2}^{k} ||f_i||_k.$$

Ideally, one would like a bound of the form $\delta = \epsilon / C_{\gamma,\mu}$ for some constant $C_{\gamma,\mu}$ depending only on γ and μ, but even a weaker relation between δ and ϵ would be interesting.

By Equation (9.1), it follows that Conjecture 9.3.1 could be used to deduce an analogue of Theorem 9.1.1 and analogues of Corollaries 9.2.1 and 9.2.2 for arbitrary functions, in the setting when $\tilde{\rho}(\Omega_1, \ldots, \Omega_k, \mu) < 1$.

9.4 Noisy Arithmetic Progressions

In this section, we discuss an intended application of the work in this chapter, had it been more successful.

A very deep and very powerful result in additive combinatorics, called Szemerédi's Theorem, states the following. For every positive integer k, define a function $r_k : \mathbb{N} \to \mathbb{N}$ by letting $r_k(n)$ be the size of the largest subset of $[n]$ which *does not* contain any arithmetic progressions of length k. Then, Szemerédi's Theorem [100]

states that $r_k(n) = o(n)$. Put differently, for every $\delta > 0$, there is an n_0 such that $r_k(n) \leq \delta n$ for every $n \geq n_0$. A very famous extension of this is a theorem of Green and Tao [48] which resolved an old conjecture in number theory, asserting that for every k, there exists arithmetic progressions of length k in the set of primes.[2]

In general, understanding the exact growth rate of the function $r_k(n)$ is a very challenging and important problem. The best known upper bound on $r_k(n)$ is

$$r_k(n) \leq n \cdot (\log_2 \log_2 n)^{-2^{-(k+9)}},$$

due to Gowers [44].

In many cases, understanding arithmetic progressions is easier in a setting which is closer to the material of this thesis, namely in vector spaces of the form $\mathbb{F}_q^n$ for some prime q (a sequence $x_1, \ldots, x_k \in \mathbb{F}_q^n$ is an arithmetic progression if there is some $y \in \mathbb{F}_q^n$ such that $x_{i+1} = x_i + y$ for every $1 \leq i < k$). This is generally referred to as the "finite field model". Often, results in the finite field model can be translated to results for the integers, though this translation can be quite complicated. See [45] for a survey of the finite field model and its relation to arithmetic progressions in the integers.

How do arithmetic progressions in $\mathbb{F}_q^n$ relate to what we have been doing in this chapter? One of the many ways of proving Szemerédi's Theorem is by Fourier analysis, using what is known as an *iteration argument*, or *energy argument*. This type of argument was first used by Roth [92] in a Fourier-analytic proof of Szemerédi's Theorem for the case $k = 3$ (several decades before Szemerédi's proof of the general case). Roth's proof was later extended to arbitrary k by Gowers [43, 44], who introduced the *Gowers norm* briefly mentioned in Section 2.4. Let us, somewhat informally, give an example of this type of proof (the proofs of Roth and Gowers do not follow the exact argument outlined below, but the spirit of the arguments remain the same). Let $(\mathbb{F}_q^k, \mu)$ be the probability space over $\mathbb{F}_q^k$ in which μ is the uniform distribution over all arithmetic progressions in $\mathbb{F}_q$ (including the trivial progressions in which the increment is 0). Furthermore, let $A : \mathbb{F}_q^n \to \{0, 1\}$ be the indicator function of a subset A of $\mathbb{F}_q^n$. Then, the total number of arithmetic progressions of length k contained in A is exactly given by

$$|\operatorname{Supp}(\mu)|^n \langle A, A, \ldots, A \rangle_{\mathcal{N}}.$$

Of these, q^n are the trivial progressions in which the increment in every coordinate is 0. Furthermore, it is easily checked that $|\operatorname{Supp}(\mu)| = q^2$. Hence if

$$\langle A, \ldots, A \rangle_{\mathcal{N}} > q^{-n},$$

it must be the case that A contains a non-trivial arithmetic progression. Now, suppose that one could conclude that A has a large Efron-Stein component $||A_S||_2 \geq$

[2] Which is far from being an immediate consequence of Szemerédi's Theorem since the primes are not dense enough.

δ for some $|S| \leq d$. Then, there exists a way to fix the variables $x_S \in \mathbb{F}_p^S$ such that the resulting subset $A' : \mathbb{F}_p^{[n]-S} \to \{0,1\}$ has density at least $\mathbb{E}[A'] \geq \mathbb{E}[A] + \delta$. We then repeat this many times. If the increment δ is sufficiently large compared to the number d of variables "lost" in every step, we will eventually end up with a set A'' of density larger than 1, which is of course a contradiction, implying that at some step along the way, we find a non-trivial arithmetic progression (which can then easily be lifted back to a non-trivial arithmetic progression in A).

The only problem then, is, how do we find a large Efron-Stein component A_S, with $|S|$ reasonably small? It is easy to see that, if $q \geq k$, then μ is in fact pairwise independent. Hence if A is low-degree, Corollary 9.2.1 tells us that we can find such a component. However, this works only for low-degree sets, which are not very interesting (they depend only on a few coordinates).

Now we will, rather than arithmetic progressions, consider *noisy arithmetic progressions*. A sequence $x_1, \ldots, x_k \in \mathbb{F}_q$ is a noisy arithmetic progression, if there exists a $y \in \mathbb{F}_q$ such that, for every $1 \leq i < n$ we have either $x_{i+1} = x_i + y$, or $x_{i+1} = x_i + y + 1$. In other words, we allow the increments between different steps to differ by 1. We again consider the space $(\mathbb{F}_q^k, \mu)$, this time with μ being the uniform distribution over all noisy arithmetic progressions in $\mathbb{F}_q$. This time, we have that $|\operatorname{Supp}(\mu)| = q^2(2^k - 1)$, and hence a set A contains a non-trivial noisy arithmetic progression if

$$\langle A, \ldots, A \rangle_{\mathscr{N}} > (q(2^k - 1))^{-n}.$$

Again one can verify that A is pairwise independent if $q \geq k$. More interestingly, it is also the case that $\rho(\Omega, \mu) < 1$, so that one can apply noise to any set A without significantly changing $\langle A, \ldots, A \rangle_{\mathscr{N}}$. Hence if Conjecture 9.3.1 is true one can pull off an iteration argument to prove that all sufficiently dense subsets of $\mathbb{F}_q^n$ contain noisy arithmetic sequences. If the relation between δ, γ and ϵ is "reasonable", one can hope that the value for "sufficiently dense" can be taken significantly smaller than corresponding bounds for arithmetic sequences. Furthermore, depending on the level of difficulty of a potential proof of Conjecture 9.3.1, the proof that dense subsets of $\mathbb{F}_q^n$ contain noisy arithmetic sequences might be significantly easier than the corresponding proof for true arithmetic sequences.

Part IV

Conclusions

Fram hinner du aldrig, men mycket i livet måste vara lek för att levas.

Harry Martinson – Leken

Chapter 10

Some Open Problems

Rather than further discussing the results to which the past 2^7 pages were devoted, let us in this concluding chapter look into the future, and at some of the many problems that remain to be solved. Here, I will list some open problems related to this thesis which I find particularly interesting. Some of these I have worked on and failed to solve, some of them I want to work on (and hopefully not fail).

Unique Games on the Boolean Hypercube How hard is it to solve Unique Games when the constraint graph is the boolean hypercube? How hard is it in the case when the label set is $\{0,1\}^l$ and every constraint is linear over the group $\mathbb{F}_2^l$, i.e., when every constraint is of the form $\pi(\ell(x)) = \ell(y) \oplus T$ for some $T \in \{0,1\}^l$, where $\oplus$ denotes bit-wise Xor? We are not aware of any better algorithm than the general algorithm by Charikar et al. [20].

Hardness Without Unique Games Can any Unique Games-based hardness result be proved under a weaker assumption, such as the d-to-1 conjecture [62], or even just $\mathsf{P} \neq \mathsf{NP}$?

k-ary Unique Games Consider the L-Unique Label Cover problem on a k-uniform hypergraph, where each constraint now asserts that $\pi_1(\ell(v_1)) = \pi_2(\ell(v_2)) = \ldots = \pi_k(\ell(v_k))$ for some permutations $\pi_1, \ldots, \pi_k$ on $[L]$ and vertices $v_1, \ldots, v_k$. How hard is this problem? It is known that for constant k, $(\gamma, 1-\gamma)$-hardness is equivalent to the UGC [65, 9], whereas for large $k = \Omega(n)$, the problem is in P. Is $(\gamma, 1-\gamma)$-hardness still equivalent to the UGC for, say, $k = \log\log n$?

Characterize Hereditary Approximation Resistance Is the sufficient condition for hereditary resistance from Chapter 5 also necessary? I.e., does pairwise independence completely characterize hereditarily resistant predicates?

The $\alpha = \beta$ Conjecture Prove or disprove Conjecture 6.1.3.

The Approximability of Max Di-Cut How hard is it to approximate the MAX DI-CUT problem? The current best algorithm is the 0.87401-approximation for MAX 2-AND [71], and the current best hardness is the $\alpha_{GW} \approx 0.87856$-UG-hardness for balanced MAX 2-AND [63]. In particular, is MAX DI-CUT harder to approximate than MAX CUT? Is it as hard as MAX 2-AND? We believe that the answer to the first question is yes, and that the methods of Chapter 6 can be used to prove it, though we have so far not succeeded in our efforts of proving this.

The Approximability of Graph Coloring What is the minimum number of colors $c(n)$ such that a 3-colorable graph can be colored using $c(n)$ colors in polynomial time? Is it the case that for every $\delta > 0$, $c(n) \leq \mathcal{O}(n^{\delta})$ (the current record is $\mathcal{O}(n^{0.2072})$ [23])? Is it the case that $c(n) > \log\log n$ (the current record is $c(n) = \omega(1)$ under a certain variant of the Unique Games Conjecture [28])?

The Approximability of TSP What is the best possible approximation ratio for TSP with triangle inequalities? For asymmetric TSP, the gap is huge, with the best current approximation algorithm having ratio $\frac{2}{3}\log_2 n$ [37], and the best current inapproximability being $117/116 - \epsilon$ [86].

The Threshold for Randomly Supported Independence How many random points from Ω^n are needed in order to be able to support a k-wise independent distribution with high probability? The results of Chapter 8 show that $(\text{poly}(q)n)^k \log(n^k)$ points suffice, and that $\Omega\left(\frac{n^k}{q^{k^2}k^k}\right)$ points are necessary.

1. Can the $\log(n^k)$ factor in the upper bound be removed? We believe the answer is yes.
2. Assuming the threshold is $c(k,q) \cdot n^k$, how does $c(k,q)$ behave for large k (say, $k = \log n$). Does it tend to 0 or infinity?

Noise Correlation Bounds for Uniform Functions Is Conjecture 9.3.1 true? If not, is something similar true?

Bibliography

[1] Farid Alizadeh. Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM Journal on Optimization*, 5:13–51, 1995. [66]

[2] Noga Alon, László Babai, and Alon Itai. A fast and simple randomized parallel algorithm for the maximal independent set problem. *Journal of Algorithms*, 7(4):567–583, 1986. ISSN 0196-6774. [15]

[3] Sanjeev Arora, Eden Chlamtac, and Moses Charikar. New Approximation Guarantee for Chromatic Number. In *ACM Symposium on Theory of Computing (STOC)*, pages 205–214, 2006. ISBN 1-59593-134-1. [61]

[4] Sanjeev Arora, Subhash Khot, Alexandra Kolla, David Steurer, Madhur Tulsiani, and Nisheeth K. Vishnoi. Unique games on expanding constraint graphs are easy. In *ACM Symposium on Theory of Computing (STOC)*, pages 21–28, 2008. [34]

[5] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof Verification and the Hardness of Approximation Problems. *Journal of the ACM*, 45(3):501–555, 1998. [31]

[6] Sanjeev Arora and Shmuel Safra. Probabilistic Checking of Proofs: A New Characterization of NP. *Journal of the ACM*, 45(1):70–122, 1998. [31]

[7] Per Austrin. Balanced Max 2-Sat Might Not be the Hardest. In *ACM Symposium on Theory of Computing (STOC)*, pages 189–197, 2007. [7, 33]

[8] Per Austrin. Towards Sharp Inapproximability For Any 2-CSP. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 307–317, 2007. [7, 22, 23, 33, 88]

[9] Per Austrin and Elchanan Mossel. Approximation Resistant Predicates From Pairwise Independence. In *IEEE Conference on Computational Complexity (CCC)*, 2008. [7, 21, 33, 58, 129]

[10] R. C. Baker, G. Harman, and J. Pintz. The Difference Between Consecutive Primes, II. *Proceedings of the London Mathematical Society*, 83(3):532–562, 2001. [37]

[11] William Beckner. Inequalities in Fourier analysis. *Annals of Mathematics*, 102(1):159–182, 1975. [95]

[12] Mihir Bellare, Don Coppersmith, Johan Håstad, Marcos A. Kiwi, and Madhu Sudan. Linearity testing in characteristic two. *IEEE Transactions on Information Theory*, 42(6):1781–1795, 1996. [46]

[13] Mihir Bellare, Oded Goldreich, and Madhu Sudan. Free Bits, PCPs, and Nonapproximability–Towards Tight Results. *SIAM Journal on Computing*, 27(3):804–915, 1998. [42]

[14] Itai Benjamini, Gil Kalai, and Oded Schramm. Noise sensitivity of boolean functions and applications to percolation. *Inst. Hautes Études Sci. Publ. Math*, 90:5–43, 1999. [21]

[15] Avrim Blum and David Karger. An $\tilde{\mathcal{O}}(n^{3/14})$-coloring algorithm for 3-colorable graphs. *Information Processing Letters*, 61(1):49–53, 1997. [61]

[16] Manuel Blum, Michael Luby, and Ronitt Rubinfeld. Self-testing/correcting with applications to numerical problems. *Journal of Computer Systems and Sciences*, 47(3):549–595, 1993. [46]

[17] Aline Bonami. Ensembles $\Gamma(p)$ dans le dual de D^∞. *Ann. Inst. Fourier*, 18 (2):193–204, 1968. [95]

[18] Aline Bonami. Étude des coefficients de Fourier des fonctions de $L^p(G)$. *Ann. Inst. Fourier*, 20:335–402, 1970. [95]

[19] Andreas Brieden, Peter Gritzmann, Ravi Kannan, Victor Klee, László Lovász, and Miklós Simonovits. Approximation of diameters: randomization doesn't help. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 244–251, 1998. [100]

[20] Moses Charikar, Konstantin Makarychev, and Yury Makarychev. Near-optimal algorithms for unique games. In *ACM Symposium on Theory of Computing (STOC)*, pages 205–214, 2006. ISBN 1-59593-134-1. [34, 129]

[21] Moses Charikar, Konstantin Makarychev, and Yury Makarychev. Near-Optimal Algorithms for Maximum Constraint Satisfaction Problems. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 62–68, 2007. [51]

[22] Moses Charikar and Anthony Wirth. Maximizing Quadratic Programs: Extending Grothendieck's Inequality. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 54–60, 2004. [30, 61]

[23] Eden Chlamtac. Approximation Algorithms Using Hierarchies of Semidefinite Programming Relaxations. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 691–701, 2007. [130]

[24] Eden Chlamtac and Gyanit Singh. Improved Approximation Guarantees through Higher Levels of SDP Hierarchies. In *APPROX-RANDOM*, pages 49–62, 2008. [65]

[25] Nicos Christofides. Worst-case analysis of a new heuristic for the travelling salesman problem. Technical report, Report 388, Graduate School of Industrial Administration, CMU, 1976. [29]

[26] Irit Dinur, Ehud Friedgut, Guy Kindler, and Ryan O'Donnell. On the Fourier tails of bounded functions over the discrete cube. In *ACM Symposium on Theory of Computing (STOC)*, pages 437–446, 2006. [96, 98, 104]

[27] Irit Dinur, Ehud Friedgut, and Oded Regev. Independent sets in graph powers are almost contained in juntas. *Geometric and Functional Analysis*, 18(1): 77–97, 2008. [22]

[28] Irit Dinur, Elchanan Mossel, and Oded Regev. Conditional hardness for approximate coloring. In *ACM Symposium on Theory of Computing (STOC)*, pages 344–353, 2006. [22, 33, 130]

[29] Bradley Efron and Charles Stein. The Jackknife Estimate of Variance. *Annals of Statistics*, 9:586–596, 1981. [18]

[30] Lars Engebretsen. The nonapproximability of non-boolean predicates. *SIAM Journal on Discrete Mathematics*, 18(1):114–129, 2004. [51]

[31] Lars Engebretsen and Jonas Holmerin. More Efficient Queries in PCPs for NP and Improved Approximation Hardness of Maximum CSP. In *Symposium on Theoretical Aspects of Computer Science (STACS)*, pages 194–205, 2005. [51]

[32] Uriel Feige. Approximating Maximum Clique by Removing Subgraphs. *SIAM J. Discrete Math.*, 18(2):219–225, 2004. [30]

[33] Uriel Feige and Michel Goemans. Aproximating the Value of Two Prover Proof Systems, With Applications to MAX 2SAT and MAX DICUT. In *Israel Symposium on Theory of Computing Systems (ISTCS)*, pages 182–189, 1995. ISBN 0-8186-6915-2. [61, 68, 72]

[34] Uriel Feige, Guy Kindler, and Ryan O'Donnell. Understanding Parallel Repetition Requires Understanding Foams. In *IEEE Conference on Computational Complexity (CCC)*, pages 179–192, 2007. [35]

[35] Uriel Feige and Michael Langberg. The RPR^2 rounding technique for semidefinite programs. *Journal of Algorithms*, 60(1):1–23, 2006. [68, 90]

[36] Uriel Feige and Daniel Reichman. On Systems of Linear Equations with Two Variables per Equation. In *APPROX-RANDOM*, pages 117–127, 2004. [34]

[37] Uriel Feige and Mohit Singh. Improved approximation ratios for traveling salesperson tours and paths in directed graphs. Manuscript, 2006. [130]

[38] Alan M. Frieze and Mark Jerrum. Improved Approximation Algorithms for MAX k-CUT and MAX BISECTION. *Algorithmica*, 18(1):67–81, 1997. [61]

[39] Zoltán Füredi. Random Polytopes in the d-Dimensional Cube. *Discrete Comput. Geom.*, 1:315–319, 1986. [102]

[40] Christophe Garban, Gábor Pete, and Oded Schramm. The Fourier Spectrum of Critical Percolation. arXiv Report math.PR/0803.3750, 2008. [21]

[41] Konstantinos Georgiou, Avner Magen, Toniann Pitassi, and Iannis Tourlakis. Integrality gaps of $2-o(1)$ for Vertex Cover SDPs in the Lovász-Schrijver Hierarchy. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 702–712, 2007. [65]

[42] Michel X. Goemans and David P. Williamson. Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming. *Journal of the ACM*, 42:1115–1145, 1995. [30, 50, 52, 61, 65, 66, 68]

[43] Tim Gowers. A new proof of Szemerédi's theorem for arithmetic progressions of length four. *Geometric and Functional Analysis*, 8:529–551, 1998. [124]

[44] Tim Gowers. A new proof of Szemerédi's theorem. *Geometric and Functional Analysis*, 11:465–588, 2001. [21, 124]

[45] Ben Green. Finite field models in additive combinatorics. In *Surveys in Combinatorics*, pages 1–27, 2005. [124]

[46] Ben Green and Terence Tao. An inverse theorem for the Gowers $U^3(G)$ norm. arXiv Report math.NT/0503014v3, to appear in Proc. Edin. Math. Soc., 2006. [22]

[47] Ben Green and Terence Tao. The distribution of polynomials over finite fields, with applications to the Gowers norms. arXiv Report math.CO/0711.3191, 2007. [22]

[48] Ben Green and Terence Tao. The primes contain arbitrarily long arithmetic progressions. *Annals of Mathematics*, 167:481–547, 2008. [124]

[49] Leonard Gross. Logarithmic Sobolev Inequalities. *American Journal of Mathematics*, 97:1061–1083, 1975. [96]

[50] Jacques Hadamard. Résolution d'une question relative aux déterminants. *Bulletin des sciences math.*, 2(17):240–248, 1893. [37]

[51] Eran Halperin, Ram Nathaniel, and Uri Zwick. Coloring k-colorable graphs using relatively small palettes. *Journal of Algorithms*, 45(1):72–90, 2002. [61]

[52] Gustav Hast. Approximating Max kCSP – Outperforming a Random Assignment with Almost a Linear Factor. In *International Colloquium on Automata, Languages and Programming (ICALP)*, pages 956–968, 2005. [51]

[53] Gustav Hast. *Beating a Random Assignment – Approximating Constraint Satisfaction Problems*. PhD thesis, KTH – Royal Institute of Technology, 2005. [52, 57, 59]

[54] Johan Håstad. Some optimal inapproximability results. *Journal of the ACM*, 48(4):798–859, 2001. [47, 48, 51, 52, 62]

[55] Johan Håstad. Every 2-CSP Allows Nontrivial Approximation. In *ACM Symposium on Theory of Computation (STOC)*, pages 740–746, 2005. [52]

[56] Johan Håstad. On the approximation resistance of a random predicate. In *APPROX-RANDOM*, pages 149–163, 2007. [33, 52, 58, 59]

[57] Thomas Holenstein. Parallel repetition: simplifications and the no-signaling case. In *ACM Symposium on Theory of Computing (STOC)*, pages 411–419, 2007. [35]

[58] Svante Janson. *Gaussian Hilbert Spaces*. Cambridge University Press, 1997. [97]

[59] Anatole Joffe. On a Set of Almost Deterministic k-Independent Random Variables. *Annals of Probability*, 2(1):161–162, 1974. [15]

[60] Jeff Kahn, Gil Kalai, and Nathan Linial. The Influence of Variables on Boolean Functions. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 68–80, 1988. [95]

[61] David R. Karger, Rajeev Motwani, and Madhu Sudan. Approximate graph coloring by semidefinite programming. *Journal of the ACM*, 45(2):246–265, 1998. [61]

[62] Subhash Khot. On the power of unique 2-prover 1-round games. In *ACM Symposium on Theory of Computing (STOC)*, pages 767–775, 2002. ISBN 1-58113-495-9. [34, 129]

[63] Subhash Khot, Guy Kindler, Elchanan Mossel, and Ryan O'Donnell. Optimal inapproximability results for max-cut and other 2-variable csps? *Siam Journal on Computing*, 37:319–357, 2007. [21, 22, 33, 34, 35, 62, 63, 64, 77, 130]

[64] Subhash Khot and Ryan O'Donnell. SDP gaps and UGC-hardness for MAXCUTGAIN. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 217–226, 2006. [30, 33]

[65] Subhash Khot and Oded Regev. Vertex cover might be hard to approximate to within $2-\epsilon$. *Journal of Computer and System Sciences*, 74(3):335–349, 2008. [33, 129]

[66] Subhash Khot and Nisheeth K. Vishnoi. The Unique Games Conjecture, Integrality Gap for Cut Problems and Embeddability of Negative Type Metrics into ℓ_1. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 53–62, 2005. [33]

[67] Guy Kindler, Assaf Naor, and Gideon Schechtman. The UGC hardness threshold of the ℓ_p Grothendieck problem. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 64–73, 2008. [33]

[68] M. Kochol. Constructive approximation of a ball by polytopes. *Math. Slovaca*, 44(1):99–105, 1994. [100]

[69] Henry Oliver Lancaster. Pairwise Statistical Independence. *Annals of Mathematical Statistics*, 36(4):1313–1317, 1965. [15]

[70] Jean B. Lasserre. An Explicit Exact SDP Relaxation for Nonlinear 0-1 Programs. In *Integer Programming & Combinatorial Optimization (IPCO)*, pages 293–303, 2001. [65]

[71] Michael Lewin, Dror Livnat, and Uri Zwick. Improved rounding techniques for the MAX 2-SAT and MAX DI-CUT problems. In *Integer Programming & Combinatorial Optimization (IPCO)*, volume 2337 of *Lecture Notes in Computer Science*, pages 67–82, 2002. [61, 68, 69, 77, 79, 84, 85, 130]

[72] Shachar Lovett, Roy Meshulam, and Alex Samorodnitsky. Inverse conjecture for the gowers norm is false. In *ACM Symposium on Theory of Computing (STOC)*, pages 547–556, 2008. [22]

[73] László Lovász and Alexander Schrijver. Cones of Matrices and Setfunctions, and 0-1 Optimization. *SIAM Journal on Optimization*, 1(2):166–190, 1991. [65]

[74] Michael Luby and Avi Wigderson. Pairwise Independence and Derandomization. *Foundation and Trends in Theoretical Computer Science*, 1(4):237–301, 2005. [15]

[75] Rajsekar Manokaran, Joseph Naor, Prasad Raghavendra, and Roy Schwartz. SDP gaps and UGC Hardness for Multiway Cut, 0-Extension, and Metric Labeling. In *ACM Symposium on Theory of Computing (STOC)*, pages 11–20, 2008. [33]

[76] Shiro Matuura and Tomomi Matsui. 0.863-Approximation Algorithm for MAX DICUT. In *RANDOM-APPROX*, pages 138–146, 2001. [61]

[77] Shiro Matuura and Tomomi Matsui. 0.935-Approximation Randomized Algorithm for MAX 2SAT and Its Derandomization. Technical Report METR 2001-03, Department of Mathematical Engineering and Information Physics, the University of Tokyo, Japan, 2001. [61]

[78] Elchanan Mossel. Gaussian bounds for noise correlation of functions. arXiv Report math/0703683v3, 2007. [13, 18, 21, 22, 23, 24, 122]

[79] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. To appear in Annals of Mathematics, 2008. [21, 22, 23, 78, 97]

[80] Edgar Nelson. Construction of quantum fields from Markoff fields. *Journal of Functional Analysis*, 12:97–112, 1973. [96]

[81] G. L. O'Brien. Pairwise Independent Random Variables. *Annals of Probability*, 8(1):170–175, 1980. [15, 36]

[82] Ryan O'Donnell. *Computational applications of noise sensitivity*. PhD thesis, Massachusetts Institute of Technology, 2003. [21]

[83] Ryan O'Donnell and Yi Wu. An optimal SDP algorithm for Max-Cut, and equally optimal Long Code tests. In *ACM Symposium on Theory of Computing (STOC)*, pages 335–344, 2008. [30, 33, 66, 68, 90]

[84] Raymond E. A. C. Paley. On orthogonal matrices. *Journal of Mathematics and Physics*, 12:311–320, 1933. [37]

[85] Christos H. Papadimitriou and Kenneth Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Dover Publications, Inc., 1998. [27]

[86] Christos H. Papadimitriou and Santosh Vempala. On the Approximability of the Traveling Salesman Problem. `http://www.cs.berkeley.edu/~christos/tsp.ps`. [130]

[87] Prasad Raghavendra. Optimal Algorithms and Inapproximability Results For Every CSP? In *ACM Symposium on Theory of Computing (STOC)*, 2008. [21, 22, 30, 33, 62, 66, 70, 90, 91]

[88] Anup Rao. Parallel Repetition in Projection Games and a Concentration Bound. In *ACM Symposium on Theory of Computing (STOC)*, 2008. [35]

[89] Ran Raz. A Parallel Repetition Theorem. *SIAM Journal on Computing*, 27 (3):763–803, 1998. [33, 34]

[90] Ran Raz. A Counterexample to Strong Parallel Repetition. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 369–373, 2008. [35]

[91] Vladimir I. Rotar. Limit theorems for polylinear forms. *J. Multivariate Anal.*, 9(4):511–530, 1979. [24]

[92] Klaus F. Roth. On certain sets of integers. *J. London Math. Soc.*, 28:245–252, 1953. [124]

[93] Alex Samorodnitsky and Luca Trevisan. A PCP characterization of NP with optimal amortized query complexity. In *ACM Symposium on Theory of Computing (STOC)*, pages 191–199, 2000. [48, 51]

[94] Alex Samorodnitsky and Luca Trevisan. Gowers uniformity, influence of variables, and PCPs. In *ACM Symposium on Theory of Computing (STOC)*, pages 11–20, 2006. ISBN 1-59593-134-1. [33, 51, 57, 58]

[95] Grant Schoenebeck. Linear Level Lasserre Lower Bounds for Certain k-CSPs. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 593–602, 2008. [34, 65]

[96] Grant Schoenebeck, Luca Trevisan, and Madhur Tulsiani. Tight integrality gaps for Lovasz-Schrijver LP relaxations of vertex cover and max cut. In *ACM Symposium on Theory of Computing (STOC)*, pages 302–310, 2007. [65]

[97] Oded Schramm and Jeffrey E. Steif. Quantitative noise sensitivity and exceptional times for percolation. To appear in Annals of Mathematics, 2007. [21]

[98] Hanif D. Sherali and Warren P. Adams. A Hierarchy of Relaxations Between the Continuous and Convex Hull Representations for Zero-One Programming Problems. *SIAM Journal on Discrete Mathematics*, 3(3):411–430, 1990. [65]

[99] J. Sylvester. Thoughts on Orthogonal Matrices, Simultaneous Sign-Successions, and Tessellated Pavements in Two or More Colours, with Applications to Newton's Rule, Ornamental Tile-Work, and the Theory of Numbers. *Philosophical Magazine*, (34):461–475, 1867. [37]

[100] Endre Szemerédi. On sets of integers containing no k elements in arithmetic progression. *Acta Arith.*, 27:299–345, 1975. [21, 123]

[101] Luca Trevisan. Parallel Approximation Algorithms by Positive Linear Programming. *Algorithmica*, 21:72–88, 1998. [51, 64]

[102] Luca Trevisan. Approximation algorithms for unique games. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 197–205, 2005. [34]

[103] Luca Trevisan, Gregory B. Sorkin, Madhu Sudan, and David P. Williamson. Gadgets, Approximation, and Linear Programming. *SIAM Journal on Computing*, 29(6):2074–2097, 2000. [62]

[104] Avi Wigderson. P, NP and Mathematics - a computational complexity perspective. *International Congress of Mathematicians (ICM 2006)*, pages 665–712, 2007. [5]

[105] Pawel Wolff. Hypercontractivity of Simple Random Variables. *Studia Math.*, 180:219–236, 2007. [96]

[106] Uri Zwick. Approximation Algorithms for Constraint Satisfaction Problems Involving at Most Three Variables Per Constraint. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 1998. [52]

[107] Uri Zwick. Analyzing the MAX 2-SAT and MAX DI-CUT approximation algorithms of Feige and Goemans. Manuscript, 2000. [88]

[108] Uri Zwick. Personal communication, 2005. [79, 84, 85]