The Dissertation Committee for Raghu Vardhan Reddy Meka
certifies that this is the approved version of the following dissertation:

Computational Applications of Invariance Principles

Committee:

David Zuckerman, Supervisor

Inderjit Dhillon

Anna Gal

Parikshit Gopalan

Adam Klivans

Computational Applications of Invariance Principles

by

Raghu Vardhan Reddy Meka, B.Tech.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2011

Dedicated to my parents.

Acknowledgments

First I would like to thank my adviser David Zuckerman - without his guidance this thesis would not have taken shape. Besides his great research advise and the many fruitful discussions we had, David's continuous encouragement and goading were important in making me not lose heart during lean patches. David has a clear and holistic view of research and life in general and I hope some of that has rubbed on to me. I am lucky to have had him as my adviser.

Besides David, another big factor in the making of this thesis was the influence of Adam Klivans. I benefited greatly from our long research meetings and discussions and I thank Adam for being a great collaborator as well as a friend. From him I have learnt a lot of things and both absorbed and sifted through a lot of opinions.

I would like to thank Parikshit Gopalan and Omer Reingold for hosting me at Microsoft Research, Silicon Valley, for two fun-filled and productive summers and patiently hearing to all the bad ideas I kept having. Besides learning a lot of technical stuff from them, I hope to have absorbed some of their (different) attitudes to research. I would like to thank Prahladh Harsha for being a good friend and collaborator - it is not a coincidence that much of the work in this thesis happened in the summer he was visiting Austin.

I am grateful to Inderjit Dhillon and Anna Gal for serving on my committee and providing helpful comments and suggestions.

I would like to thank the core group of friends who made my stay in Austin as enjoyable as it was and made the six years just a breeze - Shweta Agarwal, Siddhartha Banerjee, Aditya Gopalan, Prateek Jain, Sivaram Kalyanakrishnan, Sunil Kowlgi, Keerti Lakshminarayanan, Prince Mahajan, Vishwas Srinivasan, Rangarajan Vasudevan and others.

Last but the most important of all, I would like to thank my parents without whose encouragement and hardship I would not be here. I would like to thank my sister and brother-in-law for their constant encouragement and my nephew and niece for their constant entertainment.

Computational Applications of Invariance Principles

Publication No. _____

Raghu Vardhan Reddy Meka, Ph.D.
The University of Texas at Austin, 2011

Supervisor: David Zuckerman

This thesis focuses on applications of classical tools from probability theory and convex analysis such as limit theorems to problems in theoretical computer science, specifically to pseudorandomness and learning theory.

At first look, limit theorems, pseudorandomness and learning theory appear to be disparate subjects. However, as it has now become apparent, there's a strong connection between these questions through a third more abstract question: what do *random* objects look like. This connection is best illustrated by the study of the spectrum of Boolean functions which directly or indirectly played an important role in a plethora of results in complexity theory. The current thesis aims to take this program further by drawing on a variety of fundamental tools, both classical and new, in probability theory and analytic geometry. Our research contributions broadly fall into three categories.

Probability Theory: The central limit theorem is one of the most important results in all of probability and richly studied topic. Motivated by questions

in pseudorandomness and learning theory we obtain two new limit theorems or invariance principles. The proofs of these new results in probability, of interest on their own, have a computer science flavor and fall under the niche category of techniques from theoretical computer science with applications in pure mathematics.

Pseudorandomness: Derandomizing natural complexity classes is a fundamental problem in complexity theory, with several applications outside complexity theory. Our work addresses such derandomization questions for natural and basic geometric concept classes such as halfspaces, polynomial threshold functions (PTFs) and polytopes. We develop a reasonably generic framework for obtaining pseudorandom generators (PRGs) from invariance principles and suitably apply the framework to old and new invariance principles to obtain the best known PRGs for these complexity classes.

Learning Theory: Learning theory aims to understand what functions can be learned efficiently from examples. As developed in the seminal work of Linial, Mansour and Nisan (1994) and strengthened by several follow-up works, we now know strong connections between learning a class of functions and how sensitive to noise, as quantified by average sensitivity and noise sensitivity, the functions are. Besides their applications in learning, bounding the average and noise sensitivity has applications in hardness of approximation, voting theory, quantum computing and more. Here we address the question of bounding the sensitivity of polynomial threshold functions and intersections of halfspaces and obtain the best known results for these concept classes.

Table of Contents

Chapter 1

Introduction

An important theme in theoretical computer science over the last decade has been the usefulness of translating a combinatorial problem over a discrete domain to a problem in continuous space. The notion of convex relaxation, for example, is now a standard approach in combinatorial optimization. More recently, understanding the behavior of Boolean functions in the Gaussian space has played a crucial role in the recent breakthroughs in hardness of approximation [57], [76], [88] and social choice theory [76]. A core concept in these results is an "invariance principle" or "limit theorem" that relates and helps reduce problems over a discrete domain (typically, the hypercube) to the continuous domain (typically, $\mathbb{R}^n$ equipped with the Gaussian measure), which are often more tractable. The present work develops this high-level approach further by applying invariance principles to two basic questions in computer science:

1. Is randomness necessary for efficient computing?

2. What can we learn efficiently?

We first give an overview of the three broad topics we concern ourselves with - invariance principles, pseudorandomness and learning theory, highlight-

ing the main questions relevant to us.

Invariance Principles The classical central limit theorem says that for any sequence $X_1, \ldots, X_n$ of independent and identical random variables with finite mean μ and variance σ^2, the random variable $(X_1 + \cdots + X_n)/\sqrt{n}$ approaches the standard Gaussian distribution with mean μ and variance σ^2 in distribution. In particular, in the limit as $n \to \infty$, the distribution of $(X_1 + \cdots + X_n)/\sqrt{n}$ is *invariant* of the particular choice of the random variables $X_1, \ldots, X_n$ apart from their mean and variance.

More generally, given a collection of functions $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}$, and a collection of distributions $\mathcal{D}$ over $\mathbb{R}^n$, we will be interested in statements of the form $\mathbb{E}[f(X)] \sim \mathbb{E}[f(Y)]$, where X, Y are random variables over $\mathbb{R}^n$ with distributions in $\mathcal{D}$ and $f \in \mathcal{F}$. For instance, the classical Berry-Esséen theorem (Theorem 2.2.4) applies to the case where $\mathcal{F}$ is the class of *regular* halfspaces and $\mathcal{D}$ is the set of all *reasonable* product distributions over $\mathbb{R}^n$.

Given the above setup, it is natural to ask for what classes $\mathcal{F}$ and $\mathcal{D}$ can we have invariance principles as above and with what asymptotic error.

Pseudorandomness The use of randomness is fundamental in all of computer science and provably so for distributed computing, communication complexity and more. In spite of the ubiquity of randomness and randomized algorithms in particular, it is still unknown if the use of randomness is necessary for the design of efficient algorithms. One of the foremost open problems in

computer science is whether the complexity class BPP (the class of languages with efficient randomized algorithms) is the same as P (the class of languages with efficient algorithms). By and large, evidence suggests that indeed BPP = P.

A natural approach to the BPP vs P question is to ask for efficient pseudorandom generators (PRGs) that would "fool" polynomial time computable functions. Unfortunately, the task of constructing unconditional PRGs for all polynomial time computable functions is out of reach of current techniques. As a result, much attention has been given to the problem of constructing PRGs for simpler classes of functions. Besides being of importance on their own, explicit PRGs for simple complexity classes have found applications far beyond the goal of fooling the corresponding classes of functions.

For instance, the seminal work of Nisan [80] constructing PRGs with seed length $O(\log^2 n)$ for the class of small space machines has found applications in constructions of extractors, space efficient streaming algorithms and so on. Another prominent example is the work of Naor and Naor [78], Alon et al. [4] on explicit constructions of small-bias spaces (PRGs that fool linear functions modulo primes) that has found applications in error-correcting codes, PCP constructions and more.

The above remarks motivate the task of obtaining explicit PRGs for natural, simpler complexity classes. In this context, the present work addresses the question of designing pseudorandom generators for basic *geometric* concept classes such as halfspaces, polynomial threshold functions and polytopes.

Computational Learning Theory Learning theory aims to understand what functions can be learned efficiently from examples. Since its formalization in the seminal work of Valiant, learning theory has greatly impacted both the practical design of algorithms for real-world learning tasks, as well as the theoretical understanding of properties of functions that make them learnable.

Often, learning algorithms exploit a deep structural property of the function being learned. One such property is the Fourier spectrum of the function. Pioneered by the work of Linial, Mansour and Nisan [65], several prominent learning theory works exploit this concrete connection via two fundamental properties of Boolean functions: *average sensitivity* and *noise sensitivity*. Roughly speaking, the average sensitivity and noise sensitivity quantify the noise tolerance of functions. Besides, their applications in learning theory, understanding average sensitivity and noise sensitivity has applications in hardness of approximation, social choice theory, circuit complexity and quantum complexity.

In this context, the present work studies the sensitivity of the classes of polynomial threshold functions and intersections of halfspaces, leading to efficient algorithms for learning these classes even in the presence of noise.

1.1 Our Results

1.1.1 Invariance Principles

Viewed geometrically, the central limit theorem or more precisely its quantitative version, the Berry-Esséen theorem says the following. For a hy-

perplane in $\mathbb{R}^n$, and any product distribution over $\mathbb{R}^n$, the probability that a random point from the distribution lies on one side of the hyperplane is *invariant* in the sense of being approximately the same for all "reasonable" product distributions.

Since its first appearance in the 1940's, there have been many extensions of the Berry-Esséen theorem in the probability community. Motivated by their applications in pseudorandomness and learning theory, we obtain two generalizations of the classical Berry-Esséen theorem.

An Invariance Principle for Polytopes One class of powerful extensions of the Berry-Esséen theorem is the case where there are several hyperplanes that define a polytope, and we are interested in the probability that a random vector in $\mathbb{R}^n$ lies inside the polytope. In work with Harsha and Klivans [42], we show an invariance principle for polytopes analogous to the Berry-Esséen theorem for a single hyperplane. The novelty of our result is that the final error bound is only poly-logarithmic in the number of bounding hyperplanes (or faces) of the polytope. Previous results had at least a linear dependence on the number of faces of the polytope.

Discrete Central Limit Theorems Another class of well studied generalizations of the classical central limit theorem are the *discrete central limit theorems* [8, 7]. These results show, for instance, that a sum of independent indicator-valued random variables (or more generally, integer-valued random

variables) converges to the appropriate binomial distribution in total variation or statistical distance. In contrast, the Berry-Esséen theorem only shows convergence in the Kolmogorov-Smirnov (or cdf) distance. Most previous results from the probability community use Fourier techniques or Stein's method and are somewhat complicated. In work with Gopalan, Reingold and Zuckerman [36], we use tools developed for constructing pseudorandom generators to give a different proof of the discrete central limit theorem. Our proof relies only on the classical Berry-Esséen theorem and some elementary properties of the binomial distribution.

1.1.2 Pseudorandomness

The starting point for our results for questions related to pseudorandomness is a framework for obtaining PRGs from invariance principles for geometric concept classes. We then carefully develop this framework along with appropriate invariance principles to obtain pseudorandom generators for polynomial threshold functions, halfspaces and polytopes.

PRGs for Threshold Functions Polynomial threshold functions (PTFs) are a well studied class of functions with many applications in complexity theory [13], learning theory [61], quantum computing [11] and more. The case of degree one threshold functions, or halfspaces, have been instrumental in the development of standard learning theory tools such as perceptrons, support vector machines and boosting.

In joint work with Zuckerman, we address the natural question of constructing PRGs for PTFs and obtain the first nontrivial result for the problem. We give an explicit PRG with a seed-length of $\log n/\varepsilon^{O(d)}$ that fools degree d PTFs with error at most ε. Previously, no constructions with seed-length $o(n)$ were known even for degree two threshold functions. We also achieve substantial improvements for halfspaces obtaining a seed-length of $O(\log n + \log^2(1/\varepsilon))$. This was the first result to obtain a logarithmic dependence on the error rate ε for halfspaces.

PRGs for Polytopes Understanding the structure of integer points in polytopes (that is, solutions to integer programs) is a fundamental topic in computer science. An important problem in this area is to estimate the fraction of points on the hypercube that lie inside a given polytope. In joint work with Harsha and Klivans [42], we addressed the question of estimating the *Boolean volume* of polytopes in an oblivious manner. That is, we wanted to find a small explicit set S of points on the hypercube such that for every polytope, the fraction of points from S that lie inside the polytope is close to the fraction of points on the hypercube that lie inside the polytope.

Building on the invariance principle for polytopes, we give the first construction of an explicit set of quasi-polynomial size which preserves the Boolean volume up to an additive error ε for a broad-class of *regular* polytopes. Our notion of *regular* polytopes captures several common integer programs that arise in optimization such as dense covering problems, contingency tables.

Previous constructions had at least an exponential dependence on the number of faces of the polytope.

PRGs for Combinatorial Shapes Some of the most influential results in pseudorandomness are the PRGs for space-bounded computations. In particular, the PRGs of Nisan [80] and Impagliazzo, Nisan, and Wigderson [46] use a seed of length $O(\log^2 n)$ to fool polynomial-width branching programs. Despite much effort, these constructions have not been improved in nearly two decades. However, logarithmic-seed PRGs for weaker classes of distinguishers have been previously constructed and found many applications. In joint work with Gopalan, Reingold and Zuckerman [36], we define a natural common generalization and significant extension of many of these distinguisher classes such as small-bias spaces [78], combinatorial rectangles [32, 6, 68], 0/1 halfspaces, 0/1 modular sums [73, 67]. We name the new class *combinatorial shapes*.

Combinatorial shapes look at their inputs in consecutive chunks of $\log m$ bits (m is typically polynomial in n). On each chunk of bits the combinatorial shape may apply an arbitrary Boolean function. Nevertheless, these Boolean functions are combined into a single output by a symmetric (i.e., order independent) function. Our main result is a construction of PRGs with seed length $O(\log m + \log n + \log^2(1/\varepsilon))$ that fools combinatorial shapes with error at most ε, where m denotes the size of the universe and n the number of dimensions.

Our construction is interesting in its own right even for the special case of $m = 2$: we give the first generator of seed length $O(\log n)$ which fools all

weight-based tests, meaning that the distribution of the weight of any subset is ε-close to the appropriate binomial distribution in statistical distance. In particular, even for the special case of $m = 2$, combinatorial shapes strengthen the ever so versatile ε-biased spaces [78].

1.1.3 Computational Learning Theory

Our learning theory results stem from analyzing the noise tolerance, as quantified by *average sensitivity* and *noise sensitivity* of the concept classes we study. Roughly speaking, the average sensitivity of a function measures the expected number of bit positions needed to be flipped to change the value of the function. The noise sensitivity on the other hand measures the probability that a random *perturbation* of the input changes the value of the function.

Our sensitivity bounds along with the known connections between sensitivity bounds and learning [50] lead to efficient algorithms for learning the classes of PTFs and intersections of halfspaces with respect to the uniform distribution in the *agnostic* model.

Gotsman-Linial Conjecture: Sensitivity of PTFs Gotsman and Linial [39] conjectured that the average sensitivity of degree d PTFs is at most $O(d\sqrt{n})$. In work with Harsha and Klivans [41], [27], we obtain the first nontrivial bounds for the average and noise sensitivity of low-degree PTFs. Our work makes the first progress on the conjecture in over 15 years.

We also introduce a new *regularity lemma* about *random restrictions*

of PTFs. The regularity result we obtain can be seen as part of the high level "randomness vs structure" approach that has played a crucial role in many recent breakthroughs in additive number theory and combinatorics. Our lemma roughly says that either the behavior of the PTF under the uniform distribution over the hypercube is similar to its behavior under the Gaussian distribution, or the PTF essentially depends on only a few variables. The regularity lemma relies on the invariance principle for low-degree polynomials of Mossel et al. [76] and plays an important role in our construction of a PRG for PTFs.

Sensitivity of Intersections of Halfspaces In work with Harsha and Klivans [42] we show that the noise sensitivity of intersections *regular* halfspaces is poly-logarithmic in the number of halfspaces. Here, a halfspace is regular if none of its coefficients is much larger than the others. The previous best bound [59] had at least a linear dependence on the number of halfspaces, even for strongly regular halfspaces such as *reoriented majorities* (i.e., halfspaces with coefficients in $\{1, -1\}$). The bound on noise sensitivity is obtained by using our invariance principle for polytopes to translate the problem to the Gaussian space and invoking a nontrivial result of Nazarov [79] who essentially studies the same problem in the Gaussian setting.

Learning intersections of halfspaces is a central open problem in learning theory. Combined with the known connection between sensitivity and learning, our result gives the first quasi-polynomial time algorithm for agnos-

tically learning intersections of halfspaces under the uniform distribution for regular halfspaces. Previous algorithms had an exponential dependence on the number of halfspaces even for learning reoriented majorities without any errors.

1.1.4 Organization of the thesis

We start by discussing some preliminaries and notation in Chapter 2. Most of the definitions and other content in this chapter can be read as and when needed in the remainder of the thesis. However, we recommend reading Section 2.2.1 which gives a high level description of the Replacement method that plays a prominent role in many of our results. The remaining content broadly falls into three parts:

1. Invariance principles - Chapters 3, 4.
2. Sensitivity bounds and applications in learning theory - Chapters 5, 6.
3. Pseudorandom generators - Chapters 7, 8, 9.

We finish with a discussion of relevant open problems.

The results in Chapters 3, 8, 6 are based on joint work with Prahladh Harsha and Adam Klivans [42]. The results in Chapter 5 are based on joint work with Prahladh Harsha and Adam Klivans, [41, 27]. The results in Chapter 7 are based on joint work with David Zuckerman [74]. The results in Chapters 4, 9 are based on joint work with Parikshit Gopalan, Omer Reingold and David Zuckerman [36].

Chapter 2

Preliminaries

In this chapter we review the basic tools from probability theory, pseudorandomness and learning theory that we use. Before going into details we first highlight some notations.

2.1 Notation

The following list summarizes our frequently used conventions.

- We typically denote real-valued random variables by upper case letters $X, Y, \ldots,$.
- For a real-valued random variable X, we let $\mathbb{E}[X]$ (or $\mu(X)$), $\sigma(X)$, $\mathsf{Var}[X]$ denote it's mean, standard deviation and variance respectively.
- For a real-valued random variable X, $p > 0$, the ℓ_p norm of X (if finite) is defined by $\|X\|_p = \mathbb{E}[|X|^p]^{1/p}$.
- For two real-valued random variables X, Y, the Kolmogorov-Smirnov distance or cdf distance is defined by $\mathsf{d_{cdf}}(X, Y) = \sup_{t \in \mathbb{R}} |\mathsf{Pr}[X < t] - \mathsf{Pr}[Y < t]|$.

- For two integer-valued random variables X, Y, the statistical distance is defined by:

$$\mathsf{d_{TV}}(X,Y) \equiv \sup_{A \subseteq \mathbb{Z}} |\,\mathsf{Pr}[X \in A] - \mathsf{Pr}[Y \in A]| = \frac{1}{2}\sum_i |\,\mathsf{Pr}[X=i] - \mathsf{Pr}[Y=i]|.$$

- For a multi-set S, $x \in_u S$ denotes a uniformly random element of S.

- We typically use $\{1,-1\}$ to denote bits. In particular, unless otherwise specified, the n-dimensional hypercube will be the graph $G = (V,E)$ indexed by vertex set $V = \{1,-1\}^n$ and edge set E obtained by joining two strings $u, v \in \{1,-1\}^n$ if they differ in a single coordinate.

- For vectors $u, v \in \mathbb{R}^n$, we let $\langle u, v\rangle = \sum_i u_i v_i$ denote the inner product between u, v. For $u \in \mathbb{R}^n$, we denote its Euclidean norm by $\|u\|_2 = (\sum_i u_i^2)^{1/2}$ (we often drop the suffix 2 when its clear from context).

- For a vector $u \in \mathbb{R}^n$ and $S \subseteq [n]$, $u_S \in \mathbb{R}^S$ denotes the vector u restricted to the coordinates in the set S.

- $\mathcal{N}^n$ (where $\mathcal{N} = \mathcal{N}(0,1)$) denotes the standard multivariate spherical Gaussian distribution over $\mathbb{R}^n$ with mean 0 and identity covariance matrix. $\mathcal{N}(a,b)$ denotes the Gaussian distribution with mean a and standard deviation b.

We now start with basic notions from probability theory.

2.2 Probability Theory

We repeatedly use three fundamental tools from probability theory: hypercontractivity, Berry-Eséen theorem and the replacement method. We review these concepts below.

Definition 2.2.1. A polynomial $P : \mathbb{R}^n \to \mathbb{R}$ is a function formally defined by

$$P(x_1, \ldots, x_n) = \sum_{I \subseteq [n], I \neq \emptyset} a_I \prod_{i \in I} x_i.$$

The degree of the polynomial P is defined as $\text{degree}(P) = \max\{|I| : a_I \neq 0\}$. For a polynomial P as above and $p > 0$, we define the p'th norm of P by $\|P\|_p = \mathbb{E}_{x \in_u \{1,-1\}^n}[|P(x)|^p]^{1/p}$.

Note that by definition, polynomials for us are multilinear and that $\|P\|_2^2 = \sum_{I \subseteq [n]} a_I^2$. As we are mainly interested in the case where the inputs take values in $\{1, -1\}$, the assumption is without loss of generality. We next introduce the notion of *regular* polynomials. Intuitively, a polynomial is regular if none of the variables has too much influence on the value of the polynomial compared to the others.

Definition 2.2.2. A polynomial $P : \mathbb{R}^n \to \mathbb{R}$, $P(x) = \sum_{I \subseteq [n]} a_I \prod_{i \in I} x_i$ is ε-regular if for every $i \in [n]$,

$$\sum_{i=1}^{n} \left(\sum_{I \subseteq [n], I \ni i} a_I^2 \right)^2 \leq \varepsilon^2 \|P\|_2^4.$$

Observe that for any polynomial P as above and $0 < p < q$, by the power-mean inequality, $\|P\|_p \leq \|P\|_q$. However, we cannot in general obtain a meaningful bound on $\|P\|_q$ in terms of $\|P\|_p$. On the other hand, the hypercontractivity inequality (also known in literature as Bonami-Beckner inequality or Gross inequality or Khintchine inequalities for the case of linear polynomials) says that when P is a low-degree polynomial, we get a reverse inequality bounding $\|P\|_q$ in terms of $\|P\|_p$:

Theorem 2.2.1 (Hypercontractivity, [62])**.** *For $1 < p < q < \infty$, and $P : \mathbb{R}^n \to \mathbb{R}$ a degree d polynomial, the following holds:*

$$\|P\|_q \leq \left(\frac{q-1}{p-1}\right)^{d/2} \|P\|_p. \tag{2.2.1}$$

The above inequality and it's closely related variants have played a very important role in several recent breakthroughs in analysis of Boolean functions and its applications in hardness of approximation, social choice theory among others. In this thesis, our use of hypercontractivity will be limited to the cases where $p = 2, q = 4$ or P has degree one. We state these special cases below.

Lemma 2.2.2. *[$(2,4)$-Hypercontractivity] For any degree d polynomial $P : \mathbb{R}^n \to \mathbb{R}$, $\|P\|_4 \leq 3^{d/2}\|P\|_2$.*

Lemma 2.2.3. *[Khintchine Inequalities] For any $w_1, \ldots, w_n \in \mathbb{R}$,*

$$\mathop{\mathbb{E}}_{x \in_u \{1,-1\}^n}[\,|\langle w, x\rangle|^p\,]^{1/p} \leq \sqrt{p} \cdot \|w\|_2.$$

We next state the Berry-Esséen theorem which gives a quantitative form of the classical central limit theorem and can be seen as an *invariance principle* for halfspaces.

Theorem 2.2.4 (Theorem 1, XVI.5, [34], [93]). *Let $Y_1, \ldots, Y_t$ be independent random variables with $E[Y_i] = 0$, $\sum_i E[Y_i^2] = \sigma^2$, $\sum_i E[|Y_i|^3] \leq \rho$. Let random variable $S_n = (Y_1 + \ldots Y_n)/\sigma$, and let $Z \leftarrow \mathcal{N}(0,1)$. Then,*

$$\mathsf{d}_{\mathsf{cdf}}(S_n, Z) < \frac{\rho}{\sigma^3}.$$

The Berry-Esséen theorem along with the invariance principle for PTFs of Mossel et al. [76] play an important role in our constructions of PRGs for PTFs. In fact, the above theorem serves us as a *model* invariance principle and much of this thesis is geared toward obtaining extensions and derandomizations of the theorem. We next state the invariance principle for PTFs of Mossel et al. which can be viewed as generalizing the Berry-Esséen theorem which is applicable only to linear threshold functions (halfspaces).

Theorem 2.2.5 (IP for PTFs, Mossel et al.). *There exists a universal constant C such that the following holds. Let $P : \mathbb{R}^n \to \mathbb{R}$ be a degree d ε-regular (multi-linear) polynomial. Then, for $x \in_u \{0,1\}^n$ and $y \leftarrow \mathcal{N}(0,1)^n$,*

$$\mathsf{d}_{\mathsf{cdf}}(P(x), P(y)) \leq C\, d\, \varepsilon^{2/(4d+1)}.$$

The result stated in [76] uses $\max_i w_i^2(P)$ as the notion of regularity instead of $\sum_i w_i^4(P)$ as we do. However, their proof extends straightforwardly to the above.

Finally, we use the following anti-concentration bound for low-degree polynomials due to Carbery and Wright several times (the following is a special case of Theorem 8 of [21]; in their notation, set $q = 2d$ and the log-concave distribution μ to be $\mathcal{N}^n$).

Theorem 2.2.6 (Carbery-Wright anti-concentration bound)**.** *There exists an absolute constant C such that for any polynomial Q of degree at most d with $\|Q\| = 1$ and any interval $I \subseteq \mathbb{R}$ of length α, $\mathsf{Pr}_{X \leftarrow \mathcal{N}^n}[Q(X) \in I] \leq Cd\,\alpha^{1/d}$.*

2.2.1 The Replacement Method

Next, we briefly discuss a technique for proving invariance principles such as the Berry-Esséen theorem. Our discussion here will by choice be lacking in detail and is only meant to highlight one of the central ideas that we use recurrently in this thesis.

The original proof of the Berry-Esséen theorem (from 1942) was through Fourier analysis and made use of Esséen's inequality that gives a quantitative way of converting a bound on the (essentially, L_1-)distance between the characteristic functions of real-valued random variables to Kolmogorov-Smirnov distance between the random variables. The Fourier theoretic arguments are very precise in the sense of yielding optimal error bounds. However, they are quite intricate and are not amenable to obtaining extensions for instance, to more general classes of functions or to the case with dependencies between variables.

Of relevance to us in this context is the alternate proof of the cen-

tral limit theorem due to Lindeberg ([64]) from 1922. His proof was later refined to obtain quantitative versions of the central limit theorem in analogy to the Berry-Esséen theorem, although the exact bounds were slightly weaker. However, the Lindeberg method is much more versatile and was used in the seminal work of Mossel et al. [76] who among others things, used it to show an invariance principle for polynomial threshold functions. We roughly follow the description of the Lindeberg method due to Mossel et al., and Mossel et al. [75].

To describe the Lindeberg method, which we refer to from now on as *Replacement Method* (for reasons that will be clear later) or *hybrid argument*, let's first set up some notation. Let $f : \mathbb{R}^n \to \{0, 1\}$ be a Boolean function that we seek to show an invariance principle for. That is, we have two vector-valued product distributions (the coordinates are independent of one another) D_1, D_2 on $\mathbb{R}^n$, samples $x \leftarrow D_1$ and $y \leftarrow D_2$ and we wish to show that $\mathbb{E}[f(x)] \sim \mathbb{E}[f(y)]$. The Replacement method then involves two modular steps.

Step One: The first step involves proving an invariance principle for the distributions D_1, D_2 with respect to *smooth* functions. By this, we mean proving that

$$\left| \mathop{\mathbb{E}}_{x \leftarrow D_1}[\psi(x)] - \mathop{\mathbb{E}}_{y \in \leftarrow D_2}[\psi(y)] \right| \leq \varepsilon_n(\psi),$$

where $\psi : \mathbb{R}^n \to \mathbb{R}$ is a smooth function with well-behaved derivatives (for instance, with a universal bound on infinity norm of the first few derivatives) and $\varepsilon_n(\psi)$ is a quantity that goes to zero. The idea being that as ψ has well-

behaved derivatives, changing the input to ψ a little should not change the output too much. We can then exploit this by showing that

$$\mathbb{E}[\psi(y_1, \ldots, y_{i-1}, x_i, \ldots, x_n)] \sim \mathbb{E}[\psi(y_1, \ldots, y_{i-1}, y_i, x_{i+1}, \ldots, x_n)],$$

for $i = 1, \ldots, n$. The above description motivates the label *replacement method*, as we are showing that replacing each x_i from $(x_1, \ldots, x_n)$ with y_i from $(y_1, \ldots, y_n)$ iteratively does not change the distribution too much. Note that the above approach is similar in spirit to the traditional hybrid argument from cryptography.

Step Two: The second step involves finding a suitable smooth approximating function ψ for the original function f. Obtaining smooth approximations ψ for test functions f is a well-studied problem in approximation theory and almost optimal bounds are known for various interesting classes f.

Finally, one needs to combine the above two steps to obtain the desired invariance principle.

Our proof of the invariance principle for polytopes and the analysis of the generator for PTFs follow the above approach at a high level. However, in both cases we need a more versatile form of the replacement method that we term *block-replacement method* where we replace whole blocks of variables (instead of one at a time) for a fewer number of iterations.

2.3 Pseudorandomness

We now review some of the standard tools in derandomization that we use.

Definition 2.3.1 (k-wise independent Hash Families)**.** For $\alpha \geq 0$, a family of hash functions $\mathcal{H} = \{h : [n] \to [t]\}$ is α-almost k-wise independent if for all distinct $i_1, \ldots, i_k \in [n]$ and $\ell_1, \ldots, \ell_k \in [t]$,

$$\Pr_{h \in_u \mathcal{H}}[\, h(i_1) = \ell_1 \,\wedge\, h(i_2) = \ell_2 \,\wedge \cdots \wedge\, h(i_k) = \ell_k \,] \leq \frac{1+\alpha}{t^k}.$$

We say the $\mathcal{H}$ is k-wise independent if the above inequality holds as an equality with $\alpha = 0$.

Efficient constructions of $\mathcal{H}$ as above with $|\mathcal{H}| = O(n^k)$ are known [22]. A family of k-wise independent permutations $\mathcal{H} = \{h : [n] \to [n]\}$ is defined similarly, with the additional requirement that the hash functions $h : [n] \to [n]$ be permutations.

Definition 2.3.2 (k-wise independent spaces)**.** A generator $G : \{0,1\}^r \to [m]^n$ is said to generate a k-wise independent space if for $y \in_u \{0,1\}^r$, for all distinct $i_1, \ldots, i_k \in [n]$, $b_1, \ldots, b_k \in [m]$,

$$\Pr[\, (G(y))_{i_1} = b_1 \,\wedge\, (G(y))_{i_2} = b_2 \,\wedge \cdots \wedge\, (G(y))_{i_k} = b_k \,] = \frac{1}{m^k}.$$

It is easy to see that k-wise independent spaces are essentially a different view of k-wise independent hash functions and efficient constructions of

generators G as above with $r = O(k(\log m + \log n))$ are known. For the case when $m = 2$, we have better constructions when the equality constraint above is relaxed by allowing an error margin:

Definition 2.3.3 (Almost k-wise independent spaces)**.** For $0 < \delta < 1$, a generator $G : \{0,1\}^r \to \{0,1\}^n$ is said to generate a δ-almost k-wise independent space if for $y \in_u \{0,1\}^r$, for all distinct $i_1, \ldots, i_k \in [n]$, $b_1, \ldots, b_k \in \{0,1\}$,

$$\mathsf{Pr}\left[\,(G(y))_{i_1} = b_1 \,\wedge\, (G(y))_{i_2} = b_2 \,\wedge \cdots \wedge\, (G(y))_{i_k} = b_k\,\right] - \frac{1}{2^k}\right| < \delta.$$

Efficient generators G as above with $r = O(k + \log n + \log(1/\delta))$ [78] are known. We also use the following generalization of k-wise independence to arbitrary non-uniform distributions.

Definition 2.3.4. A collection of random variables $(X_1, \ldots, X_n)$ over a universe U is k-wise independent if for all $i_1, \ldots, i_k \in [n]$, $u_1, \ldots, u_k \in U$,

$$\mathsf{Pr}[\,X_{i_1} = u_1 \wedge X_{i_2} = u_2 \wedge \cdots \wedge X_{i_k} = u_k\,] = \mathsf{Pr}[\,X_{i_1} = u_1\,] \cdot \mathsf{Pr}[\,X_{i_2} = u_2\,] \cdots \mathsf{Pr}[\,X_{i_k} = u_k\,].$$

Much of this thesis deals with constructing pseudorandom generators (PRGs) for various classes of functions. We define PRGs in a general context below and specialize to the appropriate class when needed.

Definition 2.3.5. Let $\mathcal{C} : \{0,1\}^n \to \{0,1\}$ be a class of functions. A function $G : \{0,1\}^r \to (\{0,1\}^D)^T$ is said to ε-fool $\mathcal{C}$ if, for all $f \in \mathcal{C}$,

$$\left|\, \mathop{\mathsf{Pr}}_{x \in_u \{0,1\}^n} [f(x) = 1] - \mathop{\mathsf{Pr}}_{y \in_u \{0,1\}^r} [f(G(y)) = 1] \,\right| \leq \varepsilon.$$

The pseudorandom generators (PRGs) for polynomial-width read once branching programs (ROBPs) of Nisan [80], Impagliazzo et al. [46] play an important role in many of our constructions. We review them below.

Definition 2.3.6 (ROBP)**.** An (S, D, T)-branching program M is a layered multi-graph with a layer for each $0 \leq i \leq T$ and at most 2^S vertices (states) in each layer. The first layer has a single vertex v_0 and each vertex in the last layer is labeled with 0 (rejecting) or 1 (accepting). For $0 \leq i < T$, a vertex v in layer i has exactly 2^D outgoing edges each labeled with an element of $\{0,1\}^D$ and ending at a vertex in layer $i+1$.

Note that by definition, an (S, D, T)-branching program is read-once. We also use the following notation. Let M be an (S, D, T)-branching program and v a vertex in layer i of M.

1. For $z = (z^i, z^{i+1}, \ldots, z^T) \in (\{0,1\}^D)^{T+1-i}$ call (v, z) an *accepting* pair if starting from v and traversing the path with edges labeled z in M leads to an accepting state.

2. For $z \in (\{0,1\}^D)^T$, let $M(z) = 1$ if (v_0, z) is an accepting pair, and $M(z) = 0$ otherwise.

3. $A_M(v) = \{z : (v, z) \text{ is accepting in } M\}$ and $P_M(v)$ is the probability that (v, z) is an accepting pair for z chosen uniformly at random.

4. For brevity, let $\mathcal{U}$ denote the uniform distribution over $(\{0,1\}^D)^T$.

Nisan [80] and Impagliazzo et al. [46] gave PRGs that δ-fool (S, D, T)-branching programs with seed length $r = O((S + D) \log T + \log(T/\delta) \log T)$. For $T = \text{poly}(S, D)$, the PRG of Nisan and Zuckerman [81] fools (S, D, T)-branching programs with seed length $r = O(S + D)$. We state the bounds of the generators of Impagliazzo et al. and Nisan and Zuckerman below.

Theorem 2.3.1 (Impagliazzo et al. [46])**.** *There exists an explicit generator $G_{INW} : \{0,1\}^r \to (\{0,1\}^D)^T$ that δ-fools (S, D, T)-branching programs with seed-length $r = O(D + (S + \log(T/\delta) \log T))$.*

Theorem 2.3.2 (Nisan and Zuckerman [81])**.** *For all $0 < \gamma < 1$ and $c > 0$, there exists a constant $C = C(c, \gamma)$ such that the following holds. There exists an explicit generator $G_{NZ} : \{0,1\}^r \to (\{0,1\}^D)^T$ that δ-fools (S, D, T)-branching programs with error $\delta = 2^{\log^{1-\gamma}(S+D)}$ and seed-length $r = C(S + D)$ when $T \leq (S + D)^c$.*

2.4 Learning Theory

We next review some of the basic notions from learning theory that we use.

Average sensitivity [15] and noise sensitivity [49, 16] are two fundamental quantities that arise in the analysis of Boolean functions. Roughly speaking, the average sensitivity of a Boolean function f measures the expected number of bit positions that change the sign of f for a randomly chosen input, and the noise sensitivity of f measures the probability over a randomly cho-

sen input x that f changes sign if each bit of x is flipped independently with probability δ.

We begin by defining the (Boolean) noise sensitivity of a Boolean function:

Definition 2.4.1 (Noise Sensitivity)**.** Let f be a Boolean function $f : \{1,-1\}^n \to \{1,-1\}$. For any $\delta \in (0,1)$, let X be a random element of the hypercube $\{1,-1\}^n$ and Z a δ-perturbation of X defined as follows: for each i independently, Z_i is set to X_i with probability $1-\delta$ and $-X_i$ with probability δ. The noise sensitivity of f, denoted $\mathbb{NS}_\delta(f)$, for noise δ is then defined as follows: $\mathbb{NS}_\delta(f) = \Pr[f(X) \neq f(Z)]$.

We also study the closely related notion of average sensitivity (also known as total influence). We first define the ith influence of a Boolean function:

Definition 2.4.2 (Average Sensitivity)**.** Let f be a Boolean function, and let X be a random element of the hypercube $\{1,-1\}^n$. Define $X^{(i)}$ to be an element of $\{1,-1\}^n$ chosen as follows: $X_i^{(i)} = -X_i$ and $X_j^{(i)} = X_j$ for $j \neq i$. The influence of the i^{th} variable, denoted by $\mathbb{I}_i(f)$ is defined as follows:

$$\mathbb{I}_i(f) = \Pr\left[f(X) \neq f\left(X^{(i)}\right)\right].$$

The sum of all the influences is referred to as the average sensitivity of the function f:

$$\mathbb{AS}(f) = \sum_i \mathbb{I}_i(f).$$

Bounds on the average and noise sensitivity of Boolean functions have direct applications in hardness of approximation [44, 57], hardness amplification [82], circuit complexity [65], the theory of social choice [51], and quantum complexity [94].

In this thesis, we focus on applications in learning theory, where it is known that bounds on the noise sensitivity of a class of Boolean functions yield learning algorithms for the class that succeed in harsh models of noise, such as the agnostic model of learning, [50]. We define the *agnostic* model of learning of Haussler [45] and Kearns, Schapire and Sellie [56] below.

Definition 2.4.3. Let $\mathcal{D}$ be an arbitrary distribution on $\mathcal{X}$ and $\mathcal{C}$ a class of Boolean functions $f : \mathcal{X} \to \{-1, 1\}$. For $\delta, \varepsilon \in (0, 1)$, we say that algorithm A is a (δ, ε)-agnostic learning algorithm for $\mathcal{C}$ with respect to $\mathcal{D}$ if the following holds. For any distribution $\mathcal{D}'$ on $\mathcal{X} \times \{-1, 1\}$ whose marginal over $\mathcal{X}$ is $\mathcal{D}$, if A is given access to a set of labeled examples (x, y) drawn from $\mathcal{D}'$, then with probability at least $1 - \delta$ algorithm A outputs a hypothesis $h : \mathcal{X} \to \{-1, 1\}$ such that

$$\Pr_{(x,y)\sim\mathcal{D}'}[h(x) \neq y] \leq \mathsf{opt} + \varepsilon$$

where opt is the error made by the best classifier in $\mathcal{C}$, that is,

$$\mathsf{opt} = \inf_{g\in\mathcal{C}} \Pr_{(x,y)\sim\mathcal{D}'}[g(x) \neq y].$$

Kalai, Klivans, Mansour and Servedio [50] showed that the existence of low-degree real valued polynomial l_2-approximators to a class of functions,

implies agnostic learning algorithms for the class. In an earlier result, Klivans, O'Donnell and Servedio [59] gave a precise relationship between polynomial approximation and noise sensitivity, essentially showing that small noise sensitivity bounds imply good low-degree polynomial l_2-approximators. Combining these two results, it follows that bounding the noise sensitivity of a concept class $\mathcal{C}$ yields an agnostic learning algorithm for $\mathcal{C}$ with respect to the uniform distribution on the hypercube. We state their results (using our notation) for reference.

Theorem 2.4.1 ([50]). *Let $\mathcal{C}$ be a class of functions mapping X to $\{-1,1\}$. Let D be a distribution on $\mathcal{X} \times \{-1,1\}$ with marginal distribution D_X. Assume that for each $f \in \mathcal{C}$, there exists a polynomial P of degree d such that $\mathbb{E}_{\mathcal{D}_X}\left[(P(x) - f(x))^2\right] \leq \varepsilon^2$. Then $\mathcal{C}$ is agnostically learnable to accuracy ε in time poly(n^d/ε).*

Theorem 2.4.2 ([59]). *Let $f : \{1,-1\}^n \to \{1,-1\}$ with $\mathbb{NS}_\delta(f) \leq m(\delta)$ for some (invertible) function $m : (0,1) \to (0,1)$. Then there exists a constant C and polynomial P of degree at most $C \cdot (1/m^{-1}(\delta))$ such that*

$$\mathop{\mathbb{E}}_{x \in_u \{1,-1\}^n}\left[(P(x) - f(x))^2\right] \leq \delta.$$

Finally, it is implicit from a recent paper of Blais, O'Donnell and Wimmer [19] that bounding the Boolean noise sensitivity for a concept class $\mathcal{C}$ yields non-trivial learning algorithms for a very broad class of discrete and continuous product distributions. We believe this is additional motivation for obtaining bounds on a function's Boolean noise sensitivity.

Chapter 3

An Invariance Principle for Polytopes

3.1 Introduction

The main result of this chapter is an invariance principle for characteristic functions of polytopes generalizing the Berry-Esséen theorem. Recall that a polytope $\mathcal{K}$ is a (possibly unbounded) convex set in $\mathbb{R}^n$ formed by the intersection of some finite number of supporting halfspaces. We refer to $\mathcal{K}$ as a k-polytope if it is equal to the intersection of k halfspaces. The main theorem of this section is as follows (see Theorem 3.3.1 for exact statement):

Theorem 3.1.1 (Invariance Principle for Polytopes)**.** *For $\mathcal{K}$ a k-polytope,*

$$\left| \Pr_{x \in_u \{-1,1\}^n}[x \in \mathcal{K}] - \Pr_{x \leftarrow \mathcal{N}^n}[x \in \mathcal{K}] \right| \leq \log^{8/5} k \cdot \Delta.$$

The parameter Δ depends on the coefficients of the bounding hyperplanes of $\mathcal{K}$ and is small if these coefficients are sufficiently regular. In particular, if $\mathcal{K}$ equals $\{x \mid W^T x \leq \theta\}$ for $W \in R^{n \times k}$ and $\theta \in R^k$, and each column u of W is ε-regular, i.e., satisfies $\sum_{i=1}^n u_i^4 \leq \varepsilon^2 \|u\|_2^2$, then Δ is less than $\varepsilon^{1/6}$. Note that there is no restriction on the vector θ. The invariance principle also holds more generally for any product distribution that is hypercontractive and whose first four moments are appropriately bounded.

The novelty of our theorem is the dependence of the error on k. Applying a recent result due to Mossel [75], it is possible to obtain a statement similar to Theorem 3.1.1 with an error term that has a polynomial dependence on k. Achieving polylogarithmic dependence on k, however, is much harder, and we need to use some nontrivial results from the analysis of convex sets in Gaussian space.

The case $k = 1$, a single halfspace, corresponds to the Berry-Esséen theorem Theorem 2.2.4. We can therefore view our principle as a generalization of the Berry-Esséen theorem for polytopes. Further, understanding the structure of integer points in polytopes (that is, solutions to integer programs) is an important topic in computer science [9], optimization [100], and combinatorics [12], and we believe our invariance principle will find many applications.

As mentioned in the introduction, we use the invariance principle to derive new results in learning theory (bounding the noise sensitivity of intersections of halfspaces, sec:nsinths) and pseudorandomness (PRGs for polytopes, sec:prgpolytopes).

3.1.1 Proof Outline

In this section, we give a high level outline of the proof of our invariance principle and contrast it with the replacement method as used in the works of Mossel et al. [76] and Mossel [75]. As outlined at a high level in Section 2.2.1, the proof proceeds in two steps.

Step One: We first prove an invariance principle for smooth functions. By this we mean proving that

$$\left| \mathop{\mathbb{E}}_{X \in \{-1,1\}^n} [\Psi(\ell_1(X), \ldots, \ell_k(X))] - \mathop{\mathbb{E}}_{Y \in \mathcal{N}^n} [\Psi(\ell_1(Y), \ldots, \ell_k(Y))] \right| \leq \gamma, \quad (3.1.1)$$

where $\ell_1, \ldots, \ell_k$ are linear functions (corresponding to the normals of the faces of the k-polytope) and Ψ is a smoothing function. The value γ will depend on k, the coefficients of the ℓ_p's and the derivatives of Ψ. The function Ψ is can be viewed as a "test" function and is smooth if there is a uniform bound on its fourth derivative. Notice here that Ψ maps $\mathbb{R}^k$ to $\mathbb{R}$; in [76], they were concerned with the value $\Psi(Q(X))$ for a low-degree polynomial Q and a univariate test function Ψ.

At this point, we could take Ψ to be the k-wise product of a test function constructed by Mossel et al. to approximate the logical AND function. Further, Mossel provides a very general framework for obtaining multivariate test functions and gives bounds for the overall error incurred by the replacement method. Here we run into our first difficulty: the standard replacement method as used by Mossel et al. and Mossel results in a bad dependence on the coefficients of the ℓ_p's. In particular, the resulting error term is not small even for polytopes formed by the intersection of regular halfspaces.

To solve this problem, we use a non-standard hybrid argument that groups the input variables into blocks. We observe that in the replacement method it is irrelevant in which order we replace X_i's with Y_i's – in fact a random order would suffice. Further, we can group the X_i's into blocks and

proceed blockwise with the replacement method. To implement this intuition, we partition $[n]$ randomly into a set of blocks and replace all the X_i's within a block by the corresponding Y_i's one block at a time. Proceeding in this fashion with a random partitioning has a "smoothing effect" on the coefficients of the linear functions resulting in a much better bound on the error in terms of the coefficients.

Roughly speaking, if ℓ_{pi} denotes the i'th coefficient of ℓ_p, then the standard replacement methods of [84], [76], [75] incur an error proportional to $\sum_{i\in[n]} \left(\max_{p\in[k]} |\ell_{pi}|^4\right)$, which can be as large as $\Omega(k)$ even for regular functions ℓ_p. In contrast, our *randomized-blockwise-replacement* method only suffers an error of $(\log k) \cdot \max_{p\in[k]} \sum_i |\ell_{pi}|^4$, which is small for regular functions. It turns out that in the above analysis, we can choose the random partitioning into blocks in a $\Theta(\log k)$-wise independent manner, instead of uniformly at random, and this is crucial for our PRG constructions.

Step Two: Given the above invariance principle for smooth functions, we now aim to translate the closeness in expectation for smooth functions to closeness in cdf distance. Here the smoothness of the test function Ψ becomes important, and we run into our second problem: the natural choice of test function Ψ (the multivariate version of the test function from Mossel et al.) leads to an error bound on the order of k, rather than $\mathrm{poly}(\log k)$. To get around this problem, we first observe that in Mossel's proof of the multivariate invariance principle as in our *randomized-blockwise-*

hybrid argument, it suffices to bound the 'l_1-norm' of the fourth derivative $\sup_{x\in\mathbb{R}^k}(\sum_{p,q,r,s\in[k]}|\partial_p\partial_q\partial_r\partial_s\Psi(x)|)$, instead of uniformly bounding the fourth derivative $\sup_{x\in\mathbb{R}^k,p,q,r,s\in[k]}(|\partial_p\partial_q\partial_r\partial_s\Psi(x)|)$. Thus, it suffices to obtain a smooth approximation of the AND function for which the former quantity is small. Fortunately for us, we uncovered a beautiful result due to Bentkus [17], who constructs a smooth approximation of the AND function with precisely this property.

The final difficulty for translating closeness in expectation as in Equation 3.1.1 to closeness in cdf distance is to prove that Ψ differs from the characteristic function only on a set of small Gaussian measure. To this end, we show that it suffices to bound the Gaussian measure of l_∞-neighborhoods around the boundary of k-polytopes. For an l_∞-neighborhood of width λ, a union bound would imply Gaussian measure on the order of $k\lambda$. At this point, however, we can apply a nontrivial result due to Nazarov [79] on the Gaussian surface area of k-polytopes to get the much better bound of $\sqrt{\log k}\,\lambda$. This result of Nazarov was used before by Klivans et al. [60] in the context of learning intersections of halfspaces with respect to Gaussian distributions.

3.1.2 Related Work

As mentioned earlier, the classical Berry-Esséen theorem, Theorem 2.2.4, gives an invariance principle for the case of a single halfspace (i.e., $k = 1$).

Bentkus [18] proves a multidimensional Berry-Esséen theorem for sums of vector-valued random variables each with identity covariance matrix, whose

error term depends on the Gaussian surface area of the test set. Although his paper deals with topics related to our work, his result seems to have no implications in our setting.

Before stating our main result formally we first set up some notations.

3.2 Notation and Preliminaries

We use the following notation.

1. For $W \in \mathbb{R}^{n\times k}$, $\theta \in \mathbb{R}^k$, $\mathcal{K}(W,\theta)$ denotes the polytope $\mathcal{K}(W,\theta) = \{x : W^T x \leq \theta\}$. We say a polytope $\mathcal{K}(W,\theta)$ as above has k faces.

2. Unless stated otherwise, throughout this chapter we work with the same polytope $\mathcal{K}(W,\theta)$ and assume that the columns of the matrix W have norm one. We often shorten $\mathcal{K}(W,\theta)$ to $\mathcal{K}$ if W,θ are clear from context. We assume that $k \geq 2$.

3. For $A \in \mathbb{R}^{m_1\times m_2}$, A^T denotes the transpose of A and for $p \in [m_2]$, A^p denotes the p'th column of A.

4. The all ones vector in $\mathbb{R}^k$ is denoted by $\mathsf{1}_k$.

5. For $u \in \mathbb{R}^k$, define rectangle

$$\mathsf{Rect}(u) = (-\infty, u_1] \times (-\infty, u_2] \times \cdots \times (-\infty, u_k].$$

Note that $x \in \mathcal{K}(W,\theta)$ if and only if $W^T x \in \mathsf{Rect}(\theta)$.

6. For a smooth function $\psi : \mathbb{R}^k \to \mathbb{R}$, let

$$\|\psi^{(4)}\|_1 = \sup\{ \sum_{p,q,r,s\in[k]} |\, \partial_p\partial_q\partial_r\partial_s\, \psi(a_1,\ldots,a_k)\,| \,:\, (a_1,\ldots,a_k) \in \mathbb{R}^k \}.$$

7. In this chapter, we shall not to try to be overtly specific about the universal constants and denote all universal constants by c, C, even when we have in mind different constants in the same equation.

Below we recall the definition of regular polynomials, Definition 2.2.2, specialized to the case of halfspaces.

Definition 3.2.1. A vector $u \in \mathbb{R}^n$ is ε-regular if $\sum_i u_i^4 \le \varepsilon^2\|u\|^2$. A matrix $W \in \mathbb{R}^{n\times k}$ is ε-regular if every column of W is ε-regular. A polytope $\mathcal{K} = \mathcal{K}(W,\theta)$ is ε-regular if W is ε-regular [1].

Our main invariance principle is applicable to a large class of product distributions that satisfy the following two properties.

Definition 3.2.2 (Proper Distributions)**.** A distribution μ over $\mathbb{R}$ is proper if for $X \leftarrow \mu$, $\mathbb{E}[X] = 0$, $\mathbb{E}[X^2] = 1$ and $\mathbb{E}[X^3] = 0$.

Definition 3.2.3 (Hypercontractive Distributions)**.** A distribution μ over $\mathbb{R}$ is hypercontractive, if there exists a constant c_μ such that the following holds. For any m, vector $u \in \mathbb{R}^m$, and any $p \ge 2$,

$$\left(\mathop{\mathbb{E}}_{X\leftarrow\mu^m}[\,|\langle u, X\rangle|^p\,]\right)^{1/p} \le c_\mu\sqrt{p}\left(\mathop{\mathbb{E}}_{X\leftarrow\mu^m}[\,|\langle u, X\rangle|^2\,]\right)^{1/2}.$$

[1] "Regular polytopes" have a different meaning in combinatorics, but for the purpose of this chapter, we will abuse notation and say a polytope is ε-regular if it is formed by the intersection of ε-regular halfspaces as in Definition 3.2.1.

Note that the above property is a direct analogue of the Khintchine inequalities for the uniform distribution over the hypercube $\{1,-1\}^n$, Corollary 2.2.3. It is well known that the spherical Gaussian distribution $\mathcal{N}^n$ satisfies Khintchine inequalities with constant $c_\mu = 1$.

3.3 Invariance Principle for Polytopes

Our main invariance principle for polytopes $\mathcal{K}(W,t)$ is as follows:

Theorem 3.3.1 (Invariance Principle for Polytopes)**.** *For any proper and hypercontractive distribution μ over $\mathbb{R}$ and any ε-regular k-polytope $\mathcal{K}$,*

$$\left| \Pr_{X \leftarrow \mu^n} [X \in \mathcal{K}] - \Pr_{Y \leftarrow \mathcal{N}^n} [Y \in \mathcal{K}] \right| \leq C\, c_\mu^2\, (\log^{8/5} k)\, (\varepsilon \log(1/\varepsilon))^{1/5}. \quad (3.3.1)$$

The proof of the theorem can be divided into three parts.

1. We establish an invariance principle for smooth functions on polytopes (Theorem 3.3.2) using an extension of Replacement method; Section 3.4 is devoted to proving this part.

2. We prove that for random variables A, B over $\mathbb{R}^k$, closeness with respect to smooth functions and anti-concentration bounds for one of the variables imply closeness with respect to rectangles (Lemma 3.3.3). To do so, we use a nontrivial result of Bentkus [17] on smooth approximations for the l_∞ norm.

3. We use a result of Nazarov [79] on Gaussian surface area of polytopes to

bound the Gaussian measure of "l_∞-neighborhoods" of polytopes in $\mathbb{R}^n$ (Lemma 3.3.4).

We begin by stating an *invariance principle for smooth functions* $\psi : \mathbb{R}^k \to \mathbb{R}$. The proof is involved, making use of the randomized-blockwise-hybrid argument alluded to in the introduction. For clarity we present the proof in the next section (Section 3.4).

Theorem 3.3.2 (Invariance Principle for Smooth Functions)**.** *For any proper and hypercontractive distribution μ over $\mathbb{R}$ and any ε-regular W and smooth function $\psi : \mathbb{R}^k \to \mathbb{R}$,*

$$\left| \mathop{\mathbb{E}}_{X \leftarrow \mu^n} [\psi(W^T X)] - \mathop{\mathbb{E}}_{Y \leftarrow \mathcal{N}^n} [\psi(W^T Y)] \right| \leq C\, c_\mu^2\, \|\psi^{(4)}\|_1\, (\log^3 k)\, (\varepsilon \log(1/\varepsilon)).$$

The following lemma shows that for two random variables A, B over $\mathbb{R}^k$, closeness with respect to smooth functions and *anti-concentration bounds* for the variable B imply closeness with respect to rectangles. Note that to use the lemma we do not need anti-concentration bounds for the random variable A.

Lemma 3.3.3 (Smooth Approximation of AND)**.** *Let A, B be two random variables over $\mathbb{R}^k$ satisfying the following conditions:*

- *For all smooth functions $\psi : \mathbb{R}^k \to \mathbb{R}$, $|\mathbb{E}[\psi(A)] - \mathbb{E}[\psi(B)]| \leq \Delta \|\psi^{(4)}\|_1$.*
- *For a function $g_k : [0,1] \to [0,1]$ the following holds:*

  for all $\lambda \in [0,1]$, $\sup_{\theta \in \mathbb{R}^k} (\mathsf{Pr}[\, B \in \mathsf{Rect}(\theta + \lambda \mathbf{1}_k) \setminus \mathsf{Rect}(\theta)\,]) \leq g_k(\lambda)$.

Then, for all $\theta \in \mathbb{R}^k$, $0 < \lambda < 1$, $|\,\mathsf{Pr}[A \in \mathsf{Rect}(\theta)] - \mathsf{Pr}[B \in \mathsf{Rect}(\theta)]\,| \leq C\Delta \log^3 k/\lambda^4 + Cg_k(\lambda)$.

Finally, we use the following anti-concentration bound that follows from Nazarov's estimate on the Gaussian surface area of polytopes [79]:

Lemma 3.3.4 (Anti-concentration bound for l_∞-neighborhood of rectangles)**.** *For* $0 < \lambda < 1$,

$$\mathop{\mathsf{Pr}}_{x \leftarrow \mathcal{N}^n}\left[W^T x \in \mathsf{Rect}(\theta) \setminus \mathsf{Rect}(\theta - \lambda 1_k)\right] = O(\lambda\sqrt{\log k}).$$

We first prove Theorem 3.3.1 using the above three results and then prove Lemmas 3.3.3 and 3.3.4 in Sections 3.3.1 and 3.3.2. Theorem 3.3.2 is then proved in Section 3.4.

Proof of Theorem 3.3.1. Let $X \leftarrow \mu^n$, $Y \leftarrow \mathcal{N}^n$ and let random variables $A = W^T X$, $B = W^T Y$. Then, by Lemma 3.3.4 and Theorem 3.3.2,

$$\mathsf{Pr}\left[B \in \mathbb{R}(\theta + \lambda 1_k) \setminus \mathbb{R}(\theta)\right] \leq C\sqrt{\log k}\,\lambda,$$

$$|\mathbb{E}[\psi(A)] - \mathbb{E}[\psi(B)]| \leq C\, c_\mu^2\, (\log^3 k)\, \varepsilon\, \log(1/\varepsilon)\, \|\psi^{(4)}\|_1,$$

where $\psi : \mathbb{R}^k \to \mathbb{R}$ is any smooth function, $\theta \in \mathbb{R}^k$ and $\lambda \in (0,1)$. Therefore, by Lemma 3.3.3, for $\theta \in \mathbb{R}^k$,

$$|\mathsf{Pr}\left[A \in \mathsf{Rect}(\theta)\right] - \mathsf{Pr}\left[B \in \mathsf{Rect}(\theta)\right]| \leq C\,(\log^6 k)\,\log(1/\varepsilon)\varepsilon/\lambda^4 + C\sqrt{\log k}\,\lambda.$$

The theorem now follows by setting $\lambda = (\log^{11/10} k)\,(\varepsilon \log(1/\varepsilon))^{1/5}$. □

3.3.1 Smooth Approximation of AND

We now prove Lemma 3.3.3. For this, we use the following nontrivial result of Bentkus [17] on smooth approximations for the l_∞ norm.

Theorem 3.3.5 (Bentkus [17])**.** *For every $\alpha > 0$ and $0 < \lambda < 1$, there exists a function $\psi \equiv \psi_{\alpha,\lambda} : \mathbb{R}^k \to \mathbb{R}$ such that $\|\psi^{(4)}\|_1 \le C \log^3 k/\lambda^4$ and*

$$\psi(a) = \begin{cases} 1 & if \ \ \|a\|_\infty \le \alpha \\ 0 & if \ \ \|a\|_\infty > \alpha + \lambda \\ \in [0,1] & otherwise \end{cases}.$$

Corollary 3.3.6. *For all $u \in \mathbb{R}^k$, $0 < \lambda < 1$, $T > \|u\|_\infty$, there exists a function $\psi \equiv \psi_{u,\lambda,T} : \mathbb{R}^k \to \mathbb{R}$ such that $\|\psi^{(4)}\|_1 \le C \log^3 k/\lambda^4$ and*

$$\psi(a) = \begin{cases} 1 & if \ \ \forall l \in [k], -T + u_l \le a_l \le u_l \\ 0 & if \ \ \exists l \in [k], a_l > u_l + \lambda \\ \in [0,1] & otherwise \end{cases}.$$

Proof. Let $\psi_{T/2,\lambda}$ be the function from Theorem 3.3.5 with $\alpha = T/2$. Define $\psi \equiv \psi_{u,\lambda,T} : \mathbb{R}^k \to \mathbb{R}$ by

$$\psi_{u,\lambda,T}(a_1, \ldots, a_k) = \psi_{T/2,\lambda}(a_1 + T/2 - u_1, a_2 + T/2 - u_2, \ldots, a_k + T/2 - u_k).$$

It is easy to check that ψ satisfies the conditions of the theorem. □

Proof of Lemma 3.3.3. Fix $\theta \in \mathbb{R}^k$, $0 < \lambda < 1$. Choose $T \in \mathbb{R}$ large enough so that $T > \|\theta\|_\infty$, $\mathsf{Pr}[\|A\|_\infty \ge T] < \Delta$ and $\mathsf{Pr}[\|B\|_\infty \ge T] < \Delta$. Let $\psi : \mathbb{R}^k \to \mathbb{R}$ be the function obtained from applying Corollary 3.3.6 to θ, λ, T. Then,

$$\begin{aligned} |\mathsf{Pr}\left[A \in \mathsf{Rect}(\theta)\right] - \mathsf{Pr}\left[A \in \mathsf{Rect}_T(\theta)\right]| \le \Delta, \\ |\mathsf{Pr}\left[B \in \mathsf{Rect}(\theta)\right] - \mathsf{Pr}\left[B \in \mathsf{Rect}_T(\theta)\right]| \le \Delta, \end{aligned} \tag{3.3.2}$$

where $\mathsf{Rect}_T(\theta) = [-T+\theta_1, \theta_1] \times [-T+\theta_2, \theta_2] \times \cdots \times [-T+\theta_k, \theta_k] \subseteq \mathbb{R}^k$. Observe that from the definition of ψ in Corollary 3.3.6 and Equation 3.3.2

$$\mathsf{Pr}\left[A \in \mathsf{Rect}(\theta)\right] \leq \mathbb{E}[\psi(A)] + \Delta \leq \mathbb{E}[\psi(B)] + \Delta\|\psi^{(4)}\|_1 + \Delta.$$

Similarly,

$$\begin{aligned}\mathbb{E}[\psi(B)] &\leq \mathsf{Pr}[B \in \mathsf{Rect}(\theta + \lambda 1_k)] \\ &= \mathsf{Pr}[B \in \mathsf{Rect}(\theta)] + \mathsf{Pr}[B \in \mathsf{Rect}(\theta + \lambda 1_k) \setminus \mathsf{Rect}(\theta)] \\ &\leq \mathsf{Pr}[B \in \mathsf{Rect}(\theta)] + g_k(\lambda),\end{aligned}$$

where the last inequality follows from the definition of g_k. Combining the above two equations we get

$$\mathsf{Pr}[A \in \mathsf{Rect}(\theta)] \leq \mathsf{Pr}[B \in \mathsf{Rect}(\theta)] + 2\Delta\|\psi^{(4)}\|_1 + g_k(\lambda) \leq \mathsf{Pr}[B \in \mathsf{Rect}(\theta)] + \frac{C\Delta \log^3 k}{\lambda^4} + g_k(\lambda).$$

Proceeding similarly for the function $\psi_L : \mathbb{R}^k \to \mathbb{R}$ obtained by applying Corollary 3.3.6 to $t - \lambda 1_k, \lambda, T$, we get

$$\mathsf{Pr}\left[A \in \mathsf{Rect}(\theta)\right] \geq \mathsf{Pr}\left[B \in \mathsf{Rect}(\theta)\right] - \frac{C\Delta \log^3 k}{\lambda^4} - g_k(\lambda).$$

Therefore,

$$\left|\mathsf{Pr}\left[A \in \mathsf{Rect}(\theta)\right] - \mathsf{Pr}\left[B \in \mathsf{Rect}(\theta)\right]\right| \leq \frac{C\Delta \log^3 k}{\lambda^4} + g_k(\lambda).$$

□

3.3.1.1 Discussion of the Result of Bentkus

We now give some intuition for the smooth approximation for ℓ_∞-norm result of Bentkus, Theorem 3.3.5.

Bentkus [17] obtains his smooth approximation function ψ as in Theorem 3.3.5 by first constructing an intermediary smooth function $\varphi \equiv \varphi_\lambda : \mathbb{R}^k \to \mathbb{R}$ that approximates the ℓ_∞-norm in the following sense:

- For every $x \in \mathbb{R}^k$,
$$|\varphi(x) - \|x\|_\infty| < \lambda. \tag{3.3.3}$$
- For every $r > 0$, there exists a constant $C(r)$ such that
$$\|\varphi^{(r)}\|_1 = \sup_{x \in \mathbb{R}^k} \{ \sum_{i_1,\ldots,i_r \in [k]} |\partial_{i_1}\partial_{i_2}\cdots\partial_{i_r}\varphi(x)| \} < \frac{C(r)\log^{r/2} k}{\lambda^r}. \tag{3.3.4}$$

Bentkus's construction of the function φ is quite ingenious: for a parameter ε to be chosen later define $\varphi : \mathbb{R}^k \to \mathbb{R}$ by

$$\varphi(x) = \mathop{\mathbb{E}}_{y \leftarrow \mathcal{N}^k}[\, \|x + \varepsilon y\|_\infty \,].$$

That is, $\varphi(x)$ is the expected ℓ_∞ norm of a Gaussian perturbation of *magnitude* ε of the input vector x. The intuition is that for ε small enough, the pointwise deviation would be small due to strong concentration properties of the Gaussian distribution. Moreover, the function φ should be sufficiently smooth for ε not too small, as we are averaging over a ball around every point x.

We can bound the point-wise deviation of $\varphi(x)$ from $\|x\|_\infty$ by using the following standard estimate on the tail of a random Gaussian distribution:

$$\Pr_{y \leftarrow \mathcal{N}^k}[\, \|y\|_\infty > t \,] \leq k \cdot \exp(-t^2/4).$$

The above estimate along with a simple integration implies that:

$$\mathop{\mathbb{E}}_{y \leftarrow \mathcal{N}^k}[\|y\|_\infty] = O(\sqrt{\log k}).$$

Therefore,

$$\varphi(x) = \mathop{\mathbb{E}}_{y \leftarrow \mathcal{N}^k}[\, \|x + \varepsilon y\|_\infty \,] = \|x\| \pm \varepsilon \mathop{\mathbb{E}}_{y \leftarrow \mathcal{N}^k}[\|y\|_\infty] = \|x\|_\infty \pm O(\varepsilon \sqrt{\log k}).$$

Thus, by choosing $\varepsilon = \Omega(\lambda/\sqrt{\log k})$ sufficiently small, φ can be made to satisfy the first property from Equation 3.3.3. We now need to ensure that φ also satisfies the second property, Equation 3.3.4. This step turns out to be quite nontrivial and Bentkus shows the property by carefully calculating the partial derivatives of the function φ. We refer the reader to Bentkus's paper for the proof of Equation 3.3.4.

We believe that the $(\sqrt{\log k})^r$ bound in Equation 3.3.4 is not a coincidence and there may be a connection to the Gaussian surface area of polytopes, which by Theorem 3.3.7 is at most $O(\sqrt{\log k})$. Unfortunately, we are not able to present any rigorous argument relating the two. Nevertheless, the plausible connection suggests that perhaps Bentkus's construction can be used to obtain smooth approximations for more general families of norms and convex sets with final error bounds depending only on the Gaussian surface area of corresponding sets.

3.3.2 Anti-concentration bound for l_∞-neighborhood of rectangles

Lemma 3.3.4 follows straightforwardly from the following result of Nazarov [79]. For a convex body $K \subseteq \mathbb{R}^n$ with boundary ∂K, let $\Gamma(K)$ denote the Gaussian surface area of K defined by

$$\Gamma(K) = \int_{y \in \partial K} e^{\frac{-\|y\|^2}{2}} \, d\sigma(y),$$

where $d\sigma(y)$ denotes the surface element at $y \in \partial K$.

Theorem 3.3.7 (Nazarov (see [60, Theorem 20]))**.** *For a polytope $\mathcal{K}$ with at most k faces, $\Gamma(\mathcal{K}) \leq C\sqrt{\log k}$.*

Proof of Lemma 3.3.4. Consider an increasing (under set inclusion) family of polytopes $\mathcal{K}_\rho$ for $0 \leq \rho \leq \lambda$ such that $\mathcal{K}_0 = \{x : W^T x \in \mathsf{Rect}(\theta - \lambda 1_k)\}$ and $\mathcal{K}_\lambda = \{x : W^T x \in \mathsf{Rect}(\theta)\}$. Then,

$$\Pr_{x \leftarrow \mathcal{N}^n} \left[W^T x \in \mathsf{Rect}(\theta) \setminus \mathsf{Rect}(\theta - \lambda 1_k) \right] = \int_{\rho=0}^{\lambda} \Gamma(\mathcal{K}_\rho) d\rho \leq C\sqrt{\log k}\, \lambda,$$

where the last inequality follows from Theorem 3.3.7. □

3.4 Invariance principle for Smooth Functions over Polytopes

We now prove Theorem 3.3.2. The proof of the theorem is based on the replacement method for proving limit theorems with explicit error bounds. Let $t = 1/\varepsilon$ and let $\mathcal{H} = \{h : [n] \to [t]\}$ be a family of $(2 \log k)$-wise independent functions as defined in Definition 2.3.1.

We remark that to prove Theorem 3.3.2 we could take the hash family to be the set of all functions. However, we work with a $(2\log k)$-wise independent family as the analysis is no more complicated and we need to work with such hash families while constructing pseudorandom generators. For $S \subseteq [n]$, let W_S be the matrix formed by the rows of W with indices in S. Define

$$\mathcal{H}(W) \stackrel{\text{def}}{=} \sum_{i=1}^{t} \left(\mathop{\mathbb{E}}_{h} \left[\sum_{p=1}^{k} \|W^p_{h^{-1}(i)}\|^{4\log k} \right] \right)^{1/\log k}.$$

Theorem 3.3.2 follows immediately from the following two lemmas.

Lemma 3.4.1. *For ε-regular W, $\mathcal{H}(W) \leq C\ \log k\,(\varepsilon\ \log(1/\varepsilon))$.*

Lemma 3.4.2. *For any smooth function $\psi : \mathbb{R}^k \to \mathbb{R}$,*

$$\left| \mathop{\mathbb{E}}_{X \leftarrow \mu^n} \left[\psi(W^T X)\right] - \mathop{\mathbb{E}}_{Y \leftarrow \mathcal{N}^n} \left[\psi(W^T Y)\right] \right| \leq 4\, c_\mu^2\, (\log^2 k)\, \mathcal{H}(W) \|\psi^{(4)}\|_1.$$

Proof of Lemma 3.4.1. Fix a $l \in [t]$, $p \in [k]$. For $i \in [n]$, let X_i be the indicator random variable that is 1 if $h(i) = l$ and 0 otherwise. Then, $\mathsf{Pr}[X_i = 1] = 1/t$ and the variables $X_1, \ldots, X_n$ are $(2\log k)$-wise independent. Further,

$$Z'_p \equiv \|W^p_{|h^{-1}(l)}\|^2 = \sum_{i=1}^{n} W^2_{ip} X_i.$$

Let Y_i be i.i.d indicator random variables with $\mathsf{Pr}[Y_i = 1] = 1/t$ and let $Z_p = \sum_{i=1}^n W^2_{ip} Y_i$. Observe that Z'_p and Z_p have identical d'th moments for $d \leq 2\log k$. Moreover, by Hoeffding's inequality applied to Z_p, for any $\gamma > 0$,

$$\mathsf{Pr}\left[\left| Z_p - \frac{1}{t} \right| \geq \gamma \right] \leq 2\exp\left(-\frac{2\gamma^2}{\sum_{i=1}^n W^4_{ip}} \right) \leq 2\exp\left(-\frac{2\gamma^2}{\varepsilon^2} \right) = 2\exp(-2t^2\gamma^2).$$

The above tail bound for Z_p implies strong bounds on the moments of Z_p by standard arguments. Setting $\gamma = \sqrt{2 \log k \, \log t}/t$ in the above equation, we get

$$\mathsf{Pr}\left[|Z_p| \geq \frac{\sqrt{3 \log k \, \log t}}{t} \right] \leq \frac{1}{t^{2 \log k}}.$$

Therefore, from the above equation and the fact that $Z_p \leq 1$

$$\begin{aligned} \mathbb{E}[Z_p^{2 \log k}] &\leq \frac{(3 \log k \, \log t)^{\log k}}{t^{2 \log k}} + \mathsf{Pr}\left[|Z_p| \geq \frac{\sqrt{3 \log k \, \log t}}{t} \right] \\ &\leq \frac{(4 \log k \, \log t)^{\log k}}{t^{2 \log k}}. \end{aligned}$$

Therefore,

$$\mathop{\mathbb{E}}_{h \in_u \mathcal{H}} \left[\|W^p_{|h^{-1}(l)}\|^{4 \log k} \right] = \mathbb{E}\left[(Z'_p)^{2 \log k} \right] = \mathbb{E}\left[Z_p^{2 \log k} \right] \leq \frac{(4 \log k \log t)^{\log k}}{t^{2 \log k}}.$$

Therefore, from the definition of $\mathcal{H}(W)$ and the above equation,

$$\mathcal{H}(W) = \sum_{i=1}^{t} \left(\sum_{p=1}^{k} \mathop{\mathbb{E}}_{h} \left[\|W^p_{h^{-1}(i)}\|^{4 \log k} \right] \right)^{1/\log k} \leq t \frac{4 \log k \log t}{t^2} = 4(\log k)(\varepsilon \log(1/\varepsilon)).$$

□

The proof of Lemma 3.4.2 uses a blockwise hybrid argument and careful applications of hypercontractivity as sketched in the proof outline in the introduction. To gain some intuition of the advantage of our randomized blockwise hybrid argument over the standard replacement method, it might be helpful to compare both arguments for the following cases:

Example 1: The bounding hyperplanes of $\mathcal{K}$ are oriented majorities: $W \in \{1/\sqrt{n}, -1/\sqrt{n}\}^{n\times k}$. In this case, the standard replacement method in conjunction with Bentkus's smoothing function and Nazarov's surface area bound as used in Lemmas 3.3.3, 3.3.4 can be adapted (without having to do a blockwise hybrid argument) to get a bound as in Theorem 3.3.1.

Example 2: The bounding hyperplanes of $\mathcal{K}$ are oriented majorities on disjoint sets of variables: For $m = n/k$ and each $p \in [k]$, $m = n/k$, $W_i^p = 1/\sqrt{m}, (p-1)m + 1 \leq i \leq pm$ and $W_i^p = 0$ otherwise. In this case, when $m \geq 1/\varepsilon^2$ and each bounding hyperplane is regular, the replacement method even when used with Lemmas 3.3.3, 3.3.4 leads to an error bound that is at least linear in k.

We use the following form of the standard Taylor series expansion. For a smooth function $\psi : \mathbb{R}^k \to \mathbb{R}$, $x \in \mathbb{R}^k$ and $p_1, \ldots, p_r \in [k]$, let $\partial_{p_1,\ldots,p_r}\psi(x) = \partial_{p_1}\partial_{p_2}\cdots\partial_{p_r}\psi(x)$. For indices $p_1, \ldots, p_r \in [k]$, let $(p_1, \ldots, p_r)! = s_1!s_2!\ldots s_k!$, where, for $l \in [k]$, s_l denotes the number of occurrences of l in $(p_1, \ldots, p_r)$.

Fact 3.4.3 (Multivariate Taylor's Theorem)**.** For any smooth function $\psi : \mathbb{R}^k \to \mathbb{R}$, and $x, y \in \mathbb{R}^k$,

$$\psi(x+y) = \psi(x) + \sum_{p\in[k]} \partial_p\psi(x)\, y_p + \sum_{p,q\in[k]} \frac{1}{(p,q)!}\, \partial_{p,q}\psi(x)\, y_p y_q + \sum_{p,q,r\in[k]} \frac{1}{(p,q,r)!}\, \partial_{p,q,r}\psi(x)\, y_p y_q y_r + \mathsf{err}(x,y),$$

where $|\mathsf{err}(x,y)| \leq \|\psi^{(4)}\|_1 \cdot \max_{p\in[k]} |y_p|^4$.

Proof of Lemma 3.4.2. Let $\overline{X} \leftarrow \mu^n$ and $\overline{Y} \leftarrow \mathcal{N}^n$. We first partition $[n]$ into blocks using a random hash function $h \in_u \mathcal{H}$ and then use a blockwise-hybrid argument. Fix a hash function $h \in \mathcal{H}$. View $\overline{X}$ as $X^1, \ldots, X^t$, where each $X^l = \overline{X}_{h^{-1}(l)}$ is chosen independently and uniformly from $\mu^{|h^{-1}(l)|}$. Similarly, view $\overline{Y}$ as $Y^1, \ldots, Y^t$ where each $Y^l = \overline{Y}_{h^{-1}(l)}$ is chosen independently and uniformly from $\mathcal{N}^{|h^{-1}(l)|}$. We prove the claim via a hybrid argument where we replace the blocks $X^1, \ldots, X^t$ with $Y^1, \ldots, Y^t$ one at a time.

For $0 \leq i \leq t$, let Z^i be the distribution with $Z^i_{|h^{-1}(j)} = X^j$ for $i < j \leq t$ and $Z^i_{|h^{-1}(j)} = Y^j$ for $1 \leq j \leq i$. Then, Z^0 is distributed as μ^n and Z^t is distributed as $\mathcal{N}^n$. For $l \in [t]$, let

$$h(W, l) = \left(\sum_{p=1}^{k} \|W^p_{h^{-1}(l)}\|^{4 \log k} \right)^{1/\log k}.$$

Claim 3.4.4. *For $1 \leq l \leq t$, and fixed $h \in \mathcal{H}$,*

$$\left| \mathop{\mathbb{E}}_{\overline{X},\overline{Y}} \left[\psi(W^T Z^l) \right] - \mathop{\mathbb{E}}_{\overline{X},\overline{Y}} \left[\psi(W^T Z^{l-1}) \right] \right| \leq C\, c_\mu \, \log^2 k \, \|\psi^{(4)}\|_1 \, h(W, l).$$

Proof. Without loss of generality, suppose that $h^{-1}(l) = \{1, \ldots, m\}$. Note that Z^l, Z^{l-1} have the same random variables in positions $m+1, \ldots, n$. Let $Z^{l-1} = (X_1, \ldots, X_m, Z_{m+1}, \ldots, Z_n)$ and $Z^l = (Y_1, \ldots, Y_m, Z_{m+1}, \ldots, Z_n)$ where $(X_1, \ldots, X_m)$ is uniform over μ^m and $(Y_1, \ldots, Y_m)$ is uniform over $\mathcal{N}^m$. Note that $(Z_{m+1}, \ldots, Z_n)$ is independent of $(X_1, \ldots, X_m)$, $(Y_1, \ldots, Y_m)$.

Let $W_1 \in \mathbb{R}^{m \times k}$ be the matrix formed by the first m rows of W and similarly let $W_2 \in \mathbb{R}^{(n-m) \times k}$ be the matrix formed by the last $n - m$ rows of

W. Lastly, let $V = W_2^T(Z_{m+1}, \ldots, Z_n)$ and U be one of $X = (X_1, \ldots, X_m)$ or $Y = (Y_1, \ldots, Y_m)$. Now, by using a Taylor expansion of ψ at V as in Fact 3.4.3,

$$\begin{aligned}
&\psi(W^T(U_1, \ldots, U_m, Z_{m+1}, \ldots, Z_n)) = \psi(\, W_1^T U + V\,) \\
&\quad = \psi(V) + \sum_{p\in[k]} \partial_p \psi(V)\, \langle W_1^p, U\rangle + \sum_{p,q\in[k]} \frac{1}{(p,q)!}\, \partial_{p,q}\psi(V)\, \langle W_1^p, U\rangle\, \langle W_1^q, U\rangle \\
&\qquad + \sum_{p,q,r\in[k]} \frac{1}{(p,q,r)!}\, \partial_{p,q,r}\psi(V)\, \langle W_1^p, U\rangle\, \langle W_1^q, U\rangle\, \langle W_1^r, U\rangle + \mathsf{err}(V, W_1^T U).
\end{aligned} \tag{3.4.1}$$

Now, using the fact that $\|z\|_\infty \le \|z\|_{\log k}$ for $z \in \mathbb{R}^k$,

$$\left|\mathsf{err}(V, W_1^T U)\right| \le \|\psi^{(4)}\|_1 \cdot \max_{p\in[k]} |\langle W_1^p, U\rangle|^4 \le \|\psi^{(4)}\|_1 \left(\sum_{p=1}^k |\langle W_1^p, U\rangle|^{4\log k}\right)^{1/\log k}. \tag{3.4.2}$$

Now, by hypercontractivity of μ,

$$\begin{aligned}
\mathop{\mathbb{E}}_X \left[\left(\sum_{p=1}^k |\langle W_1^p, X\rangle|^{4\log k}\right)^{1/\log k}\right] &\le \left(\mathop{\mathbb{E}}_X \left[\sum_{p=1}^k |\langle W_1^p, X\rangle|^{4\log k}\right]\right)^{1/\log k} \\
&\qquad \text{(by power-mean inequality)} \\
&= \left(\sum_{p=1}^k \mathop{\mathbb{E}}_X \left[|\langle W_1^p, X\rangle|^{4\log k}\right]\right)^{1/\log k} \\
&\le \left(\sum_{p=1}^k (c_\mu \log k)^{2\log k}\, \|W_1^p\|^{4\log k}\right)^{1/\log k} \\
&\qquad \text{(by hypercontractivity of } \mu) \\
&\le C c_\mu^2 (\log^2 k)\, h(W, l).
\end{aligned} \tag{3.4.3}$$

Similarly, by hypercontractivity of $\mathcal{N}$,

$$\mathop{\mathbb{E}}_{Y}\left[\left(\sum_{p=1}^{k}|\langle W_1^p, Y\rangle|^{4\log k}\right)^{1/\log k}\right] \leq C(\log^2 k)\, h(W,l). \tag{3.4.4}$$

Since μ is proper, for any $u^1, u^2, u^3 \in \mathbb{R}^m$,

$$\mathbb{E}[\langle u^1, X\rangle] = \mathbb{E}[\langle u^1, Y\rangle], \quad \mathbb{E}[\langle u^1, X\rangle \langle u^2, X\rangle] = \mathbb{E}[\langle u^1, Y\rangle \langle u^2, Y\rangle]$$

$$\mathbb{E}[\langle u^1, X\rangle \langle u^2, X\rangle \langle u^3, X\rangle] = \mathbb{E}[\langle u^1, Y\rangle \langle u^2, Y\rangle \langle u^3, Y\rangle].$$

From the above equations, Equations (3.4.1), (3.4.2), (3.4.3), (3.4.4) and the fact that X, Y, V are independent of one another, it follows that

$$\left|\mathbb{E}\left[\psi(W^T Z^l) - \psi(W^T Z^{l-1})\right]\right| \leq C c_\mu^2 (\log^2 k)\, \|\psi^{(4)}\|_1 h(W,l).$$

□

Lemma 3.4.2 now follows from the above claim, summing from $l = 1, \ldots, t$, and taking expectation with respect to $h \in_u \mathcal{H}$. □

Chapter 4

Discrete Central Limit Theorems

4.1 Introduction

The classical Central Limit Theorem (CLT) says that a sum of independent random variables should be close, in Kolmogorov distance, to the corresponding Gaussian or Binomial random variable. The Kolmogorov distance is weaker than statistical (total variation) distance $\mathsf{d_{TV}}$, since Kolmogorov distance allows only special types of statistical tests, namely threshold functions. Nevertheless, if the random variables are integer-valued, then under some reasonable conditions it is known that a sum of independent variables approaches the appropriate binomial distribution in statistical distance. Such theorems are called *discrete central limit theorems*.

For clarity, We first state our discrete central limit theorem for the case of multinomial distributions.

Theorem 4.1.1. *Let $X_1, \ldots, X_n$ be independent indicator random variables with $\mathsf{Pr}[X_i = 1] = p_i$. Let $X = \sum_i X_i$, $\mathbb{E}[X] = \mu, \mathsf{Var}(X) = \sum_i p_i(1-p_i) = \sigma^2$. Then, for $Z \leftarrow \mathsf{BIN}(m, q)$, where $m = \mu^2/(\mu - \sigma^2)$, $q = (\mu - \sigma^2)/\mu$, $\mathsf{d_{TV}}(X, Z) = O\left(\sqrt{\log(\sigma)/\sigma}\right)$.*

The parameters m, q above are chosen so that $\mathbb{E}[Z] = \mathbb{E}[X]$ and $\mathsf{Var}[Z] =$

$\mathsf{Var}[X]$. (A similar statement holds with $m = \lfloor \mu^2/(\mu - \sigma^2) \rfloor$ and $q = \mu/m$; we avoid this minor technicality.) Limit theorems as above with almost optimal error estimates ($\Theta(1/\sigma)$) are known in the probability literature (see [8, 7] and references therein). However, most previous results use Fourier techniques or Stein's method and appear more complicated, at least from a computer science perspective. In contrast our proof is elementary, relying only on the classical Berry-Esséen theorem and few simple properties of the binomial distribution. We also obtain a more general *invariance principle*, Theorem 4.4.2, for the case of sums of integer-valued random variables.

Discrete central limit theorems as above have, at least implicitly, been used before in computer science. Two prominent instances are the works of Daskalakis and Papadimitriou [24, 25]. A main technical result in these works can be viewed as a discrete limit theorem and roughly says the following: given a multinomial distribution (or more generally, a multivariate-multinomial distribution), the probabilities of each of the indicator variables can be rounded to multiples of a parameter $1/\varepsilon$, so as to not incur too much of a loss in statistical distance. Their arguments for showing the discrete CLT are quite involved and use a variety of sampling and Poisson approximation techniques. Given the generality of our argument for proving Theorem 4.1.1, it is conceivable that a similar argument can be extended to the more nuanced discrete limit theorems of [24, 25].

4.2 Preliminaries

We start with some basic properties of the binomial and multi-nomial distributions.

Corollary 4.2.1 (Berry-Esséen for Multinomials)**.** *For $Y = \sum_i Y_i$ a sum of independent indicator variables, $Z \leftarrow \mathbb{N}(0,1)$,*

$$\mathsf{d}_{\mathsf{cdf}}\left((Y - \mathbb{E}(Y))/\sigma(Y), Z \right) \leq 1/\sigma(Y).$$

Proof. Follows from Theorem 2.2.4, as for $0,1$ valued Y_i, $\sum_i \mathbb{E}[|Y_i - \mathbb{E}[Y_i]|^4] \leq \sum_i \mathbb{E}[|Y_i - \mathbb{E}[Y_i]|^2]$. □

Fact 4.2.2. For $Z_1 \leftarrow \mathbb{N}(\mu_1, \sigma_1)$, $Z_2 \leftarrow \mathbb{N}(\mu_2, \sigma_2)$, for $\sigma_1 \geq 1$,

$$\mathsf{d}_{\mathsf{cdf}}(Z_1, Z_2) = O\left(\frac{|\mu_1 - \mu_2|}{\sigma_1} + \frac{\sqrt{|\sigma_1^2 - \sigma_2^2| \log(\sigma_1)}}{\sigma_1} \right).$$

Proof. We'll use the following anti-concentration property of Gaussians: for $Z \leftarrow \mathbb{N}(0, \sigma)$, $\delta > 0$, $\mathsf{Pr}[Z \in [\theta, \theta + \delta]] = O(\delta/\sigma)$. Suppose that $\mu_2 > \mu_1$. Then,

$$\mathsf{d}_{\mathsf{cdf}}(Z_1, \mathbb{N}(\mu_2, \sigma_1)) \leq \mathsf{Pr}[Z_1 \in [\mu_1, \mu_2]] = O(|\mu_2 - \mu_1|/\sigma_1).$$

Thus, it suffices to study the case when $\mu_1 = \mu_2 = 0$. Let $\sigma_2 > \sigma_1$ and $\lambda = \sqrt{\sigma_2^2 - \sigma_1^2}$. Observe that Z_2 can be generated as $Z_2 = Z_1 + Z'$, where Z' is an independent $\mathbb{N}(0, \lambda)$ random variable. Now, $\mathsf{Pr}[|Z'| > 3\lambda\sqrt{\log \sigma_1}] \leq 1/\sigma_1$.

Therefore, for any $\theta \in \mathbb{R}$,

$$\begin{aligned}
\mathsf{Pr}[Z_2 < \theta] &= \mathsf{Pr}[Z_1 + Z' < \theta] \\
&\leq \mathsf{Pr}[Z_1 < \theta + 3\lambda\sqrt{\log \sigma_1}] + \mathsf{Pr}[|Z'| > 3\lambda\sqrt{\log \sigma_1}] \\
&\leq \mathsf{Pr}[Z_1 < \theta] + \mathsf{Pr}[Z_1 \in [\theta, \theta + 3\lambda\sqrt{\log \sigma_1}] + 1/\sigma_1 \\
&\leq \mathsf{Pr}[Z_1 < \theta] + O(3\lambda\sqrt{\log \sigma_1}/\sigma_1) + 1/\sigma_1.
\end{aligned}$$

The claim now follows from a similar argument by starting from Z_1 instead of Z_2. □

Fact 4.2.3. Any multinomial distribution X with $\mathsf{Var}(X) = \sigma^2$ is $(2/\sigma)$-shift invariant.

Proof. A simple induction shows that multinomial distributions are unimodal, with the density function being maximized either at a unique value j or at j and $j+1$. For this value j, it holds that $\mathsf{d_{TV}}(X, X+1) = \mathsf{Pr}[X \leq j] - \mathsf{Pr}[X+1 \leq j] = \mathsf{Pr}[X = j]$. We now use the anti-concentration of X, which follows from the Berry-Esséen theorem. Indeed by Theorem 2.2.4, if $Z \leftarrow \mathbb{N}(0,1)$, then $\mathsf{Pr}[X = j] \leq \mathsf{Pr}\left[Z = (j-\mu)/\sigma\right] + 2/\sigma = 2/\sigma$. □

Finally we use the following inequality that follows, for instance, from Bernstein's large deviation bound.

Fact 4.2.4. For any multinomial distribution X, and $\delta > 0$, $\mathsf{Pr}[\ |X - \mathbb{E}[X]| \geq 3\sigma(X)\sqrt{\log(1/\delta)}\ ] \leq \delta$.

4.3 Main Convolution Lemma

We say that a random variable Y is α-shift invariant if $\mathsf{d_{TV}}(Y, Y+1) \leq \alpha$. Several common distributions, such as binomial, Gaussian, and multinomial distributions, are all shift-invariant, roughly, inversely proportional to their standard deviation.

The starting point for our results is the following lemma, which says that two distributions that are close in Kolmogorov distance when convolved with a shift-invariant distribution become close in statistical distance. The lemma also plays a central role in our construction of a PRG for combinatorial shapes from Chapter 9.

Lemma 4.3.1 (Main Convolution Lemma)**.** *Let X be a α-shift invariant distribution and let Y, Z be integer-valued distributions with support contained in $[a, a+b]$ for some $a \in \mathbb{R}, b > 0 \in \mathbb{R}$. Then,*

$$\mathsf{d_{TV}}(X+Y, X+Z) \leq 4\sqrt{\alpha b \mathsf{d_{cdf}}(Y, Z)}.$$

Proof. Without loss of generality suppose that Y, Z are supported in $[0, b)$. For $d \in \mathbb{Z}_+$ to be chosen later, let Y_d be the integer random variable with support over $S_d = \{id : i \in \mathbb{Z}_+, i \leq \lfloor b/d \rfloor\}$, with pdf p_d defined by, $p_d(id) = \mathsf{Pr}[Y \in [id, (i+1)d)]$. We first show that

$$\mathsf{d_{TV}}(X+Y, X+Y_d) \leq \alpha d. \tag{4.3.1}$$

There is a natural coupling of Y and Y_d: we set $Y_d = id$ with probability $p_d(id)$ and then sample $Y = Y_d + \bar{Y}$ from the interval $[id, (i+1)d)$ according to the

marginal distribution of Y conditioned on the event that $Y \in [id, (i+1)d)$. Note that $\bar{Y} \in \{0, 1, \ldots, d-1\}$ and it is an integer. We have

$$\mathsf{d_{TV}}(X + Y, X + Y_d) = \mathsf{d_{TV}}(X + Y_d + \bar{Y}, X + Y_d).$$

Further, conditioned on a particular value of $Y_d = id$,

$$\mathsf{d_{TV}}(X + Y_d + \bar{Y}, X + Y_d) = \mathsf{d_{TV}}(X + \bar{Y}, X) \leq \alpha d,$$

where the last inequality follows from the shift invariance of X and the fact that $\bar{Y} \in \{0, \ldots, d-1\}$. Therefore,

$$\mathsf{d_{TV}}(X + Y, X + Y_d) = \mathsf{d_{TV}}(X + Y_d + \bar{Y}, X + Y_d) \leq \alpha d.$$

We define Z_d similarly. It follows that $\mathsf{d_{TV}}(X + Z, X + Z_d) \leq \alpha d$. Next we bound $\mathsf{d_{TV}}(Y_d, Z_d)$.

Observe that Y_d, Z_d both have supports of size at most b/d. For any i,

$$|\ \mathsf{Pr}[Y_d = id] - \mathsf{Pr}[Z_d = id]\ | = |\ \mathsf{Pr}[Y \in [id, (i+1)d)\,] - \mathsf{Pr}[Z \in [id, (i+1)d)\,]\ | \leq 2\mathsf{d_{cdf}}(Y, Z).$$

Hence $\mathsf{d_{TV}}(Y_d, Z_d) \leq (2b/d)\mathsf{d_{cdf}}(Y, Z)$. Combining the above equations,

$$\begin{aligned}\mathsf{d_{TV}}(X + Y, X + Z) &\leq \mathsf{d_{TV}}(X + Y, X + Y_d) + \mathsf{d_{TV}}(X + Y_d, X + Z_d) + \\ &\qquad \mathsf{d_{TV}}(X + Z_d, X + Z) \\ &\leq 2\alpha d + \frac{2b\mathsf{d_{cdf}}(Y, Z)}{d}.\end{aligned}$$

The lemma now follows by setting $d = \lceil \sqrt{b\mathsf{d_{cdf}}(Y, Z)/\alpha} \rceil$. □

One can weaken the boundedness requirement to say that Y and Z rarely exceed b. We record the following easy corollary without proof.

Corollary 4.3.2. *Let X be a α-shift invariant distribution and let Y, Z be two integer-valued distributions. Then, for $a \in R$ and $b \in \mathbb{R}^+$*

$$\mathsf{d_{TV}}(X+Y, X+Z) \leq 4\sqrt{\alpha b \mathsf{d_{cdf}}(Y,Z)} + \mathsf{Pr}[Y \notin [a, a+b)] + \mathsf{Pr}[Z \notin [a, a+b)].$$

4.4 Discrete Central Limit Theorems

We now prove the discrete central limit theorem, Theorem 4.1.1. The proof proceeds by partitioning the variables appropriately and using the convolution lemma. We partition the variables into two sets S and T such that $X_S = \sum_{i \in S} X_i$ and $X_T = \sum_{j \in T} X_i$ have approximately the same mean and variance. We introduce variables Y_S and Y_T which are two independent copies of $\mathsf{BIN}(m/2, q)$. Then, the Berry-Esséen theorem, which is a quantitative form of the classical central limit theorem, guarantees the closeness of X_S, Y_S and X_T, Y_T in Kolmogorov distance. Secondly, multinomial distributions are shift-invariant. Hence we bound the statistical distance between $X_S + X_T$ and $Y_S + Y_T$, by using our Convolution lemma to show that each of them is close to $X_S + Y_T$ in statistical distance.

The following easy fact (whose proof we omit) is used to partition the variables.

Fact 4.4.1. Let $0 \leq a_1 \leq \cdots \leq a_n \leq 1$. Let $S \subset [n]$ consist of all odd indices. Then $|\sum_{i \in S} a_i - (\sum_j a_j)/2| \leq a_n/2$.

Proof of Theorem 4.1.1. Without loss of generality suppose that $\sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_n$, where $\sigma_i = \sigma(X_i)$. Let S and T consist of odd and even indices respectively. Let $X_S = \sum_{i \in S}(X_i - \mathbb{E}[X_i])$ and $X_T = \sum_{i \in T}(X_i - \mathbb{E}[X_i])$. Let $\sigma_S^2 = \mathsf{Var}(X_S)$. Then, from Fact 4.4.1 $|\, \sigma_S^2 - \sigma^2/2 \,| \leq 1/2$.

Let Y_S, Y_T denote two independent copies of $(\mathsf{BIN}(m/2, q) - \mu/2)$ for m, q as in the theorem statement. Note that $Y_S + Y_T$ has distribution $\mathsf{BIN}(m, q) - \mu$ and that $\mathbb{E}[Y_S] = \mathbb{E}[Y_T] = 0$ and $\mathsf{Var}(Y_S) = \mathsf{Var}(Y_T) = \sigma^2/2$.

We proceed to bound the various quantities (α, B and $\mathsf{d_{cdf}}$) required to apply the convolution lemma. By Fact 4.2.3, X_S, Y_S, X_T, Y_T are all $\alpha = (2/\sigma)$-shift invariant. By Theorem 2.2.4 and Fact 4.2.2,

$$\mathsf{d_{cdf}}(X_S, Y_S) \leq \mathsf{d_{cdf}}(X_S, \mathbb{N}(0, \sigma_S^2)) + \mathsf{d_{cdf}}(Y_S, \mathbb{N}(0, \sigma^2/2)) + \mathsf{d_{cdf}}(\mathbb{N}(0, \sigma_S^2), \mathbb{N}(0, \sigma^2/2)) \tag{4.4.1}$$

$$\leq \frac{1}{\sigma} + \frac{1}{\sigma} + O\left(\frac{\sqrt{\log(\sigma)}}{\sigma}\right) = O\left(\frac{\sqrt{\log(\sigma)}}{\sigma}\right). \tag{4.4.2}$$

A similar bound holds for $\mathsf{d_{cdf}}(X_T, Y_T)$.

Next we show that X_S, X_T, Y_S, Y_T are bounded in a range $[-B, B]$ with probability $(1 - 1/\sigma)$. By Fact 4.2.4, for $B = 12(\sigma\sqrt{\log \sigma})$, $\mathsf{Pr}[\,|X_S| > B\,] \leq 1/4\sigma$, and a similar statement holds for X_T, Y_S, Y_T. We then apply the union bound. Therefore, applying Corollary 4.3.2,

$$\begin{aligned}
\mathsf{d_{TV}}(X_S + X_T, Y_S + Y_T) &\leq \mathsf{d_{TV}}(X_S + X_T, X_S + Y_T) + \mathsf{d_{TV}}(X_S + Y_T, Y_S + Y_T) \\
&\leq 4\sqrt{\alpha B \mathsf{d_{cdf}}(X_T, Y_T)} + 4\sqrt{\alpha B \mathsf{d_{cdf}}(X_S, Y_S)} + \frac{1}{\sigma} \\
&= O\left(\sqrt{\log(\sigma)/\sigma}\right). \quad \text{(By Equation 7.2.1)}
\end{aligned}$$

□

We next generalize Theorem 4.1.1 to sums of independent integer-valued variables (as opposed to indicator random variables). The error term in the statistical distance guarantee we get depends on the Kolmogorov distance guarantee given by the Berry-Esséen theorem and on the shift invariance of the individual random variables. The dependence on these terms is in some sense unavoidable (as explained below). As for the case of indicator random variables our bound is weaker those those of the more fine-grained results of [8, 7]. However, the arguments and exact technical conditions of [8, 7] are complicated and the parameters we get are comparable up to $\Omega(1)$ factors in the exponents.

Theorem 4.4.2. *Let $\bar{X} = (X_1, \ldots, X_n), \bar{Y} = (Y_1, \ldots, Y_m)$ be two sets of independent integer-valued variables. Let $X = \sum_i X_i, Y = \sum_i Y_i$ and let $\mathbb{E}[X] = \mathbb{E}[Y]$, $\sigma^2 = \mathsf{Var}(X) = \mathsf{Var}(Y)$. Further, let*

$$\max_i\{\mathsf{Var}(X_i), \mathsf{Var}(Y_i)\} \leq \sigma^2/2, \quad \max(\sum_i \mathbb{E}[\,|X_i - \mathbb{E}[X_i]|^3\,], \sum_i \mathbb{E}[\,|Y_i - \mathbb{E}[Y_i]|^3\,]) \leq \rho,$$

$$4 \leq U = \min(\sum_i (1 - \mathsf{d_{TV}}(X_i, X_i + 1)), \sum_j (1 - \mathsf{d_{TV}}(Y_j, Y_j + 1))\,).$$

Then,

$$\mathsf{d_{TV}}(X, Y) = O\left(\left(\frac{\rho \log(1/\sigma)}{\sigma^2 U^{1/2}}\right)^{1/2} + \frac{\rho}{\sigma^3} + \frac{1}{\sigma}\right).$$

Note that for a limit theorem as above to hold, we need assumptions on X, Y stronger than matching means and variances which was enough for the

Berry-Esséen theorem. For instance, the X_i's could be supported on even integers and Y_i's on odd integers with X, Y having the same mean and variances. In this case the statistical distance between X, Y is 1, whereas the Kolmogorov distance could still be small. Thus, the additional assumption that X_i's, Y_i's have some shift-invariance is a natural restriction to have.

We use the following tricky generalization of Fact 4.4.1 whose proof uses Hall's theorem.

Lemma 4.4.3. *Given $a_1, \ldots, a_n > 0$ and $b_1, \ldots, b_n > 0$, there exists a set $S \subseteq [n]$ such that*

$$|\sum_{i \in S} a_i - \frac{\sum_j a_j}{2}| \leq \frac{\max_i a_i - \min_i a_i}{2}, \quad |\sum_{i \in S} b_i - \frac{\sum_j b_j}{2}| \leq \frac{\max_i b_i - \min_i b_i}{2}.$$

Proof. Let n be even, the case of n odd is similar. Let $A = \sum_i a_i, B = \sum_i b_i$. Suppose that $a_1 \leq a_2 \leq \cdots a_n$ and let $\pi : [n] \to [n]$ be such that $b_{\pi(1)} \leq b_{\pi(2)} \leq \cdots \leq b_{\pi(n)}$. Form a bipartite graph $G = (L, R, E)$, where $|L| = |R| = [n/2]$ with vertices on left corresponding to pairs $\{(a_1, a_2), (a_3, a_4), \ldots, (a_{n-1}, a_n)\}$ and vertices on right corresponding to $\{(b_{\pi(2i-1)}, b_{\pi(2i)}) : i \in [n/2]\}$. Finally, add an edge in G between vertices (a_i, a_{i+1}) and $(b_{\pi(j)}, b_{\pi(j+1)})$ if and only if $\{i, i+1\} \cap \{\pi(j), \pi(j+1)\} \neq \emptyset$.

Observe that G is a 2-regular graph and hence by Hall's theorem there exists perfect matching M in G. For each $i \in [n/2]$, let M connect vertex $(a_{2i-1}, a_{2i}) \in L$ to a vertex $(b_j, b_{j'}) \in R$ so that index $r_i \in \{2i-1, 2i\} \cap \{j, j'\}$. Let $S = \{r_i : i \in [n/2]\}$. We claim that S satisfies the required properties.

Note that

$$A_o = \sum_{i \in [n/2]} a_{2i-1} \leq \sum_{i \in [n/2]} a_{r_i} \leq \sum_{i \in [n/2]} a_{2i} = A_e.$$

Further, $A_e - A_o \leq a_n - a_1$. Thus,

$$\frac{A - (a_n - a_1)}{2} \leq A_o \leq \sum_i a_{r_i} \leq A_e \leq \frac{A + (a_n - a_1)}{2}.$$

The lemma now follows by a similar argument applied to b_{r_i} for $i \in [n/2]$. □

We also use the following elegant lemma of Barbour and Xia [8] which they show using an elementary coupling argument. Intuitively, the lemma says that shift-invariance *amplifies* when taking sums of independent shift-invariant variables.

Lemma 4.4.4 (Barbour and Xia, Proposition 4.6)**.** *Let $Z_1, \ldots, Z_n$ be integer valued random variables, $Z = \sum_i Z_i$ and $U_Z = \sum_i (1 - \mathsf{d_{TV}}(Z_i, Z_i + 1))$. Then $\mathsf{d_{TV}}(Z, Z+1) \leq 2/\sqrt{U_Z}$.*

Proof of Theorem 4.4.2. Let $\nu = \max_i(\mathsf{Var}(X_i), \mathsf{Var}(Y_i))$. Let $U_X = \sum_i (1 - \mathsf{d_{TV}}(X_i, X_i+1))$ and let U_Y be defined similarly. Now, by Lemma 4.4.3 applied to $Var(X_1), \ldots, Var(X_n)$ and $(1-\mathsf{d_{TV}}(X_1, X_1+1)), \ldots, (1-\mathsf{d_{TV}}(X_n, X_n+1))$, there exists a subset $S \subseteq [n]$ such that

$$|\sum_{i \in S} Var(X_i) - \frac{\sigma^2}{2}| \leq \frac{\nu}{2}, \quad |\sum_{i \in S} (1 - \mathsf{d_{TV}}(X_i, X_i + 1)) - \frac{U_X}{2}| \leq \frac{1}{2}.$$

Similarly, there exists a subset $T \subseteq [n]$ such that

$$|\sum_{i \in T} Var(Y_i) - \frac{\sigma^2}{2}| \leq \frac{\nu}{2}, \quad |\sum_{i \in T} (1 - \mathsf{d_{TV}}(Y_i, Y_i + 1)) - \frac{U_Y}{2}| \leq \frac{1}{2}.$$

Let $X_S = \sum_{i \in S} X_i$, $X'_S = \sum_{i \notin S} X_i$ and let Y_T, Y'_T be defined similarly. Without loss of generality suppose that $\mathbb{E}[X_S] = \mathbb{E}[Y_T] = \mathbb{E}[X'_S] = \mathbb{E}[Y'_T] = 0$ (if not, we can translate the variables accordingly). Then, by the above equations and Lemma 4.4.4 it follows that X_S, X'_S, Y_T, Y'_T are α-shift invariant for $\alpha = 4/\sqrt{U}$.

Let $\delta = \rho/(\sigma^2 - \nu)^{3/2}$. Now, by an argument similar to that of Equation 7.2.1 and the Berry-Esséen theorem,

$$\begin{aligned}
\mathsf{d}_{\mathsf{cdf}}(X_S, Y_T - \mathbb{E}[Y_T]) &\leq \mathsf{d}_{\mathsf{cdf}}(X_S, \mathcal{N}(0, Var(X_S))) + \mathsf{d}_{\mathsf{cdf}}(Y_T, \mathcal{N}(0, Var(Y_T))) \\
&\quad + \mathsf{d}_{\mathsf{cdf}}(\mathcal{N}(0, Var(X_S)), \mathcal{N}(0, Var(Y_T))) \\
&\leq \frac{2\rho}{(\sigma^2 - \nu)^{3/2}} + \frac{2\rho}{(\sigma^2 - \nu)^{3/2}} + O\left(\frac{\sqrt{\log \sigma}}{\sigma}\right). \\
&\leq 4\delta + O\left(\frac{\sqrt{\log \sigma}}{\sigma}\right).
\end{aligned}$$

Now, by the Berry-Esséen theorem, for $B = O(\sigma\sqrt{\log(\sigma)})$,

$$\mathsf{Pr}[|X_S - \mathbb{E}[X_S]| > B] \leq 2\delta + 1/\sigma, \quad \mathsf{Pr}[|Y_T - \mathbb{E}[Y_T| > B] \leq 2\delta + 1/\sigma.$$

Further, similar inequalities hold for X'_S, Y'_T as well. Therefore, by Corollary 4.3.2, and the above inequalities,

$$\begin{aligned}
\mathsf{d}_{\mathsf{TV}}(X_S + X'_S, Y_T + Y'_T) &\leq \mathsf{d}_{\mathsf{TV}}(X_S + X'_S, X_S + Y'_T) + \mathsf{d}_{\mathsf{TV}}(X_S + Y'_T, Y_T + Y'_T) \\
&\leq 4\sqrt{\alpha B \mathsf{d}_{\mathsf{cdf}}(X'_S, Y'_T)} + 4\sqrt{\alpha B \mathsf{d}_{\mathsf{cdf}}(X_S, Y_T)} + \\
&\qquad O(\delta) + O(1/\sigma) \\
&= O\left(\frac{\sigma \log(1/\sigma)\rho}{(\sigma^2 - \nu)^{3/2} U^{1/2}}\right)^{1/2} + O(\delta) + O(1/\sigma).
\end{aligned}$$

The theorem now follows as $\rho \leq \sigma^2/2$. □

Chapter 5

Gotsman-Linial Conjecture and Random Restrictions of PTFs

5.1 Introduction

As mentioned in the introduction, average and noise sensitivity are two fundamental notions in the analysis of Boolean functions with a variety of important applications. In this chapter we study the sensitivity of threshold functions and give the first nontrivial bounds for low-degree PTFs. We then use our results along with existing learning theory machinery to get better agnostic learning algorithms for these classes. In spirit of most of our results, the results in this chapter will essentially be obtained by using the invariance principles for polynomials of Mossel et al. to translate the problem to the Gaussian setting and solve the Gaussian problem directly.

Recall the definition of noise sensitivity and average sensitivity from Section 2.4. It is well known (and easy to prove via elementary methods) that the noise sensitivity of linear threshold functions is at most $O(\sqrt{\delta})$ [85]. On the other hand, we do not know of any nontrivial bounds on the noise sensitivity and average sensitivity for degrees 2 and higher.

Our results give the first nontrivial bounds for degrees 2 and higher.

Theorem 5.1.1 (Boolean Noise Sensitivity). *For any degree d PTF* $f : \{1,-1\}^n \to \{1,-1\}$ *and* $\delta \in (0,1)$,

$$\mathbb{NS}_\delta(f) = O_d\left(\delta^{1/(4d+6)}\right).$$

The subscript d in the $O_d(\cdot)$ notation indicates that the hidden constant depends on d.

Our next set of results bound the average sensitivity of degree d PTFs. Clearly, for any function f, $\mathbb{AS}(f)$ is at most n (the parity function on n variables achieves this bound). It is well known that the average sensitivity of linear threshold functions is $O(\sqrt{n})$ (the Majority function has average sensitivity $\Theta(\sqrt{n})$).

In 1994, Gotsman and Linial [39] conjectured that the average sensitivity of any degree d polynomial threshold function f is $O(d\sqrt{n})$. We are not aware of any progress on this conjecture until now.

We give two upper bounds on the average sensitivity of degree d PTFs. We first use a simple translation lemma for bounding average sensitivity in terms of noise sensitivity of a Boolean function and Theorem 5.1.1 to obtain the following bound.

Theorem 5.1.2 (average sensitivity). *For a degree d PTF* $f : \{1,-1\}^n \to \{1,-1\}$, $\mathbb{AS}(f) = 2^{O(d)}\,(n^{1-1/(4d+6)})$.

We also give an elementary combinatorial argument, to show that the average sensitivity of any degree d PTF is at most $3n^{1-1/2^d}$. The combi-

natorial proof is based on the following lemma for general Boolean functions that may prove useful elsewhere. For $x \in \{1,-1\}^n$, and $i \in [n]$, let $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

Lemma 5.1.3. *For Boolean functions $f_i : \{1,-1\}^n \to \{1,-1\}$ with f_i not depending on the i'th coordinate x_i, and $X \in_u \{1,-1\}^n$, $\mathbb{E}_X[\,|\sum_i X_i f_i(X_{-i})\,|]^2 \leq 2\sum_i \mathbb{AS}(f_i) + n$.*

We believe that when the functions f_i in the above lemma are LTFs, the above bound can be improved to $O(n)$, which in turn would imply the Gotsman-Linial conjecture for quadratic threshold functions.

5.1.1 Random Restrictions of PTFs – a structural result

An important ingredient of our sensitivity bounds for PTFs are new structural theorems about random restrictions of PTFs obtained via *hypercontractivity*. The structural results we obtain can be seen as part of the high level "randomness vs structure" paradigm that has played a fundamental role in many recent breakthroughs in additive number theory and combinatorics. Specifically, we obtain the following structural result (Theorem 5.3.4): arbitrary low-degree PTFs can be approximated by small depth decision trees in which the leaf nodes either compute a regular PTF or a function with high bias.

We remark that our structural results, though motivated by similar results of Servedio [92] and Diakonikolas et al. [26] for the simpler case of

LTFs, do not follow from a generalization of their arguments for LTFs to PTFs. The structural results for random restrictions of low-degree PTFs provide a reasonably generic template for reducing problems involving arbitrary PTFs to ones on regular PTFs. In fact, these structural properties are used precisely for the above reason both in this work and in our construction of pseudorandom generators for PTFs, Chapter 7.

5.1.2 Related Work

Independent of this work, Diakonikolas, Raghavendra, Servedio, and Tan [29] have obtained nearly identical results to ours for both the average and noise sensitivity of PTFs. The broad outline of their proof is also similar to ours.

Regarding our structural result described in Section 5.1.1, Diakonikolas, Servedio, Tan and Wan [30] have independently obtained similar results to ours. As an application, they prove the existence of low-weight approximators for polynomial threshold functions.

In a beautiful result, Daniel Kane [52] showed that the Gotsman-Linial conjecture is true in the Gaussian setting. Specifically, define Gaussian noise sensitivity as follows:

Definition 5.1.1 (Gaussian Noise Sensitivity)**.** Let $f : \mathbb{R}^n \to \{-1, 1\}$ be any Boolean function on $\mathbb{R}^n$. Let X, Y be two independent random variable drawn from the multivariate Gaussian distribution $\mathcal{N}^n$ (where $\mathcal{N} = \mathcal{N}(0, 1)$ is the univariate Gaussian distribution on $\mathbb{R}$ with mean 0 and variance 1) and Z

a δ-perturbation of X defined as follows: $Z = (1-\delta)X + \sqrt{2\delta - \delta^2}Y$. The Gaussian noise sensitivity $\mathbb{GNS}_\delta(f)$ of f for noise δ is defined as follows:

$$\mathbb{GNS}_\delta(f) = \Pr\left[f(X) \neq f(Z)\right].$$

Daniel Kane showed that for any degree d PTF f (even those not necessarily defined by multi-linear polynomials as assume in the Boolean setting), $\mathbb{GNS}_\delta(f) \sim d\sqrt{2\varepsilon}/\pi$.

5.1.3 Proof Outline

The proofs of our theorems are inspired by the use of the *invariance principle* in the proof of the "Majority is Stablest" theorem [76]. As in the proof of the "Majority is Stablest" theorem, our main technical tools are the invariance principle and the anti-concentration bounds (also called small ball probabilities) of Carbery and Wright [21].

Bounding the probability that a threshold function changes value either when it is perturbed slightly (in the case of noise sensitivity) or when a variable is flipped (average sensitivity) involves bounding probabilities of the form $\Pr\left[|Q(X)| \leq |R(X)|\right]$ where $Q(X), R(X)$ are low degree polynomials and R has small l_2-norm relative to that of Q. The event $|Q(X)| \leq |R(X)|$ implies that either $|Q(X)|$ is small or $|R(X)|$ is large. In other words, for every γ

$$\Pr\left[|Q(X)| \leq |R(X)|\right] \leq \Pr\left[|Q(X)| \leq \gamma\right] + \Pr\left[|R(X)| > \gamma\right].$$

Since R has small norm, the second quantity in the above expression can be easily bounded using a tail bound (even Markov's inequality suffices). Bound-

ing the first quantity is trickier. Our first observation is that if the random variable X were distributed according to the Gaussian distribution as opposed to the uniform distribution on the hypercube, bounds on probabilities of the form $\Pr[|Q(X)| \leq \gamma]$ immediately follow from the anti-concentration bounds of Carbery and Wright [21]. We then transfer these bounds to the Boolean setting using the invariance principle.

Unfortunately, the invariance principle holds only for regular polynomials (i.e., polynomials in which no single variable has large influence). We thus obtain the required bounds on noise sensitivity and average sensitivity for the special case of regular PTFs. We then extend these results to an arbitrary PTF f using our structural results on random restrictions of the PTF f. The structural results state that either the restricted PTF is a regular polynomial or is a very biased function. In the former case, we resort to the above argument for regular PTFs and bound the noise sensitivity of the given PTF. In the latter case, we merely note that the noise sensitivity of a biased function can be easily bounded. This in turn lets us extend the results for regular PTFs to all PTFs.

5.1.4 Learning Theory Applications

In this section, we briefly elaborate on the learning theory applications of our results. As noted earlier, our bounds on noise sensitivity imply learning results in the challenging *agnostic* model of learning (see Section 2.4 for the definition of the model).

Combining our noise sensitivity bound, Theorem 5.1.1 and the results of Klivans et al. [59] (Theorem 2.4.2), Kalai et al. [50] (Theorem 2.4.1) we obtain the following:

Theorem 5.1.4. *The concept class of degree d PTFs is agnostically learnable to within ε with respect to the uniform distribution on $\{-1,1\}^n$ in time $n^{1/\varepsilon^{O(d)}}$.*

These are the first polynomial-time algorithms for agnostically learning constant degree PTFs with respect to the uniform distribution on the hypercube (to within any constant error parameter).

Finally, using the results of Blais et al. [19], our learning result can be extend a very broad class of discrete and continuous product distributions. We do not delve into these details here.

5.2 Notation and Preliminaries

1. For a subset $I \subseteq [n]$, the monomial X^I is defined by $X^I = \prod_{i \in I} X_i$.

2. Unless otherwise stated, we work with a PTF f of degree d and a degree d polynomial $P(X) = \sum_I a_I X^I$ with zero constant term (i.e., $a_\emptyset = 0$) such that $f(X_1, \ldots, X_n) = \mathsf{sign}(P(X_1, \ldots, X_n) - \theta)$. In case of ambiguity, we will refer to the coefficients a_I as $a_I(P)$.

3. For $i \in [n]$, $x^i = (x_1, \ldots, x_i) \in \{1,-1\}^i$, $f_{x^i} : \{1,-1\}^{n-i} \to \{1,-1\}$ is defined by $f_{x^i}(X_{i+1}, \ldots, X_n) = \mathsf{sign}(P(x_1, \ldots, x_i, X_{i+1}, \ldots, X_n) - \theta)$.

4. For $i \in [n]$, $P_{|i}(X_1, \ldots, X_i) = \sum_{I \subseteq [i]} a_I X^I$ is the *restriction* of P to the variables $X_1, \ldots, X_i$.

5. For clarity, we suppress the exact dependence of the constants on the degree d in this chapter; a more careful examination of our proofs shows that all constants depending on the degree d are at worst $2^{O(d)}$.

Definition 5.2.1. A partial assignment $x^i = (x_1, \ldots, x_i)$ is ε-*determining* for f, if there exists $b \in \{1, -1\}$ such that $\mathsf{Pr}_{(X_{i+1}, \ldots, X_n) \in_u \{1,-1\}^{n-i}}[\, f_{x^i}(X_{i+1}, \ldots, X_n) \neq b\,] \leq \varepsilon$.

Recall the notion of regular polynomials Section 2.2. We shall use the following notation for this chapter: For a polynomial Q, the *weight* of the i^{th} coordinate is defined by $w_i^2(Q) = \sum_{I \ni i} a_I^2$. For $i \in [n]$, let $\sigma_i(Q)^2 = \sum_{j \geq i} w_j^2(Q)$. Observe that, from Definition 2.2.2, Q is ε-regular if $\sum_i w_i^4(P) \leq \varepsilon^2 \left(\sum_i w_i^2(P)\right)^2 = \varepsilon^2 \sigma_1^4(P)$. We also assume without loss of generality that the variables are ordered such that $w_1(P) \geq w_2(P) \geq \cdots \geq w_n(P)$.

In addition, $(2, 4)$-hypercontractivity, the invariance principle of Mossel et al. [76] and the anti-concentration bounds of Carbery and Wright [21] described in Section 2.2 will play a prominent role in this chapter.

5.3 Random Restrictions of PTFs

We now establish our structural results on random restrictions of low-degree PTFs. The use of *critical indices* ($K(P, \varepsilon)$) in our analysis is motivated

by the results of Servedio [92] and Diakonikolas et al. [26] who obtain similar results for LTFs. At a high level, we show the following.

Given any $\varepsilon > 0$, define the ε-critical index of a multilinear polynomial P, $K = K(P, \varepsilon)$, to be the least index i such that $w_j^2(P) \leq \varepsilon^2 \, \sigma_{i+1}^2(P)$ for all $j > i$. We consider two cases depending on how large $K(P, \varepsilon)$ is and roughly, show the following (here $c, \alpha > 0$ are some universal constants).

1. $K \leq 1/\varepsilon^{cd}$. In this case we show that for $x^K = (x_1, \ldots, x_K) \in_u \{1, -1\}^K$, the PTF f_{x^K} is ε-regular with probability at least α.

2. $K > 1/\varepsilon^{cd}$. In this case we show that with probability at least α, the value of the threshold function is determined by the *top* $L = 1/\varepsilon^{cd}$ variables.

More concretely, we show the following.

Lemma 5.3.1. *For every integer d, there exist constants $a_d \in \mathbb{R}$, $\gamma_d > 0$ such that for any multilinear polynomial P of degree at most d and $K = K(P, \varepsilon)$ as defined above, the following holds. The polynomial $P_{x^K}(Y_{k+1}, \ldots, Y_n) \stackrel{def}{=} P(x_1, \ldots, x_K, Y_{K+1}, \ldots, Y_n)$ in variables $Y_{K+1}, \ldots, Y_n$ obtained by randomly choosing $x^K = (x_1, \ldots, x_K) \in_u \{1, -1\}^K$ is $a_d\varepsilon$-regular with probability at least γ_d.*

Lemma 5.3.2. *For every d, there exist constants $b_d, c_d \in \mathbb{R}$, $\delta_d > 0$, such that for any multilinear polynomial P of degree at most d the following holds. If*

$K(P,\varepsilon) \geq c_d \log(1/\varepsilon)/\varepsilon^2 = L$, *then a random partial assignment* $(x_1,\ldots,x_L) \in_u \{1,-1\}^L$ *is* $b_d\varepsilon$*-determining for* P *with probability at least* δ_d.

By repeatedly applying the above lemmas, we show that arbitrary low-degree PTFs can be approximated by small depth decision trees in which the leaf nodes either compute a regular PTF or a function with high bias. We first introduce some notation to this end. Though, we do not need it in this section, we state our main structural lemma for a more general class of distributions with limited independence as this will be important later on in our construction of PRGs for PTFs in Chapter 7.

Lemma 5.3.3. *There exist universal constants* $c, c_d, \delta_d > 0$ *such that for* $K(P,\varepsilon) \geq c\log(1/\varepsilon)/\varepsilon^2 = L$, *the following holds for all* $\theta \in \mathbb{R}$. *For a random partial assignment* $(x_1,\ldots,x_L) \in_u \{1,-1\}^L$ *with probability at least* δ_d *the following happens. There exists* $b \in \{1,-1\}$ *such that*

$$\Pr_{(Y_{L+1},\ldots,Y_n)\leftarrow D}[\,\mathsf{sign}(P(x_1,x_2,\ldots,x_L,Y_{L+1},\ldots,Y_n)-\theta) \neq b\,] \leq c_d\varepsilon, \quad (5.3.1)$$

for any $2d$*-wise independent distribution* D *over* $\{1,-1\}^{n-L}$.

When D is the uniform distribution over $\{1,-1\}^{n-L}$, the above lemma is equivalent to Lemma 5.3.2. However, the argument for Lemma 5.3.2 extends straightforwardly to $2d$-wise independent distributions D and we skip the details.

Definition 5.3.1. A block decision tree T with block-size L is a decision tree with the following properties. Each internal node of the decision tree reads at

most L variables. For each leaf node $\rho \in T$, the output upon reaching the leaf node ρ is a function $f_\rho : \{1,-1\}^{V_\rho} \to \{1,-1\}$, where V_ρ is the set of variables not occurring on the path to the node ρ. The depth of T is the length of the longest path from the root of T to a leaf in T.

Definition 5.3.2. Given a block decision tree T computing a function f, we say that a leaf node $\rho \in T$ is (ε, d)-*good* if the function f_ρ satisfies one of the following two properties.

1. There exists $b \in \{1,-1\}$, such that for any $2d$-wise independent distribution D over $\{1,-1\}^{V_\rho}$,

$$\Pr_{Y \leftarrow D}[f_\rho(Y) \neq b] \leq \varepsilon.$$

2. f_ρ is a ε-regular degree d PTF.

We are ready to state and prove our main structural result on writing low-degree PTFs as a "decision tree of regular PTFs" assuming Lemmas 5.3.1, 5.3.3. As mentioned before, Diakonikolas et al. [30] obtain a similar result to ours (although, their definition regularity is technically a bit different from ours).

Theorem 5.3.4 (Main Restriction Theorem)**.** *There exist universal constants c'_d, c''_d such that the following holds for any degree d polynomial P and PTF $f = \mathsf{sign}(P(\,) - \theta)$. There exists a block decision tree T computing f of block-size $L = c'_d \log(1/\varepsilon)/\varepsilon^2$ and depth at most $c''_d \log(1/\varepsilon)$, such that with probability*

at least $1-\varepsilon$ *a uniformly random walk on the tree leads to an* (ε, d)*-good leaf node.*

Proof. The proof is by recursively applying Lemmas 5.3.1 and 5.3.3. Let $c, c_d, \gamma_d, \delta_d$ be constants from the above lemmas. Let L be defined as in Lemma 5.3.2 and let $\alpha = \min(\gamma_d, \delta_d)$. For $S \subseteq [n]$ and a partial assignment $y \in \{1,-1\}^S$, let $P_y : \{1,-1\}^{[n]/S} \to \mathbb{R}$ be the degree at most d polynomial defined by $P_y(Y) = P(Z)$, where $Z_i = y_i$ for $i \in S$ and $Z_i = Y_i$ for $i \notin S$. Let $L(y) = \min(K(P_y, \varepsilon), L)$ and let $I(y)$ be the $L(y)$ largest influence coordinates in the polynomial P_y. We now define a block-decision tree computing f inductively.

Let $y_0 = \emptyset$ and let $I_0 = I(y_0)$. The root of the decision tree reads the variables in I_0. For $0 \leq q \leq \log_{1/(1-\alpha)}(1/\varepsilon)$ suppose that after q steps we are at a node β having read the variables in $S(\beta) \subseteq [n]$ and a corresponding partial assignment y. Then, if P_y is $c_d\varepsilon$-regular or if P_y satisfies Equation (5.3.1) we stop. Else, we make another step and read the values of variables in $I(y)$.

For any leaf node ρ, let $y(\rho)$ denote the partial assignment that leads to ρ. Then the leaf node ρ outputs the function $f_\rho(Y) = \mathsf{sign}(P_{y(\rho)}(Y) - \theta)$.

It follows from the construction that T is a block-decision tree computing f with block-size L and depth at most $\log_{1/(1-\alpha)}(1/\varepsilon)$. Further, for any internal node $\beta \in T$, by Lemmas 5.3.1, 5.3.2 at least α fraction of its children are $(c_d\varepsilon, d)$-good. Since any leaf node that is not $(c_d\varepsilon, d)$-good is at least $\log_{1/(1-\alpha)}(1/\varepsilon)$ far away from the root of T, it follows that a uniformly

random walk on T leads to a $(c_d\varepsilon, d)$-good node with probability at least $1-\varepsilon$. The lemma now follows. □

To prove the main lemmas we use the following simple results.We use the following lemma of Alon et al. [5].

Lemma 5.3.5 ([5, Lemma 3.2])**.** *Let A be a real valued random variable satisfying $\mathbb{E}[A] = 0$, $\mathbb{E}[A^2] = \sigma^2$ and $\mathbb{E}[A^4] \leq b\sigma^4$. Then, $\mathsf{Pr}[\, A \geq \sigma/4\sqrt{b}\,] \geq 1/4^{4/3}b$.*

Lemma 5.3.6. *For $d > 0$ there exist constants $\alpha_d, \beta_d > 0$ such that for any degree at most d polynomial Q, and $X \in_u \{1,-1\}^n$, $\mathsf{Pr}[\, Q(X) \geq \mathbb{E}[Q] + \alpha_d\sigma(Q)\,] \geq \beta_d$, where $\sigma^2(Q)$ is the variance of $Q(X) = \|Q\|^2 - (\mathbb{E}_X[Q])^2$. In particular, $\mathsf{Pr}[\, Q(X) \geq \mathbb{E}[Q]\,] \geq \beta_d$.*

Proof. Let random variable $A = Q(X) - \mathbb{E}_X[Q(X)]$. Then, $\mathbb{E}[A] = 0$, $\mathbb{E}[A^2] = \sigma^2(Q)$ and by $(2,4)$-hypercontractivity, $\mathbb{E}[A^4] \leq 9^d\, \mathbb{E}[A^2] = 9^d\sigma^4(Q)$. The claim now follows from Lemma 5.3.5. □

5.3.1 Proof of Lemma 5.3.1

Proof. Let $X \equiv (X_1, \ldots, X_K)$. We prove the lemma as follows: (1) Bound the expectation of $\sum_{j>K} w_j^4(P_X)$ using hypercontractivity and use Markov's inequality to show that with high probability $\sum_{j>K} w_j^4(P_X)$ is small. (2) Use the fact that $\sigma^2_{K+1}(P_X) = \sum_{j>K} w_j^2(P_X)$ is a degree at most $2d$ polynomial in X and Lemma 5.3.6 to lower bound the probability that $\sigma^2_{K+1}(P_X)$ is large.

Let

$$P_X(Y_{K+1},\ldots,Y_n) = P(X_1,\ldots,X_K,Y_{K+1},\ldots,Y_n) = \\ R(X_1,\ldots,X_K) + \sum_{J\subseteq[K+1,n],0<|J|\leq d} Q_J(X_1,\ldots,X_K)\prod_{j\in J} Y_j.$$

We now bound $\mathbb{E}[\sum_{j>K} w_j^4(P_X)]$. Fix a $j > K$ and observe that $w_j^2(P_X) = \sum_{J\ni j} Q_J^2(X)$. Thus,

$$\mathop{\mathbb{E}}_X\left[w_j^2(P_X)\right] = \sum_{J\ni j}\mathop{\mathbb{E}}_X\left[Q_J^2(X)\right] = \sum_{J\ni j}\|Q_J\|^2 = w_j^2(P). \tag{5.3.2}$$

Further, by $(2,4)$-hypercontractivity, Corollary 2.2.2,

$$\mathop{\mathbb{E}}_X\left[w_j^4(P_X)\right] = \mathop{\mathbb{E}}_X\left[\sum_{J_1,J_2\ni j} Q_{J_1}^2(X)\,Q_{J_2}^2(X)\right] = \sum_{J_1,J_2\ni j}\mathop{\mathbb{E}}_X\left[\,Q_{J_1}^2(X)\,Q_{J_2}^2(X)\,\right] \\ \leq \sum_{J_1,J_2\ni j} 9^d\mathop{\mathbb{E}}_X\left[Q_{J_1}^2(X)\right]\cdot\mathop{\mathbb{E}}_X\left[Q_{J_2}^2(X)\right] = \sum_{J_1,J_2\ni j} 9^d\,\|Q_{J_1}\|^2\,\|Q_{J_2}\|^2 = 9^d\,w_j^4(P).$$

Hence, $\mathbb{E}[\,\sum_{j>K} w_j^4(P_X)\,] \leq 9^d\sum_{j>K} w_j^4(P)$. Now, from the definition of $K(P,\varepsilon)$, $w_j^2(P) \leq \varepsilon^2\sigma_{K+1}^2(P)$ for all $j > K$. Thus,

$$\sum_{j>K} w_j^4(P) \leq \varepsilon^2\sigma_{K+1}^2(P)\sum_{j>K} w_j^2(P) = \varepsilon^2\sigma_{K+1}^4(P).$$

Combining the above inequalities and applying Markov's inequality we get

$$\Pr_X\left[\,\sum_{j>K} w_j^4(P_X) \geq \gamma 9^d\varepsilon^2\sigma_{K+1}^4(P)\,\right] \leq 1/\gamma. \tag{5.3.3}$$

Observe that $Q(X) = \sum_{j>K} w_j^2(P_X)$ is a degree at most $2d$ polynomial in $X_1,\ldots,X_k$ and by (5.3.2), $\mathbb{E}[Q] = \sum_{j>K} w_j^2(P) = \sigma_{K+1}^2(P)$. Thus, by applying Lemma 5.3.6 to Q, $\Pr[\,\sum_{j>K} w_j^2(P_X) \geq \sigma_{K+1}^2(P)\,] \geq \beta_{2d}$. Setting

$\gamma = 2/\beta_{2d}$ in (5.3.3) and using the above equation, we get

$$\Pr_X \left[\sum_{j>K} w_j^4(P_X) \leq a_d^2 \varepsilon^2 \left(\sum_{j>K} w_j^2(P_X) \right)^2 \right] \geq \beta_{2d}/2,$$

where $a_d^2 = 2 \cdot 9^d/\beta_{2d}$. Thus, the polynomial $P_X(Y_{K+1}, \ldots, Y_n)$ is $(a_d\varepsilon)$-regular with probability at least $\gamma_d = \beta_{2d}/2$. □

5.3.2 Proof of Lemma 5.3.2

Proof of Lemma 5.3.2. Suppose that $K(P, \varepsilon) \geq L = c \log(1/\varepsilon)/\varepsilon^2$ for a constant c to be chosen later and let $Q(X_1, \ldots, X_n) = P(X_1, \ldots, X_n) - P_{|L}(X_1, \ldots, X_L)$. The proof proceeds as follows. We first show that $\|Q\|$ is significantly smaller than $\|P_{|L}\|$. We then use Lemma 5.3.6 applied to $P_{|L} - \theta$ and Markov's inequality applied to $|Q(X)|$ to show that $|P_{|L}(X_1, \ldots, X_L) - \theta|$ is larger than $|Q(X)|$, so that $Q(X)$ cannot flip the sign of $P_{|L}(X_1, \ldots, X_L) - \theta$, with at least a constant probability. We first bound $\|Q\|$.

Lemma 5.3.7. *For $1 \leq i < j < K(P, \varepsilon)$, $\sigma_j^2(P) \leq (1 - \varepsilon^2)^{j-i} \sigma_i^2(P)$.*

Proof. For $1 \leq i < K(P, \varepsilon)$, we have

$$\sigma_i^2(P) = w_i^2(P) + \sigma_{i+1}^2(P) \geq \varepsilon^2 \sigma_i^2(P) + \sigma_{i+1}^2(P).$$

Thus, $\sigma_{i+1}^2(P) \leq (1 - \varepsilon^2)\sigma_i^2(P)$. The lemma follows. □

Claim 5.3.8. *For a suitably large enough constant c_d, $\|Q\| \leq \sqrt{\varepsilon}\, \alpha_d \|P_{|L}\|$.*

Proof. Let α_d, β_d be the constants from Lemma 5.3.6. By definition $\|Q\|^2 = \sum_{I:I\not\subseteq[L]} a_I^2 \le \sigma_L^2(P)$. Now,

$$\begin{aligned}
\sigma_1^2(P) &= \sum_{j<L} w_j^2(P) + \sigma_L^2(P) \\
&\le d \sum_{I:I\cap[L]\neq\emptyset} a_I^2 + \sigma_L^2(P) \\
&\le d \sum_{I:\emptyset\neq I\subseteq[L]} a_I^2 + d \sum_{I:I\not\subseteq[L]} a_I^2 + \sigma_L^2(P) \\
&\le d \sum_{I:\emptyset\neq I\subseteq[L]} a_I^2 + d\sum_{j>L} w_j^2(P) + \sigma_L^2(P) \\
&\le d \sum_{I:\emptyset\neq I\subseteq[L]} a_I^2 + (d+1)\,\sigma_L^2(P).
\end{aligned}$$

Further, by Lemma 5.3.7,

$$\sigma_L^2(P) \le (1-\varepsilon^2)^{L-1}\sigma_1^2(P).$$

Combining the above inequalities we get,

$$\sigma_L^2(P) \le O_d\left((1-\varepsilon^2)^{L-1}\right) \sum_{I:\emptyset\neq I\subseteq[L]} a_I^2 = O_d\left((1-\varepsilon^2)^{L-1}\right)\,\sigma^2(P). \tag{5.3.4}$$

Choosing $L = c_d \log(1/\varepsilon)/\varepsilon^2$ for large enough c_d, we get the claim. □

By Claim 5.3.8 and Markov's inequality,

$$\begin{aligned}
\Pr_{x\in_u\{1,-1\}^n}\left[\,|Q(x_1,\ldots,x_n)| \ge \alpha_d\,\|P_{|L}\|\,\right] \le \\
\Pr_{x\in_u\{1,-1\}^n}\left[\,|Q(x_1,\ldots,x_n)| \ge \frac{\|}{Q}\|/\sqrt{\varepsilon}\,\right] \le \varepsilon.
\end{aligned} \tag{5.3.5}$$

Let $S \subseteq \{1,-1\}^L$ be the set of all *bad* $x^L \in \{1,-1\}^L$ such that,

$$\Pr_{(X_{L+1},\ldots,X_n)\in_u\{1,-1\}^n}\left[\,|Q(x_1,\ldots,x_L,X_{L+1},\ldots,X_n)| \ge \alpha_d\,\|P_{|L}\|\,\right] \ge 2\varepsilon/\beta_d.$$

Then, from (5.3.5) and the above equation, $\mathsf{Pr}_{x^L \in_u \{1,-1\}^L}\left[x^L \in S\right] \leq \beta_d/2$. Now, let $T \subseteq \{1,-1\}^L$ be such that for $x^L \in T$, $|\,P_{|L}(x_1,\ldots,x_L) - \theta\,| \geq \alpha_d \|P_L\|$ and $x^L \notin S$. Observe that all $x^L \in T$ are $(2\varepsilon/\beta_d)$-determining and by Lemma 5.3.6 and the above equations,

$$\begin{aligned}&\mathop{\mathsf{Pr}}_{x^L \in_u \{1,-1\}^L}\left[x^L \in T\right] \geq \\ &\mathop{\mathsf{Pr}}_{x^L \in_u \{1,-1\}^L}[\left|P_{|L}(x_1,\ldots,x_L) - \theta\right| \geq \alpha_d \|P_L\|] - \mathop{\mathsf{Pr}}_{x^L \in_u \{1,-1\}^L}\left[x^L \in S\right] \geq \beta_d/2.\end{aligned}$$

The lemma now follows. □

5.4 Noise sensitivity of PTFs

We now bound the noise sensitivity of PTFs and prove Theorem 5.1.1. We do so by first bounding the noise sensitivity of regular PTFs and then use the results of the previous section to reduce the general case to the regular case.

5.4.1 Noise sensitivity of Regular PTFs

At a high level, we bound the noise sensitivity of regular PTFs as follows: (1) Reduce the problem to that of proving certain anti-concentration bounds for regular PTFs over the hypercube. (2) Use the invariance principle of Mossel et al. [76] to reduce proving anti-concentration bounds over the hypercube to that of proving anti-concentration bounds over Gaussian distributions. (3) Use the Carbery-Wright anti-concentration bounds [21] for polynomials over log-concave distributions.

For the rest of this section, we fix degree d multilinear polynomial P and a corresponding degree d PTF f. Recall that it suffices to consider multilinear polynomials as we are working over the hypercube. We first reduce bounding noise sensitivity to proving anti-concentration bounds.

Lemma 5.4.1. *For $0 < \rho < 1$, $\delta > 0$, $\mathbb{NS}_\rho(f) \leq (d+1)\,\delta + \mathsf{Pr}_{x\in\{1,-1\}^n}[\,|P(x)-\theta| \leq 2\sqrt{\rho}/\delta\,]$.*

Proof. Let S be a random subset $S \subseteq [n]$ where each $i \in [n]$ is in S independently with probability ρ. From the definition of noise sensitivity it easily follows that

$$\begin{aligned}
\mathbb{NS}_\rho(f) &= \mathop{\mathsf{Pr}}_{X\in_u\{1,-1\}^n,S}[\,\mathsf{sign}\,(P(X)-\theta) \neq \mathsf{sign}(\,P(X) - 2\sum_{I:|I\cap S| \text{ is odd}} a_I X^I - \theta\,)\,] \\
&= \mathop{\mathsf{Pr}}_{X\in_u\{1,-1\}^n,S}[\,|P(x)-\theta| \leq 2\,|\sum_{I:|I\cap S| \text{ is odd}} a_I X^I\,|\,] \\
&\leq \mathop{\mathsf{Pr}}_{X\in_u\{1,-1\}^n,S}[\,|\sum_{I:|I\cap S| \text{ is odd}} a_I X^I\,| \geq \sqrt{\rho}/\delta\,] + \\
&\qquad \mathop{\mathsf{Pr}}_{X\in_u\{1,-1\}^n}[\,|\,P(X)-\theta\,| \leq 2\sqrt{\rho}/\delta\,] \qquad (5.4.1)
\end{aligned}$$

Define a non-negative random variable P_S as follows: $P_S^2 = \sum_{I:|I\cap S| \text{ is odd}} a_I^2$. We can then bound the first quantity in the above expression using P_S as follows:

$$\begin{aligned}
&\mathop{\mathsf{Pr}}_{X\in_u\{1,-1\}^n,S}[\,|\sum_{I:|I\cap S| \text{ is odd}} a_I X^I\,| \geq \sqrt{\rho}/\delta\,] \leq \\
&\qquad \mathop{\mathsf{Pr}}_{X\in_u\{1,-1\}^n,S}[\,|\sum_{I:|I\cap S| \text{ is odd}} a_I X^I\,| \geq P_S/\sqrt{\delta}\,] + \mathop{\mathsf{Pr}}_{S}[\,P_S \geq \sqrt{\rho}/\sqrt{\delta}\,] \qquad (5.4.2)
\end{aligned}$$

Since $\mathbb{E}_X(\sum_{I|I\cap S| \text{ is odd}} a_I X^I)^2 = P_S^2$, by Markov's inequality, we have

$$\Pr_{x\in_u\{1,-1\}^n}[\,|\sum_{I:|I\cap S| \text{ is odd}} a_I X^I| \geq P_S/\sqrt{\delta}\,] \leq \delta. \tag{5.4.3}$$

Now, note that $P_S^2 \leq \sum_{i\in S} w_i^2(P)$. Thus, $\mathbb{E}_S[P_S^2] \leq \mathbb{E}_S[\sum_{i\in S} w_i^2(P)] = \rho\sum_i w_i^2(P) \leq d\,\rho$. Hence, by Markov's inequality, $\Pr_S[P_S \geq \sqrt{\rho}/\sqrt{\delta}\,] \leq d\,\delta$. The lemma now follows by combining Equations (5.4.1), (5.4.2), (5.4.3) and the above equation. □

We now prove an anti-concentration bound for regular PTFs.

Lemma 5.4.2. *If P is ε-regular, then for any interval $I \subseteq \mathbb{R}$ of length at most α, $\Pr_{X\in_u\{1,-1\}^n}[\,P(X) \in I\,] = O_d(\,\alpha^{1/d} + \varepsilon^{2/(4d+1)}\,)$.*

Proof. Let $Z_1 = P(X), Z_2 = P(Y)$ for $X \in_u \{1,-1\}^n, Y \leftarrow \mathcal{N}^n$. Then, since P is ε-regular, by Theorem 2.2.5, for all $t \in \mathbb{R}$, $|\,\Pr[Z_1 > t] - \Pr[Z_2 > t]\,| = O_d(\varepsilon^{2/(4d+1)})$. Now, by the above equation and Theorem 2.2.6 applied to the random variable Y for interval I, $\Pr[Z_1 \in I] = \Pr[Z_2 \in I] + O_d(\,\varepsilon^{2/(4d+1)}\,) = O_d(\,\alpha^{1/d} + \varepsilon^{2/(4d+1)}\,)$. □

We can now obtain a bound on noise sensitivity of regular PTFs.

Theorem 5.4.3. *If f is an ε-regular PTF of degree d, then $\mathbb{NS}_\varepsilon(f) \leq O_d\left(\varepsilon^{1/(2d+2)}\right)$.*

Proof. Let $\delta > 0$ to be chosen later. Then, by Lemma 5.4.1 and Lemma 5.4.2 above, $\mathbb{NS}_\varepsilon(f) = O_d(\,\delta + \varepsilon^{2/(4d+1)} + \varepsilon^{1/2d}/\delta^{1/d}\,)$. Choosing $\delta = \varepsilon^{1/(2d+2)}$ we get $\mathbb{NS}_\varepsilon(f) = O_d(\,\varepsilon^{1/(2d+2)}\,)$. □

5.4.2 Noise Sensitivity of arbitrary PTFs

We prove Theorem 5.1.1 by recursively applying the following lemma.

Lemma 5.4.4. *For every d there exist universal constants $c_d, \Delta_d \in \mathbb{N}, \alpha_d \in (0,1)$ such that for $M = \min(K(P,\varepsilon), c_d \log(1/\varepsilon)/\varepsilon^2)$ and $X^M = (X_1, \ldots, X_M) \in_u \{1,-1\}^M$,*

$$\Pr_{X^M}\left[\, \mathbb{NS}_\varepsilon(f_{X^M}) \leq \Delta_d \varepsilon^{1/(2d+2)} \,\right] \geq \alpha_d. \tag{5.4.4}$$

Proof. Let $a_d, b_d, c_d, \gamma_d, \delta_d$ be the constants from Lemmas 5.3.1, 5.3.2. Let $\alpha_d = \min(\gamma_d, \delta_d)$. We consider two cases.

Case (i): $M = K(P,\varepsilon)$. Then, by Lemma 5.3.1 and Theorem 5.4.3, for $X^K \in_u \{1,-1\}^K$, with probability at least α_d, $\mathbb{NS}_\varepsilon(f_{x^K}) \leq \Delta_d \varepsilon^{1/(2d+2)}$ for some constant Δ_d.

Case (ii): $M = c_d \log(1/\varepsilon)/\varepsilon^2$. Then, by Lemma 5.3.2, $X^M \in_u \{1,-1\}^M$ is $b_d\varepsilon$-determining with probability at least α_d. Further, if X^M is $b_d\varepsilon$-determining, with f_{X^M} biased towards $b \in \{1,-1\}$, then

$$\mathbb{NS}_\varepsilon(f_{X^M}) = \Pr_{Z_1 \in_u \{1,-1\}^{n-M}, Z_2 \in_\varepsilon Z_1}\left[\, f_{X^M}(Z_1) \neq f_{X^M}(Z_2) \,\right] \leq$$
$$2 \Pr_{Z \in_u \{1,-1\}^{n-M}}\left[\, f_{X^M}(Z) \neq b \,\right] \leq 2b_d\varepsilon,$$

where $Z_2 \in_\varepsilon Z_1$ is an ε-perturbation of Z_1. The lemma now follows. □

Proof of Theorem 5.1.1. Let c_d, Δ_d, α_d be as in the above lemma and let $L = c_d \log(1/\varepsilon)/\varepsilon^2$, $t = \log_{1-\alpha_d}(1/\varepsilon)$. We will show that for $\delta = \varepsilon^{1/(2d+2)}/(L\,t) =$

$O_d(\varepsilon^{(4d+5)/(2d+2)}/\log^2(1/\varepsilon))$,

$$\mathbb{NS}_\delta(f) = O_d(\,\varepsilon^{1/(2d+2)}\,).$$

For $S \subseteq [n]$ and $x \in \{1,-1\}^n$ let $P_{x,S} : \{1,-1\}^{\bar{S}} \to \mathbb{R}$ be the degree at most d polynomial defined by $P_{x,S}(X_{\bar{S}}) = P(x_{|S}, X_{\bar{S}})$.. Fix a $x = (x_1, \ldots, x_n) \in \{1,-1\}^n$ and define $S_{x,i} \subseteq [n]$ for $i \geq 1$, recursively as follows. $S_{x,1}$ is the set of $M_1 \leq L$ largest weight coordinates in P given by applying Lemma 5.4.4 to P. For $i \geq 1$, let $S^{x,i} = S_{x,1} \cup S_{x,2} \cup \ldots \cup S_{x,i}$.

For $i > 1$, let $S_{x,i+1}$ be the set of $M_{i+1} \leq L$ largest weight coordinates in $P_{x,S^{x,i}}$ given by applying Lemma 5.4.4 to the polynomial $P_{x,S^{x,i}}$. Define $f_{x,i}$ by $f_{x,i}(\cdot) \equiv \mathsf{sgn}(P_{x,S^{x,i}}(\cdot) - \theta)$. Note that the definition of $f_{x,i}$ only depends on x_j for $j \in S^{x,i}$ and that $|S^{x,i}| \leq L \cdot i$.

Call $x \in \{1,-1\}^n$ (ε, f)-good if there exists an i, $1 \leq i \leq t$ such that $\mathbb{NS}_\varepsilon(f_{x,i}) \leq \Delta_d\, \varepsilon^{1/(2d+2)}$ and let t_x be such an i for a (ε, f)-good x. Then, from the definition of $f_{x,i}$ and Lemma 5.4.4,

$$\mathop{\mathsf{Pr}}_{x \in_u \{1,-1\}^n}[\,x \text{ is } (\varepsilon, f)\text{-good}\,] \geq 1 - \varepsilon. \tag{5.4.5}$$

Let $y \in_\delta x$ be a δ-perturbation of $x \in_u \{1,-1\}^n$. Then, since $|S^{x,t_x}| \leq L\,t$,

$$\mathop{\mathsf{Pr}}_{x,y}[\,x_{|S^{x,t_x}} \neq y_{|S^{x,t_x}}\,] \leq L\,t\,\delta = \varepsilon^{1/(2d+2)}. \tag{5.4.6}$$

Also note that for any $i \geq 1$, conditioned on an assignment for the values in $x_{|S^{x,i}}$ and $x_{|S^{x,i}} = y_{|S^{x,i}}$, $\mathsf{Pr}_{x,y}[f(x) \neq f(y)] = \mathbb{NS}_\delta(f_{x,i}) \leq \mathbb{NS}_\varepsilon(f_{x,i})$.

Thus, conditioned on x being (ε, f)-good and $x_{|S^{x,t_x}} = y_{|S^{x,t_x}}$,

$$\Pr_{x,y}[\, f(x) \neq f(y)\,] \leq \mathbb{NS}_{\varepsilon}(f_{x,t_x}) \leq \Delta_d\, \varepsilon^{1/(2d+2)}. \tag{5.4.7}$$

Combining (5.4.5), (5.4.6), (5.4.7), we get

$$\mathbb{NS}_{\delta}(f) \leq \varepsilon + L\, t\, \delta + \Delta_d \varepsilon^{1/(2d+2)} = O_d\left(\, \varepsilon^{1/(2d+2)}\,\right).$$

Since $\delta = O_d\left(\varepsilon^{\frac{4d+5}{2d+2}} / \log^2(1/\varepsilon)\right)$ and the above is applicable for all $\varepsilon > 0$, we get that for all $\rho > 0$,

$$\mathbb{NS}_{\rho}(f) = O_d\left(\, \log(1/\rho)\rho^{1/(4d+5)}\,\right) = O_d\left(\, \rho^{1/(4d+6)}\,\right).$$

□

5.5 Average sensitivity of PTFs

In this section we bound the average sensitivity of PTFs on the Boolean hypercube, proving Theorem 5.1.2. We first prove a lemma bounding the average sensitivity of a Boolean function in terms of its noise sensitivity. Theorem 5.1.2 follows immediately from Theorem 5.1.1 and the following lemma:

Lemma 5.5.1 (noise sensitivity to average sensitivity)**.** *For any Boolean function $f : \{1,-1\}^n \to \{1,-1\}$, $\mathbb{AS}(f) \leq 2ne\, \mathbb{NS}_{(1/n)}(f)$.*

Proof. Let $\delta = 1/n$. Let $X \in_u \{1,-1\}^n$ and let $S \subseteq [n]$ be a random set with each element $i \in [n]$ present in S independently with probability δ. Let $X(S)$ be the vector obtained by flipping the coordinates of X in S. Then,

$\mathbb{NS}(f) = \Pr_{X,S}[f(X) \neq f(X(S))]$. Observe that for $i \in [n]$, $\Pr[S = i] = \delta(1-\delta)^{n-1} = (1/n)\,(1-1/n)^{n-1} > 1/2ne$. Therefore,

$$\begin{aligned}
\mathbb{NS}_\delta(f) &= \Pr_{X,S}[f(X) \neq f(X(S))] \\
&= \sum_i \Pr_S[\, S = \{i\}\,] \cdot \Pr_X[\, f(X) \neq f(X(S)) \,|\, S = i\,] + \\
&\qquad \Pr_S[\, |S| \neq 1\,] \cdot \Pr_{X,S}[\, f(X) \neq f(X(S)) \,|\, |S| \neq 1\,] \\
&> \sum_i \frac{1}{2ne} \Pr_X[f(X) \neq f(X(\{i\})] = \frac{1}{2ne}\mathbb{AS}(f).
\end{aligned}$$

□

5.6 Average sensitivity using a combinatorial argument

In this section, we give a combinatorial argument for the following bound on average sensitivity.

Theorem 5.6.1. *For any degree d PTF $f : \{1,-1\}^n \to \{1,-1\}$, $\mathbb{AS}(f) \leq 3\,n^{1-2^{-d}}$.*

We first show the theorem using Lemma 5.1.3.

Proof. Let $P(x) = x_i P_i(x_{-i}) + Q_i(x_{-i})$, where $P_i(\,), Q_i(\,)$ are degree $d-1$ and degree d polynomials respectively that do not depend on x_i. Define $f_i(x_{-i}) = \mathsf{sgn}(P_i(x_{-i}))$ and $g_i(x) = f(x)f_i(x_{-i})$. Then,

$$\begin{aligned}
\mathbb{I}_i(f) &= \Pr_{X \in_u \{1,-1\}^n}[f(X) \neq f(X^{(i)})] = \Pr_{X \in_u \{1,-1\}^n}[f(X)f_i(X_{-i}) \neq f(X^{(i)})f_i(X_{-i})] \\
&= \Pr_{X \in_u \{1,-1\}^n}[f(X)f_i(X_{-i}) \neq f(X^{(i)})f_i((X^{(i)})_{-i})] = \Pr_{X \in_u \{1,-1\}^n}[g_i(X) \neq g_i(X^{(i)})] \\
&= \mathbb{I}_i(g_i).
\end{aligned}$$

Observe that g_i is monotone increasing in x_i for $i \in [n]$ and hence $\mathbb{I}_i(g_i) = \mathbb{E}_X[X_i g_i(X)]$. Thus,

$$\mathbb{AS}(f) = \sum_i \mathbb{I}_i(f) = \sum_i \mathbb{I}_i(g_i) = \sum_i \mathop{\mathbb{E}}_X[X_i g_i(X)] = \sum_i \mathop{\mathbb{E}}_X[X_i f(X) f_i(X_{-i})] = \mathop{\mathbb{E}}_X\left[f(X)\sum_i X_i f_i(X_{-i})\right].$$

Since $|f(x)| \le 1$ for all x, we have

$$\mathbb{AS}(f) \le \mathop{\mathbb{E}}_X\left[\left|\sum_i X_i f_i(X_{-i})\right|\right]. \tag{5.6.1}$$

We now use induction and Lemma 5.1.3. For an LTF f, f_i as defined above are constants. Therefore, by Equation (5.6.1),

$$\mathbb{AS}(f) \le \mathop{\mathbb{E}}_X\left[\left|\sum_i X_i f_i(X_{-i})\right|\right] = \mathop{\mathbb{E}}_X\left[\left|\sum_i X_i\right|\right] = O(\sqrt{n}).$$

Suppose the theorem is true for degree d PTFs and let f be a degree $d+1$ PTF and let f_i be as defined before. Then, by Equation (5.6.1) and Lemma 5.1.3

$$\mathbb{AS}(f)^2 \le 2\sum_i \mathbb{AS}(f_i) + n \le \sum_i 6\, n^{1-2^{-d}} + n \le 7\, n^{2-2^{-d}}.$$

Therefore, $\mathbb{AS}(f) \le 3\, n^{1-2^{-(d+1)}}$. The theorem follows by induction. □

Proof of Lemma 5.1.3. For brevity, let $f_i(x) = f_i(x_{-i})$. By Cauchy-Schwarz,

for any random variable Z we have $\mathbb{E}[|Z|]^2 \leq \mathbb{E}[Z^2]$. Thus,

$$\begin{aligned}\mathop{\mathbb{E}}_{X}\left[\left|\sum_i X_i f_i(X_{-i})\right|\right]^2 &\leq \mathop{\mathbb{E}}_{X}\left[\left(\sum_i X_i f_i(X_{-i})\right)^2\right] \\ &= \mathop{\mathbb{E}}_{X}[\sum_{i,j} X_i X_j f_i(X) f_j(X)\,] \\ &= n + \sum_{i\neq j} \mathop{\mathbb{E}}_{X}[\,X_i X_j f_i(X) f_j(X)\,]. \qquad (5.6.2)\end{aligned}$$

For $i \neq j \in [n]$, let $x_{-ij} = (x_k : k \in [n], k \neq i, j)$ and let $S_i^j = \{x \in \{1,-1\}^n : f_i(x) \neq f_i(x \oplus e_j)\}$. Note that $\mathbb{I}_j(f_i) = \mathsf{Pr}_X[X \in S_i^j]$. Now,

$$\mathop{\mathbb{E}}_{X}[\,X_i X_j f_i(X) f_j(X)\,] = \sum_{x \in S_i^j \cup S_j^i} \mu(x)\, x_i x_j f_i(x) f_j(x) + \sum_{x \notin S_i^j \cup S_j^i} \mu(x)\, x_i x_j f_i(x) f_j(x), \qquad (5.6.3)$$

where $\mu(x) = 1/2^n$ is the probability of choosing x under the uniform distribution. We bound the first term in the above expression by the average sensitivity of the f_i's and show that the second term vanishes. Observe that,

$$\sum_{x \in S_i^j \cup S_j^i} \mu(x) x_i x_j f_i(x) f_j(x) \leq \mu(S_i^j \cup S_j^i) \leq \mu(S_i^j) + \mu(S_j^i) = \mathbb{I}_j(f_i) + \mathbb{I}_i(f_j). \qquad (5.6.4)$$

Note that for $x \notin S_i^j \cup S_j^i$, $f_i(x), f_j(x)$ are both independent of the values of x_i, x_j. For such x (abusing notation) let $f_i(x_{-ij}) = f_i(x)$, $f_j(x_{-ij}) = f_j(x)$ and let $T_{ij} = \{(x_k : k \neq i, j) : x \notin S_i^j \cup S_j^i\}$. Then, since for $x \notin S_i^j \cup S_j^i$, $f_i(x), f_j(x)$ depend only on x_{-ij}, we get that $x \notin S_i^j \cup S_j^i$ if and only if $x_{-ij} \notin T_{ij}$.

Therefore,

$$\sum_{x \notin S_i^j \cup S_j^i} \mu(x)\, x_i x_j f_i(x) f_j(x) = \sum_{x \notin S_i^j \cup S_j^i} \mu(x_{-ij})\, \mu(x_i)\, \mu(x_j)\, f_i(x_{-ij})\, f_j(x_{-ij})\, x_i x_j$$
$$= \sum_{x_{-ij} \notin T_{ij}} \mu(x_{-ij}) f_i(x_{-ij})\, f_j(x_{-ij}) \mathop{\mathbb{E}}_{x_i, x_j} [x_i x_j] = 0. \tag{5.6.5}$$

From Equations (5.6.2), (5.6.3), (5.6.4),(5.6.5) we have,

$$\mathop{\mathbb{E}}_{X} \left[\left| \sum_i X_i f_i(X_{-i}) \right| \right]^2 \leq n + \sum_{i \neq j} (\mathbb{I}_j(f_i) + \mathbb{I}_i(f_j)) = n + 2 \sum_i \sum_{j : j \neq i} \mathbb{I}_j(f_i) =$$
$$n + 2 \sum_i \mathbb{AS}(f_i).$$

□

Remark 5.6.1. The bound of Lemma 5.1.3 is tight up to a constant factor if we only have bounds on the average sensitivity of the f_i's to go with. For example, consider f_i defined as follows. Divide $[n]$ into $m = \sqrt{n}$ blocks $B_1, \ldots, B_m$ of size m each and for $1 \leq j \leq m$, $i \in B_j$, let $f_i = \prod_{k \in B_j : k \neq i} x_k$. Then, the left hand side of the lemma is $\Theta(n^{3/2})$ and $\mathbb{AS}(f_i) = m - 1 = \Theta(\sqrt{n})$ for all i.

Chapter 6

Noise Sensitivity of Polytopes

In this chapter we use our invariance principle for polytopes to bound the noise sensitivity of polytopes or intersections of halfspaces.

6.1 Introduction

As discussed in Section 2.4, noise sensitivity of Boolean functions introduced in the seminal works of [49], [16] is an important notion in the *analysis* of Boolean functions with a variety of applications in complexity theory.

A direct application of our invariance principle Theorem 3.1.1 gives the following new bound on the noise sensitivity of intersections of regular halfspaces:

Theorem 6.1.1 (noise sensitivity of intersections of halfspaces)**.** *Let f be computed by the intersection of k, ε-regular halfspaces. Then the Boolean noise sensitivity of f for noise rate ε is at most* $(\log k)^{O(1)} \cdot \varepsilon^{1/6}$.

Applying the results of Kalai et al. [50] and Klivans et al. [59] (see Section 2.4), the above theorem implies that intersections of k, ε-regular halfspaces are agnostically learnable with respect to the uniform distribution on

$\{-1,1\}^n$ in time $n^{(\log^{O(1)} k)}$ for any constant error parameter. In particular, intersections of $\{-1,1\}$ halfspaces (oriented majorities) are ε-regular and fall into this class. The previous best algorithm for learning this concept class, even in the easier PAC model, ran in time $n^{O(k^2)}$ ([59, 50]).

The current best bound for the noise sensitivity of intersection of k arbitrary halfspaces is $O(k\sqrt{\varepsilon})$. This bound is obtained by starting with the $\sqrt{\varepsilon}$ noise sensitivity bound for a single halfspace due to Peres [85] and applying a union bound over k halfspaces. On the other hand, optimal bounds of $\Theta(\sqrt{\log k}\sqrt{\varepsilon})$ for the related Gaussian noise sensitivity were obtained recently by Klivans et al. [60]. We believe that the right order for Boolean noise sensitivity of intersection of k halfspaces is $\Theta(\sqrt{\log k}\sqrt{\varepsilon})$ as well.

Improving the bounds for Boolean noise sensitivity would be of considerable interest, particularly for the learning theory applications, as learning the class of intersections of halfspaces even with respect to specific distributions such as the uniform distribution over $\{1,-1\}^n$ is an important open problem in learning theory. We feel that our result is an important step towards improving noise sensitivity bounds for intersections of arbitrary (not necessarily regular) halfspaces.

6.2 Noise Sensitivity of Intersections of Regular Halfspaces

We now describe how our invariance principle yields a bound on the average and noise sensitivity of intersections of regular halfspaces. We ask

the reader to recall the definition of noise sensitivity, Definition 2.4.1, from Section 2.4. For $W \in \mathbb{R}^{n\times k}$ and $p \in [k]$, let W^p denote the p'th column of W. We will assume throughout that the matrix W is normalized so that each column has ℓ_2 norm 1.

Let $f^1, \ldots, f^k : \{1,-1\}^n \to \{1,-1\}$ be halfspaces with $f^p(x) = \mathsf{sign}(\langle W^p, x\rangle - \theta_p)$ and let $f^{\wedge k} : \{1,-1\}^n \to \{1,-1\}$ be their intersection, $f^{\wedge k} = f^1 \wedge f^2 \wedge \ldots \wedge f^k$.

Theorem 6.2.1. *For $f^{\wedge k}$ ε-regular, $\mathbb{NS}_\delta(f^{\wedge k}) \leq C(\log^{1.6}(k/\delta))\,(\varepsilon^{1/6} + \delta^{1/2})$.*

We prove the theorem by first reducing bounding noise sensitivity of $f^{\wedge k}$ to bounding the Boolean volume of l_∞-neighborhoods of polytopes. We then use our invariance principle, Theorem 3.3.1, to prove the required bounds on the Boolean volume of boundaries of polytopes.

As mentioned before, the above theorem implies a $n^{\log^{O(1)} k}$ algorithm for learning intersections of regular halfspaces in the agnostic model for any constant error rate.

We use the following tail bound that follows from Pinelis's subgaussian tail estimates [86].

Fact 6.2.2. There exist absolute constants $c_1, c_2 > 0$ such that all $w \in \mathbb{R}^m$, $t > 0$,

$$\Pr_{x \in_u \{1,-1\}^m}[\,|\langle w, x\rangle| > t\|w\|\,] \leq c_1 \exp(-c_2 t^2).$$

The following claim says that for W ε-regular, random $x \in_u \{1,-1\}^n$, and a δ-perturbation y of x, $W^T x$ is close to $W^T y$ in l_∞ distance.

Claim 6.2.3. *For $x \in \{1,-1\}^n$, let $y(x)$ be a random δ-perturbation of $y(x)$ of x. Then,*

$$\Pr_{x \in_u \{1,-1\}^n, y(x)} \left[\|W^T x - W^T y(x)\|_\infty \geq \lambda \right] \leq 2\delta,$$

where $\lambda = C \log(k/\delta)^{1/2} \delta^{1/2} + C \log(k/\delta)^{3/4} \varepsilon^{1/2}$.

Proof. Let $Y = (Y_1, \ldots, Y_n)$ be i.i.d indicator variables with $\Pr[Y_i = 1] = \delta$. Let $S(Y) = support(Y)$. Now, for $p \in [k]$, $\|W^p_{S(Y)}\|^2 = \sum_{i=1}^n W^2_{ip} Y_i$ and $\mathbb{E}[\|W^p_{S(Y)}\|^2] = \delta$. Further, since W is ε-regular, by Hoeffding's inequality, for all $t > 0$,

$$\Pr\left[\left| \|W^p_{S(Y)}\|^2 - \delta \right| \geq \gamma \right] \leq 2 \exp\left(\frac{-2\gamma^2}{\sum_i W^4_{ip}} \right) \leq 2 \exp\left(\frac{-2\gamma^2}{\varepsilon^2} \right).$$

Thus, by a union bound

$$\Pr_Y \left[\exists p \in [k],\ \|W^p_{S(Y)}\|^2 \geq \delta + 2\sqrt{\log(k/\delta)}\, \varepsilon \right] \leq \delta. \tag{6.2.1}$$

Note that for a fixed Y and sufficiently large C, by Fact 6.2.2 and a union bound,

$$\Pr_{x \in_u \{1,-1\}^n} \left[\exists p \in [k],\ |\langle W^p_{S(Y)}, x_{S(Y)} \rangle| \geq C\sqrt{\log(k/\delta)}\, \|W^p_{S(Y)}\| \right] \leq \delta.$$

From Equation 6.2.1 and the above equation, we get that for a sufficiently large constant C

$$\Pr_{x \in_u \{1,-1\}^n, Y} \left[\exists p \in [k],\ |\langle W^p_{S(Y)}, x_{S(Y)} \rangle| \geq C \log(k/\delta)^{1/2} \delta^{1/2} + C \log(k/\delta)^{3/4} \varepsilon^{1/2} \right] \leq 2\delta. \tag{6.2.2}$$

Now, observe that that for $x \in \{1,-1\}^n$, to generate a δ-perturbation of x, $y(x)$, we can first generate a random Y as above and flip the bits of x in the support of Y. Thus, from Equation 6.2.2,

$$\Pr_{x\in_u\{1,-1\}^n,Y}\left[\,\exists p\in[k]\,|\langle W^p,x\rangle-\langle W^p,y(x)\rangle|\geq\lambda\,\right]=$$
$$\Pr_{x\in_u\{1,-1\}^n,Y}\left[\,\exists p\in[k]\,|\,\langle W^p_{S(Y)},x_{S(Y)}\rangle|\geq\lambda\,\right]\leq 2\delta,$$

where $\lambda = C\log(k/\delta)^{1/2}\delta^{1/2} + C\log(k/\delta)^{3/4}\varepsilon^{1/2}$. Therefore,

$$\Pr_{x\in_u\{1,-1\}^n,Y}\left[\,\|W^Tx - W^Ty(x)\|_\infty\geq\lambda\,\right]\leq 2\delta.$$

□

The following claim can be seen as an anti-concentration bound for regular polytopes over the hypercube and could be of use elsewhere.

Claim 6.2.4. *For ε-regular $W\in\mathbb{R}^{n\times k}$, $\theta\in\mathbb{R}^k$, and $0<\lambda<1$,*

$$\Pr_{x\in_u\{1,-1\}^n}[\,W^Tx\in\mathsf{Rect}(\theta+\lambda\,1_k)\setminus\mathsf{Rect}(\theta-\lambda\,1_k)\,]\leq$$
$$C(\log^{1.6}k)\,(\varepsilon\,\log(1/\varepsilon))^{1/5}+\sqrt{\log k}\,\lambda.$$

Proof. Follows directly from Theorem 3.3.1 and Lemma 3.3.4. □

We can now prove Theorem 6.2.1.

of Theorem 6.2.1. Note that for $x, y \in \mathbb{R}^n$, $f^{\wedge k}(x) \neq f^{\wedge k}(y)$ implies that $W^T x \in \mathsf{Rect}(\theta + \gamma 1_k) \setminus \mathsf{Rect}(\theta - \gamma 1_k)$, where $\gamma = \|W^T x - W^T y\|_\infty$. Hence,

$$\begin{aligned}
\mathbb{NS}_\delta(f^{\wedge k}) &= \mathop{\mathsf{Pr}}_{x \in_u \{1,-1\}^n, Y}[\, f^{\wedge k}(x) \neq f^{\wedge k}(y(x))\,] \\
&\leq \mathop{\mathsf{Pr}}_{x \in_u \{1,-1\}^n, Y}[\, f^{\wedge k}(x) \neq f^{\wedge k}(y(x)) \mid \|W^T x - W^T y(x)\|_\infty \leq \lambda\,] + 2\delta \\
&\qquad \text{(Claim 6.2.3)} \\
&\leq \mathop{\mathsf{Pr}}_{x \in_u \{1,-1\}^n}[\, W^T x \in \mathsf{Rect}(\theta + \lambda\, 1_k) \setminus \mathsf{Rect}(\theta - \lambda\, 1_k)\,] + 2\delta \\
&\leq C(\log^{1.6} k)\,(\varepsilon\, \log(1/\varepsilon))^{1/5} + \sqrt{\log k}\,\lambda + 2\delta. \\
&\qquad \text{(Claim 6.2.4)}
\end{aligned}$$

The theorem now follows. □

Chapter 7

Pseudorandom Generators for Polynomial Threshold Functions

7.1 Introduction

In this chapter we study the question of constructing PRGs for PTFs and present the first nontrivial generators for low-degree PTFs. Along the way we will develop pretty generic framework for obtaining PRGs from invariance principles and develop these techniques further to obtain PRGs for polytopes Chapter 8 and combinatorial shapes Chapter 9. .

We first recall the definition of a PRG specialized for the class of PTFs.

Definition 7.1.1. A function $G : \{0,1\}^r \to \{1,-1\}^n$ is a PRG with error ε for (or ε-fools) PTFs of degree d, if

$$\left| \mathop{\mathbb{E}}_{x \in_u \{1,-1\}^n} [f(x)] - \mathop{\mathbb{E}}_{y \in_u \{0,1\}^r} [f(G(y))] \right| \leq \varepsilon,$$

for all PTFs f of degree at most d.

We refer to the parameter r as the seed-length of the generator G and say the generator is explicit if it is computable by a (deterministic) polynomial time algorithm. It can be shown by the probabilistic method that there exist PRGs that ε-fool degree d PTFs with *seed length* $r = O(d \log n + \log(1/\varepsilon))$

(see Section 7.8). However, despite their long history, until recently very little was known about explicitly constructing such PRGs, even for the special class of halfspaces.

In this work, we present a PRG that ε-fools degree d PTFs with seed length $\log n/\varepsilon^{O(d)}$. Previously, PRGs with seed length $o(n)$ were not known even for degree 2 PTFs and constant ε.

Theorem 7.1.1. *For $0 < \varepsilon < 1$, there exists an explicit PRG fooling PTFs of degree d with error at most ε and seed length $2^{O(d)} \log n/\varepsilon^{8d+3}$.*

Independent of our work, Diakonikolas et al. [28] showed that bounded independence fools degree 2 PTFs and in particular give a PRG with seed-length $(\log n) \cdot \tilde{O}(1/\varepsilon^9)$ for degree 2 PTFs (here $\tilde{O}$ hides poly-logarithmic factors). In another independent work, Ben-Eliezer et al. [14] showed that bounded independence fools certain special classes of PTFs.

For the $d = 1$ case of halfspaces, Diakonikolas et al. [26] constructed PRGs with seed length $O(\log n)$ for constant error rates. PRGs with seed length $O(\log^2 n)$ for halfspaces with polynomially bounded weights follow easily from known results. However, nothing nontrivial was known for general halfspaces, for instance, when $\varepsilon = 1/\sqrt{n}$. In this work we construct PRGs with exponentially better dependence on the error parameter ε.

Theorem 7.1.2. *For all constants c, $\varepsilon \geq 1/n^c$, there exists an explicit PRG fooling halfspaces with error at most ε and seed length $O(\log n + \log^2(1/\varepsilon))$.*

We also obtain results similar to the above for spherical caps. The problem of constructing PRGs for spherical caps was brought to our attention by Amir Shpilka; Karnin et al. [55] were the first to obtain a PRG with similar parameters using different methods. They achieve a seed-length of $(1 + o(1)) \log n + O(\log^2(1/\varepsilon))$.

Theorem 7.1.3. *There exists a constant $c > 0$ such that for all $\varepsilon > c \log n / n^{1/4}$, there exists an explicit PRG fooling spherical caps with error at most ε and seed length $O(\log n + \log^2(1/\varepsilon))$.*

We briefly summarize the previous constructions for halfspaces.

1. Halfspaces with polynomially bounded integer weights can be computed by polynomial width read-once branching programs (ROBPs). Thus, the PRGs for ROBPs such as those of Nisan [80] and Impagliazzo et al. [46] fool halfspaces with polynomially bounded integer weights with seed length $O(\log^2 n)$. However, a simple counting argument ([69], [43]) shows that almost all halfspaces have exponentially large weights.

2. Diakonikolas et al. [26] showed that k-wise independent spaces fool halfspaces for $k = O(\log^2(1/\varepsilon)/\varepsilon^2)$. By using the known efficient constructions of k-wise independent spaces they obtain PRGs for halfspaces with seed length $O(\log n \log^2(1/\varepsilon)/\varepsilon^2)$.

3. Rabani and Shpilka [87] gave explicit constructions of polynomial size *hitting sets* for halfspaces.

The overarching theme behind all our constructions is the use of *invariance principles* to get pseudorandom generators. Intuitively, invariance principles could be helpful in constructing pseudorandom generators as we can hope to exploit the invariance with respect to product distributions by replacing a product distribution with a "smaller product distribution" that still satisfies the conditions for applying the invariance principle. We believe that the above technique could be helpful for other derandomization problems.

Another aspect of our constructions is what we call the "monotone trick". The PRGs for small-width read-once branching programs (ROBP) from the works of Nisan [80], Impagliazzo et al. [46], and Nisan and Zuckerman [81], have been a fundamental tool in derandomization with several applications [95], [90], [38]. An important ingredient in our PRG for halfspaces is our observation that any PRG for small-width ROBPs fools arbitrary width "monotone" ROBPs. Roughly speaking, we say an ROBP is monotone if there exists an ordering on the nodes in each layer of the program so that the corresponding sets of accepting strings *respect the ordering* (see Definition 7.2.1). We believe that this notion of monotone ROBP is quite natural and combined with the "monotone trick" could be useful elsewhere.

We now give a high level view of our constructions and their analyses.

7.1.1 Outline of Constructions

Our constructions build mainly on the hitting set construction for halfspaces of Rabani and Shpilka. Although the constructions and analyses are

similar in spirit for halfspaces and higher degree PTFs, for clarity, we deal with the two classes separately, at the cost of some repetition. The analysis is simpler for halfspaces and provides intuition for the more complicated analysis for higher degree PTFs.

7.1.1.1 PRGs for Halfspaces

Our first step in constructing PRGs for halfspaces is to use our "monotone trick" to show that PRGs for polynomial width read-once branching programs (ROBPs) also fool halfspaces. Previously, PRGs for polynomial width ROBPs were only known to fool halfspaces with polynomially bounded weights. Although the natural simulation of halfspaces by ROBP may require polynomially large width, we note that the resulting ROBP is what we call *monotone* (see Definition 7.2.1). We show that PRGs for polynomial width ROBP fool monotone ROBPs of arbitrary width.

Theorem 7.1.4. *A PRG that δ-fools monotone ROBP of width $\log(4T/\varepsilon)$ and length T fools monotone ROBP of arbitrary width and length T with error at most $\varepsilon + \delta$.*

See Theorem 7.2.1 for a more formal statement. As a corollary we get the following.

Corollary 7.1.5. *For all $\varepsilon > 0$, a PRG that δ-fools width $\log(4n/\varepsilon)$ and length n ROBPs fools halfspaces on n variables with error at most $\varepsilon + \delta$.*

The above result already improves on the previous constructions for small ε, giving a PRG with seed length $O(\log^2 n)$ for $\varepsilon = 1/\mathrm{poly}(n)$. However, the randomness used is $O(\log^2 n)$ even for constant ε.

We next improve the dependence of the seed length on the error parameter ε to obtain our main results for fooling halfspaces. Following the approach of Diakonikolas et al. [26] we first construct PRGs fooling *regular* halfspaces. A halfspace with coefficients $(w_1, \ldots, w_n)$ is regular if no coefficient is significantly larger than the others. Such halfspaces are easier to analyze because for regular w, the distribution of $\langle w, x \rangle$ with x uniformly distributed in $\{1, -1\}^n$ is close to a normal distribution by the Central Limit Theorem. Using a quantitative form of the above statement, the Berry-Esséen theorem, we show that a simplified version of the hitting set construction of Rabani and Shpilka gives a PRG fooling *regular* halfspaces.

Having fooled regular halfspaces, we use the structural results on halfspaces of Servedio [91] and Diakonikolas et al. [26] to fool arbitrary halfspaces. The structural results of Servedio and Diakonikolas et al. roughly show that either a halfspace is regular or is close to a function depending only on a small number of coordinates. Given this, we proceed by a case analysis as in Diakonikolas et al.: if a halfspace is regular, we use the analysis for regular halfspaces; else, we argue that bounded independence suffices.

The above analysis gives a PRG fooling halfspaces with seed length $O(\log n \log^2(1/\varepsilon)/\varepsilon^2)$, matching the PRG of Diakonikolas et al. [26]. However, not only is our construction simpler to analyze (for the regular case), but

we can also apply our "monotone trick" to derandomize the construction. Derandomizing using the PRG for ROBPs of Impagliazzo et al. [46] gives Theorem 7.1.2.

For spherical caps, we give a simpler more direct construction based on our generator for regular halfspaces. We use an idea of Ailon and Chazelle [1] and the invariance of spherical caps with respect to unitary rotations to convert the case of arbitrary spherical caps to *regular spherical caps*. We defer the details to Section 7.6.

7.1.1.2 PRGs for PTFs

We next extend our PRG for halfspaces to fool higher degree polynomial threshold functions. The construction we use to fool PTFs is a natural extension of our *underandomized* PRG for halfspaces. The analysis, though similar in outline, is significantly more complicated and at a high level proceeds as follows.

As was done for halfspaces we first study the case of regular PTFs. The mainstay of our analysis for regular halfspaces is the Berry-Esséen theorem for sums of independent random variables. By using the generalized Berry-Esséen type theorem, or *invariance principle*, for low-degree multi-linear polynomials, proved by Mossel et al. [76], we extend our analysis for regular halfspaces to regular PTFs. We remark that unlike the case for halfspaces, we cannot use the invariance principle of Mossel et al. directly, but instead adapt their proof technique for our generator. In particular, we crucially use the fact that most

of the arguments of Mossel et al. work even for distributions with bounded independence.

We then use structural results for PTFs of Diakonikolas et al. [30] and Harsha et al. [41] that generalize the results of Servedio [91] and Diakonikolas et al. [26] for halfspaces. Roughly speaking, these results show the following: with at least a constant probability, upon randomly restricting a small number of variables, the resulting restricted PTF is either regular or has high bias. However, we cannot yet use the above observation to do a case analysis as was done for halfspaces; instead, we give a more delicate argument with recursive application of the results on random restrictions.

We first present our result on fooling arbitrary width *monotone* ROBPs with PRGs for small-width ROBPs.

7.2 PRGs for Monotone ROBPs

We refer the reader to Section 2.3 to recall the definitions of read once branching programs, Definition 2.3.6 and the PRGs of Nisan [80] and Impagliazzo, Nisan and Wigderson [46], Theorem 2.3.1.

We also recall the following notations from Section 2.3. Let M be an (S, D, T)-branching program and v a vertex in layer i of M.

1. For $z = (z^i, z^{i+1}, \ldots, z^T) \in (\{0,1\}^D)^{T+1-i}$ call (v, z) an *accepting* pair if starting from v and traversing the path with edges labeled z in M leads to an accepting state.

2. For $z \in (\{0,1\}^D)^T$, let $M(z) = 1$ if (v_0, z) is an accepting pair, and $M(z) = 0$ otherwise.

3. $A_M(v) = \{z : (v, z) \text{ is accepting in } M\}$ and $P_M(v)$ is the probability that (v, z) is an accepting pair for z chosen uniformly at random.

4. For brevity, let $\mathcal{U}$ denote the uniform distribution over $(\{0,1\}^D)^T$.

Here we show that the PRGs for small-width ROBPs in fact fool arbitrary width monotone branching programs as defined below.

Definition 7.2.1 (Monotone ROBP)**.** An (S, D, T)-branching program M is said to be monotone if for all $0 \leq i < T$, there exists an ordering $\{v_1 \prec v_2 \prec \ldots \prec v_{l_i}\}$ of the vertices in layer i such that for $1 \leq j < k \leq l_i$, $A_M(v_j) \subseteq A_M(v_k)$.

Theorem 7.2.1. *Let $0 < \varepsilon < 1$ and $G : \{0,1\}^R \to (\{0,1\}^D)^T$ be a PRG that δ-fools monotone $(\log(2T/\varepsilon), D, T)$-branching programs. Then G fools monotone (S, D, T)-branching programs for arbitrary S with error at most $\varepsilon + \delta$.*

In particular, for $\delta = 1/\mathrm{poly}(T)$ the above theorem gives a PRG fooling monotone (S, D, T)-branching programs with error at most $\delta + \varepsilon$ and seed length $O(D + \log(T/\varepsilon) \log T)$. Note that the seed length does not depend on the space S. Given the above result, Corollary 7.1.5 follows easily.

Proof of Corollary 7.1.5. A halfspace with weight vector $w \in \mathbb{R}^n$ and threshold $\theta \in \mathbb{R}$ can be naturally computed by an $(S, 1, n)$-branching program $M_{w,\theta}$,

for S large enough, by letting the states in layer i correspond to the partial sums $\sum_{j=1}^{i} w_j x_j$. It is easy to check that $M_{w,\theta}$ is monotone. The theorem now follows from Theorem 7.2.1. □

We now prove Theorem 7.2.1. The proof is based on the simple idea of "sandwiching" monotone branching programs between small-width branching programs. To this end, let M be a monotone (S, D, T)-branching program and call a pair of (s, D, T)-branching programs (M_{down}, M_{up}), ε-sandwiching for M if the following hold.

1. For all $z \in (\{0,1\}^D)^T$, $M_{down}(z) \leq M(z) \leq M_{up}(z)$.
2. $\mathsf{Pr}_{z \leftarrow \mathcal{U}}[M_{up}(z) = 1] - \mathsf{Pr}_{z \leftarrow \mathcal{U}}[M_{down}(z) = 1] \leq \varepsilon$.

We first show that to fool monotone branching programs it suffices to fool small-width sandwiching programs between which the monotone branching program is sandwiched. We then show that every monotone branching program can be sandwiched between two small-width branching programs.

Lemma 7.2.2. *If a PRG G δ-fools (s, D, T)-branching programs, and there exist (s, D, T)-branching programs (M_{down}, M_{up}) that are ε-sandwiching for M, then G $(\varepsilon + \delta)$-fools M.*

Proof. Let $\mathcal{D}$ denote the output distribution of G. Then,

$$\mathop{\mathsf{Pr}}_{z \leftarrow \mathcal{U}}[M_{down}(z) = 1] \leq \mathop{\mathsf{Pr}}_{z \leftarrow \mathcal{U}}[M(z) = 1], \quad \mathop{\mathsf{Pr}}_{z \leftarrow \mathcal{D}}[M(z) = 1] \leq \mathop{\mathsf{Pr}}_{z \leftarrow \mathcal{D}}[M_{up}(z) = 1].$$

Further, since $\mathcal{D}$ δ-fools M_{up},

$$\Pr_{z \leftarrow \mathcal{D}}[M_{up}(z) = 1] \leq \Pr_{z \leftarrow \mathcal{U}}[M_{up}(z) = 1] + \delta.$$

Thus,

$$\Pr_{z \leftarrow \mathcal{D}}[M(z) = 1] - \Pr_{z \leftarrow \mathcal{U}}[M(z) = 1] \leq \Pr_{z \leftarrow \mathcal{U}}[M_{up}(z) = 1] - \Pr_{z \leftarrow \mathcal{U}}[M_{down}(z) = 1] + \delta \leq \varepsilon + \delta.$$

By a similar argument with the roles of M_{up}, M_{down} interchanged, we get

$$|\Pr_{z \leftarrow \mathcal{D}}[M(z) = 1] - \Pr_{z \leftarrow \mathcal{U}}[M(z) = 1]| \leq \varepsilon + \delta.$$

□

Lemma 7.2.3. *For any monotone (S, D, T)-branching program M, there exist $(\log(2T/\varepsilon), D, T)$-branching programs (M_{down}, M_{up}) that are ε-sandwiching for M.*

Proof. We first set up some notation. For $0 \leq i \leq T$, let the vertices in layer i of M be $V^i = \{v_1^i \prec v_2^i \prec \ldots \prec v_{l_i}^i\}$. For $J \subseteq V^i$, let $\min(J), \max(J)$ denote the minimum and maximum elements of J under $\prec$. Call $J \subseteq V^i$ an interval if there exist indices $p \leq q$ such that $J = \{v_p^i, v_{p+1}^i, \ldots, v_q^i\}$.

For each $1 \leq i \leq T$, partition the vertices of layer i into at most $t_i \leq 2T/\varepsilon$ intervals $J_1^i, J_2^i, \ldots, J_{t_i}^i$ so that for any interval J_k^i and $v, v' \in J_k^i$,

$$|P_M(v) - P_M(v')| \leq \frac{\varepsilon}{2T}. \tag{7.2.1}$$

Let $s = \log(2T/\varepsilon)$ and define an (s, D, T)-branching program M_{up} as follows. The vertices in layer i of M_{up} are $B^i = \{\max(J_1^i), \max(J_2^i), \ldots, \max(J_{t_i}^i)\}$ and

the edges are placed by *rounding* the edges of M *upwards* as follows. For $v \in B^i$ suppose there is an edge labeled z between v and a vertex $w \in J = J_k^{i+1}$. Then, we place an edge labeled z between v and $\max(J)$. M_{down} is defined similarly by using $\min(J)$ instead of $\max(J)$ as above. We claim that M_{up}, M_{down} are ε-sandwiching for M. We analyze M_{up} below; the analysis for M_{down} is similar.

Claim 7.2.4. *For $0 \leq i \leq T$ and $v \in B^i$, $A_M(v) \subseteq A_{M_{up}}(v)$. In particular, for any z, $M(z) \leq M_{up}(z)$.*

Proof. Follows from the monotonicity of M. □

Claim 7.2.5. *For $0 \leq i \leq T$, and $v \in B^i$, $P_{M_{up}}(v) - P_M(v) \leq (T-i)\frac{\varepsilon}{2T}$. In particular, for z chosen uniformly at random, $\mathsf{Pr}[M_{up}(z) = 1] - \mathsf{Pr}[M(z) = 1] \leq \varepsilon/2$.*

Proof. The second part of the claim follows from the first. The proof is by downward induction on i. For $i = T$, the statement is true trivially. Now, suppose the claim is true for all $j \geq i+1$. Let $v \in B^i$ and let $z = (z^{i+1}, \bar{z})$ be uniformly chosen from $(\{0,1\}^D)^{T-i}$ with $z^{i+1} \in_u \{0,1\}^D$. Let $\Gamma(v, z^{i+1}) \in J(v, z^{i+1}) = J_k^{i+1}$ for one of the intervals of layer $i+1$. Then, the edge labeled

z^{i+1} from v goes to $\max(J(v, z^{i+1}))$ in M_{up}. Now,

$$\begin{aligned}
P_M(v) &= \sum_{u \in \{0,1\}^D} \mathsf{Pr}[z^{i+1} = u]\, P_M(\Gamma(v,u)) \\
&\geq \sum_{u \in \{0,1\}^D} \mathsf{Pr}[z^{i+1} = u] \left(P_M(\max(J(v,u))) - \frac{\varepsilon}{2T} \right) \\
&\quad \text{(Equation (7.2.1))} \\
&\geq \sum_{u \in \{0,1\}^D} \mathsf{Pr}[z^{i+1} = u] \left(P_{M_{up}}(\max(J(v,u))) - \frac{(T-i-1)\varepsilon}{2T} - \frac{\varepsilon}{2T} \right) \\
&\quad \text{(Induction hypothesis)} \\
&= \sum_{u \in \{0,1\}^D} \mathsf{Pr}[z^{i+1} = u]\, P_{M_{up}}(\max(J(v,u))) - \frac{(T-i)\varepsilon}{2T} \\
&= P_{M_{up}}(v) - \frac{(T-i)\varepsilon}{2T}. \\
&\quad \text{(Definition of } M_{up})
\end{aligned}$$

The claim now follows from the above equation and induction. □

Lemma 7.2.3 now follows from Claims 7.2.4, 7.2.5 and similar arguments for M_{down}.

□

7.3 Main Generator Construction

We now describe our main construction G that serves as a blueprint for our constructions in this chapter as well as the next. The generator G is essentially a simplification of the hitting set construction for halfspaces by Rabani and Shpilka [87]. We use the following as building blocks. We refer to

Section 2.3 for the appropriate definitions.

- A family $\mathcal{H} = \{h : [n] \to [t]\}$ of hash functions that is α-almost pairwise independent (see Definition 2.3.1).
- A generator $G_k : \{0,1\}^{r_0} \to \{1,-1\}^m$ of a δ-almost k-wise independent space over $\{1,-1\}^m$ (see Definition 2.3.2).

Although efficient constructions of hash families $\mathcal{H}$ and generators G_k as above are known even for $\alpha = 0$, $\delta = 0$ and constant k, we work with small but non-zero α, δ, as we will need the more general objects for our analyses.

The basic idea behind the generator is as follows. We first use the hash functions to distribute the *coordinates* ($[n]$) into buckets. The purpose of this step is to spread out the "influences" of the coordinates across buckets. Then, for each bucket we use an independently chosen sample from a δ-almost k-wise independent distribution to generate the bits for the coordinate positions mapped to the bucket. The purpose of this step is, roughly, to "match the first few moments" of functions restricted to the coordinates in each bucket. The hope then is to subsequently use invariance principles to show closeness in distribution.

Fix the error parameter $\varepsilon > 0$ and let t at most $\text{poly}(\log(1/\varepsilon))/\varepsilon^2$ to be chosen later. Let $m = n/t$ (assuming without loss of generality that t divides n) and let $\mathcal{H}$ be an α-pairwise independent hash family. To avoid some technicalities that can be overcome easily, we assume that every hash function

$h \in \mathcal{H}$ is evenly distributed, meaning $\forall h, i \in [t]$, $|\{j : h(j) = i, j \in [n]\}| = n/t$. Let $G_k : \{0,1\}^{r_0} \to \{1,-1\}^m$ generate a δ-almost k-wise independent space for $\delta \geq \text{poly}(\varepsilon, 1/n)$ to be chosen later.

Define $G \equiv G_{\mathcal{H},k,t} : \mathcal{H} \times (\{0,1\}^{r_0})^t \to \{0,1\}^n$ by

$$G(h, z^1, \ldots, z^t) = x, \text{ where } x_{|h^{-1}(i)} = G_k(z^i) \text{ for } i \in [t]. \tag{7.3.1}$$

The generator $G_{\mathcal{H},k,t}$ will be used again in our construction of PRGs for polytopes, Chapter 8 and Chapter 9 as well. In this chapter we focus on fooling PTFs using $G_{\mathcal{H},k,t}$.

We will show that for the parameters t, α, δ, k and $\mathcal{H}, G_k$ chosen appropriately, the above generator fools halfspaces as well as degree d PTFs. In particular, we fool progressively stronger classes, from halfspaces to degree d PTFs by choosing $\mathcal{H}$ and G_k progressively stronger. The table below gives a simplified summary of the results we get for different choices of $\mathcal{H}, G_k$. We define *balanced* hash functions in Definition 7.4.2.

Hash Family $\mathcal{H}$	Generator G_k	Fooling class
Pairwise independent	4-wise independent	Regular halfspaces, Theorem 7.4.2
Pairwise independent, *Balanced*	$\Theta(\log t)$-wise independent	Halfspaces, Theorem 7.4.8
Pairwise independent	$4d$-wise independent	Regular degree d PTFs, Theorem 7.5.1
Pairwise independent, *Balanced*	$\Theta(t)$-wise independent	Degree d PTFs, Theorem 7.5.9.

7.4 PRGs for Halfspaces

In this section we show that for appropriately chosen parameters, G fools halfspaces. We first show that G fools "regular" halfspaces to obtain a PRG with seed length $O(\log n/\varepsilon^2)$ for regular halfspaces. We then extend the analysis to arbitrary halfspaces to get a PRG with seed length $O(\log n \log^2(1/\varepsilon)/\varepsilon^2)$ and apply the monotone trick to prove Theorem 7.1.2.

In the following let $H_{w,\theta} : \{1,-1\}^n \to \{1,-1\}$ denote a halfspace $H_{w,\theta}(x) = \mathsf{sign}(\langle w, x\rangle - \theta)$. Unless stated otherwise, we assume throughout that a halfspace $H_{w,\theta}$ is normalized, meaning $\|w\| = 1$ (here $\|\cdot\|$ is the l_2-norm). In the following, we say two real-valued distributions P, Q are ε-close if $\mathsf{d_{cdf}}(P,Q) \leq \varepsilon$. We use the fact that Kolmogorov-Smirnov distance is convex.

Lemma 7.4.1. *For fixed Q, the distance function $\mathsf{d_{cdf}}(P,Q)$ defined for probability distributions over $\mathbb{R}$ is a convex function.*

For $\sigma > 0$, let $\mathcal{N}(0,\sigma)$ denote the normal distribution with mean 0 and variance σ^2. We also assume that $\varepsilon > 1/n^{.49}$ as otherwise, Theorem 7.1.2 follows from Corollary 7.1.5.

7.4.1 PRGs for Regular Halfspaces

As was done in Diakonikolas et al. we first deal with regular halfspaces.

Definition 7.4.1. A vector $w \in \mathbb{R}^n$ ε-regular if $|w_i| \leq \varepsilon\|w\|$ for all i. A halfspace $H_{w,\theta}$ is ε-regular if w is ε-regular.

Let $t = 1/\varepsilon^2$. We claim that for $\mathcal{H}$ pairwise independent and G_k generating an almost 4-wise independent distribution, G fools regular halfspaces. Note that the randomness used by G in this setting is $O(\log n/\varepsilon^2)$.

Theorem 7.4.2. *Let $\mathcal{H}$ be an α-almost pairwise independent family for $\alpha = O(1)$ and let G_k generate a δ-almost 4-wise independent distribution for $\delta = \varepsilon^2/4n^5$. Then, G defined by Equation 7.3.1 fools ε-regular halfspaces with error at most $O(\varepsilon)$ and seed length $O(\log n/\varepsilon^2)$. In particular, for $x \in \{1,-1\}^n$ generated from G and ε-regular w with $\|w\| = 1$, the distribution of $\langle w, x\rangle$ is $O(\varepsilon)$-close to $\mathcal{N}(0,1)$.*

To prove the theorem we will need the following corollary of the Berry-Esséen theorem, Theorem 2.2.4.

Corollary 7.4.3. *Let $Y_1, \ldots, Y_t$ be independent random variables with $E[Y_i] = 0$, $\sum_i E[Y_i^2] = \sigma^2$, $\sum_i E[|Y_i|^4] \leq \rho_4$. Let $F(.)$ denote the cdf of the random variable $S_n = (Y_1 + \ldots Y_n)/\sigma$, and $\Phi(.)$ denote the cdf of the normal distribution $\mathcal{N}(0,1)$. Then,*

$$\|F - \Phi\|_\infty = \sup_z |F(z) - \Phi(z)| \leq \frac{\sqrt{\rho_4}}{\sigma^2}.$$

Proof. For $1 \leq i \leq n$, by Cauchy-Schwarz, $E[|Y_i|^3] \leq \sqrt{E[Y_i^2]} \cdot \sqrt{E[Y_i^4]}$. Therefore,

$$\sum_i E[|Y_i|^3] \leq \sum_i \sqrt{E[Y_i^2]} \cdot \sqrt{E[Y_i^4]} \leq \left(\sum_i E[Y_i^2]\right)^{1/2} \left(\sum_i E[Y_i^4]\right)^{1/2}.$$

The claim now follows from Theorem 2.2.4. □

Lemma 7.4.4. *For ε-regular w with $\|w\| = 1$ and $x \in_u \{1, -1\}^n$, the distribution of $\langle w, x\rangle$ is ε-close to $\mathcal{N}(0, 1)$.*

Proof. Let $Y_i = w_i x_i$. Then, $\sum_i \mathbb{E}[Y_i^2] = 1$ and $\sum_i \mathbb{E}[Y_i^4] = \sum_i w_i^4 \leq \varepsilon^2$. The lemma now follows from Corollary 7.4.3. □

The following lemma says that for a pairwise-independent family of hash functions $\mathcal{H}$ and $w \in \mathbb{R}^n$, the weight of the coefficients is *almost equidistributed* among the buckets.

Lemma 7.4.5. *Let $\mathcal{H}$ be an α-almost pairwise independent family of hash functions from $[n]$ to $[t]$. For ε-regular w with $\|w\| = 1$, $\sum_{i=1}^{t} \mathbb{E}[\|w_{h^{-1}(i)}\|^4] \leq (1 + \alpha)\varepsilon^2 + \frac{1+\alpha}{t}$.*

Proof. Fix $i \in [t]$. For $1 \leq j \leq n$, let X_j be the indicator variable that is 1 if $h(j) = i$ and 0 otherwise. Then, $\mathbb{E}[\|w_{h^{-1}(i)}\|^2] = 1/t$ and

$$\|w_{h^{-1}(i)}\|^4 = \left(\sum_{j=1}^{n} (X_j w_j)^2\right)^2 = \sum_{j=1}^{n} X_j^4 w_j^4 + \sum_{j \neq k} X_j^2 X_k^2 w_j^2 w_k^2.$$

Now, $\mathbb{E}[X_j^4] \leq (1 + \alpha)/t$ and for $j \neq k$, $\mathbb{E}[X_j^2 X_k^2] \leq (1 + \alpha)/t^2$. Thus, taking expectations of the above equation,

$$\begin{aligned}
\mathbb{E}[\|w_{h^{-1}(i)}\|^4] &\leq \frac{1+\alpha}{t} \sum_j w_j^4 + \frac{1+\alpha}{t^2} \sum_{j \neq k} w_j^2 w_k^2 \\
&\leq \frac{1+\alpha}{t} (\max_i |w_i|^2) + \frac{1+\alpha}{t^2} \\
&\leq \frac{(1+\alpha)\,\varepsilon^2}{t} + \frac{1+\alpha}{t^2}.
\end{aligned}$$

The lemma follows by summing over all $i \in [t]$. □

Proof of Theorem 7.4.2. Fix a hash function $h \in \mathcal{H}$. Let $w^i = w_{|h^{-1}(i)}$ for $i \in [t]$. Then,

$$\langle w, G(h,z)\rangle = \sum_{i=1}^{t} \langle w^i, G_k(z^i)\rangle.$$

Let random variables $Y_i^h \equiv Y_i \equiv \langle w^i, G_k(z^i)\rangle$ and $Y^h = Y_1 + \ldots + Y_t$. Then, $\mathbb{E}[Y_i] = 0$ and since $G_k(z^i)$ is δ-almost 4-wise independent, $|\,\mathbb{E}[Y_i^2] - \|w^i\|^2| \leq \delta n^2$. Further, for $1 \leq i \leq t$,

$$\mathop{\mathbb{E}}_{x \in_u \{1,-1\}^m}[\,\langle w^i, x\rangle^4\,] = \sum_{j=1}^{m} (w_j^i)^4 + 3 \sum_{p \neq q \in [m]} (w_p^i)^2 (w_q^i)^2 \leq 3\|w^i\|^4.$$

Since, the above equation depends only on the first four moments of random variable x and $G_k(Z^i)$ is δ-almost 4-wise independent, it follows that $\mathbb{E}[Y_i^4] \leq 3\|w^i\|^4 + \delta n^4$. Thus, $\sum_i \mathbb{E}[Y_i^2] \geq 1 - \delta n^2 t \geq 1/2$ and $\sum_{i=1}^t \mathbb{E}[Y_i^4] \leq 3\sum_{i=1}^t \|w^i\|^4 + \delta n^5$. Let $\rho_h = \sum_i \|w^i\|^4$. Then, by Corollary 7.4.3, since $\delta \leq \varepsilon^2/4n^5$, for a fixed h the distribution of Y^h is $(\sqrt{3\rho_h} + \varepsilon)$-close to $\mathcal{N}(0,1)$.

Observe that for random h, z the distribution of $Y = \langle w, G(h,z)\rangle$ is a convex-combination of the distributions of Y^h for $h \in \mathcal{H}$. Thus, from Lemma 7.4.1, the distribution of Y is $O(\mathbb{E}[\sqrt{\rho_h}] + \varepsilon)$-close to $\mathcal{N}(0,1)$. Now, by Cauchy-Schwarz $\mathbb{E}[\sqrt{\rho_h}] \leq \sqrt{\mathbb{E}[\rho_h]}$. Further, since w is ε-regular and $t = 1/\varepsilon^2$, it follows from Lemma 7.4.5 that $\mathbb{E}[\rho_h] = \sum_i \mathbb{E}[\|w^i\|^4] = \sum_i \mathbb{E}[\|w_{h^{-1}(i)}\|^4] \leq 2(1+\alpha)\varepsilon^2$. Thus, the distribution of Y is $O(\varepsilon)$-close to $\mathcal{N}(0,1)$. The theorem now follows from combining this with Lemma 7.4.4. □

7.4.2 PRGs for Arbitrary Halfspaces

We now study arbitrary halfspaces and show that the generator G fools arbitrary halfspaces if the family of hash functions $\mathcal{H}$ and generator G_0 satisfy certain stronger properties. We use the following structural result on halfspaces that follows from the results of Servedio [91] and Diakonikolas et al. [26].

Theorem 7.4.6. *Let $H_{w,\theta}$ be a halfspace with $w_1 \geq \ldots \geq w_n$, $\sum w_i^2 = 1$. There exists $K = K(\varepsilon) = O(\log^2(1/\varepsilon)/\varepsilon^2)$ such that one of the following two conditions holds.*

1. *$w^K = (w_{K(\varepsilon)+1}, \ldots, w_n)$ is ε-regular.*
2. *Let $w' = (w_1, \ldots, w_{K(\varepsilon)})$ and let $H_{w',\theta}(x) = \mathsf{sgn}(\sum_{i=1}^{K} w_i x_i - \theta)$. Then,*

$$|\Pr_{x \leftarrow \mathcal{D}}[H_{w,\theta}(x) \neq \mathsf{H}_{w',\theta}(x)]| \leq 2\varepsilon, \tag{7.4.1}$$

where $\mathcal{D}$ is any distribution satisfying the following conditions for $x \leftarrow \mathcal{D}$.

   (a) The distribution of $(x_1, \ldots, x_K)$ is ε-close to uniform.

   (b) With probability at least $1 - \varepsilon$ over the choice of $(x_1, \ldots, x_K)$, the distribution of $(x_{K+1}, \ldots, x_n)$ conditioned on $(x_1, \ldots, x_K)$ is $(1/n^2)$-almost pairwise independent.

In particular, for distributions $\mathcal{D}$ as above

$$|\mathop{\mathbb{E}}_{x \leftarrow \mathcal{D}}[H_{w,\theta}(x)] - \mathop{\mathbb{E}}_{x \leftarrow \mathcal{D}}[\mathsf{H}_{w',\theta}(x)]| \leq 2\varepsilon. \tag{7.4.2}$$

Servedio and Diakonikolas et al. show the above result when $\mathcal{D}$ is the uniform distribution. However, their arguments extend straightforwardly to any distribution $\mathcal{D}$ as above.

Given the above theorem, we use a case analysis to analyze G. If the first condition of the theorem above holds, we use the results of the previous section, Theorem 7.4.2, showing that G fools regular halfspaces. If the second condition holds, we argue that for x distributed as the output of the generator, the distribution of $(x_1, \ldots, x_{K(\varepsilon)})$ is $O(\varepsilon)$-close to uniform.

Let $t = K(\varepsilon)$. We need the family of hash functions $\mathcal{H} : [n] \to [t]$ in the construction of G to be *balanced* along with being α-almost pairwise independent. Intuitively, a hash family is balanced if with high probability the maximum size of a bucket is small.

Definition 7.4.2 (Balanced Hash Functions)**.** A family of hash functions $\mathcal{H} = \{h : [n] \to [t]$ is (K, L, β)-balanced if for any $S \subseteq [n]$, $|S| \leq K$,

$$\Pr_{h \in_u \mathcal{H}}[\max_{j \in [t]} (|h^{-1}(j) \cap S|) \geq L] \leq \beta. \tag{7.4.3}$$

We use the following construction of balanced hash families due to Lovett et al. [67].

Theorem 7.4.7 (See Lemma 2.12 in [67])**.** *Let $t = \log(1/\varepsilon)/\varepsilon^2$ and $K = K(\varepsilon)$ as in Theorem 7.4.6. Then, there exists a $(K, O(\log(1/\varepsilon)), 1/t^2)$-balanced hash family $\mathcal{H} : [n] \to [t]$ that is also pairwise independent with $|\mathcal{H}| = \exp(O(\log n + \log^2(1/\varepsilon)))$. Moreover, $\mathcal{H}$ is efficiently samplable.*

Let $m = n/t$ and fix L to be one of $O(\log t), O(\log n)$. We also need the generator $G_0 : \{0,1\}^{r_0} \to \{1,-1\}^m$ to be exactly 4-wise independent and δ-almost $(L+4)$-wise independent for $\delta = \varepsilon^3/tn^5$. From Definition 2.3.2, we have generators G_0 as above with $r_0 = O(\log n + \log(1/\delta) + L) = O(\log(n/\varepsilon))$.

We now show that with $\mathcal{H}, G_0$ as above, G fools halfspaces with error $O(\varepsilon)$. The randomness used by the generator is $\log|\mathcal{H}| + r_0 t = O(\log n \log^2(1/\varepsilon)/\varepsilon^2)$ and matches the randomness used in the results of Diakonikolas et al. [26].

Theorem 7.4.8. *With $\mathcal{H}, G_0$ chosen as above, G defined by Equation (7.3.1) fools halfspaces with error at most $O(\varepsilon)$ and seed length $O(\log n \log^2(1/\varepsilon)/\varepsilon^2)$.*

Proof. Let $H_{w,\theta}$ be a halfspace and without loss of generality suppose that $w_1 \geq \ldots \geq w_n$ and $\sum_i w_i^2 = 1$. Let $S = \{1, \ldots, K(\varepsilon)\}$. Call a hash function *$S$-good* if for all $j \in [t]$, $|S_j| = |S \cap h^{-1}(j)| \leq L$. From Definition 7.4.2, a random hash function $h \in_u \mathcal{H}$ is S-good with probability at least $1 - 1/t^2$. Recall that $G(h, z^1, \ldots, z^t) = x$, where $x_{|h^{-1}(j)} = G_0(z^j)$ for $j \in [t]$. Let $\mathcal{D}$ denote the distribution of the output of G and let $x \leftarrow \mathcal{D}$.

Claim 7.4.9. *Given an S-good hash function h, the distribution of $x_{|S}$ is ε-close to uniform. Moreover, with probability at least $1 - \varepsilon$ over the random choices of $x_{|S}$, the distribution of x in the coordinates not in S conditioned on $x_{|S}$ is $(\varepsilon^2/4n^5)$-almost 4-wise independent.*

Proof. Fix an S-good hash function h. Since $z^1, \ldots, z^t$ are chosen independently, given the hash function h, $x|_{S_1}, \ldots, x|_{S_t}$ are independent of each other.

Moreover, since the output of G_0 is δ-almost $(L+4)$-wise independent and $|S_j| \leq L$ for all $j \in [t]$, $x|_{S_j}$ is δ-close to uniform for all $j \in [t]$. It follows that given an S-good hash function h, $x|_S$ is $(t\delta)$-close to uniform. Further, by a similar argument, for any set $I \subseteq [n] \setminus S$ with $|I| = 4$, the distribution of $x_{|(S \cup I)}$ is $(t\delta)$-close to uniform. It follows that, with probability at least $1 - \varepsilon$, the distribution of $x_{|I}$ conditioned on $x_{|S}$ is $(t\delta/\varepsilon)$-close to uniform. The claim now follows from the above observations and noting that $t\delta = \varepsilon^3/4n^5$. □

We can now prove the theorem by a case analysis. Suppose that the weight vector w satisfies condition (2) of Theorem 7.4.6. Observe that from the above claim, $\mathcal{D}$ satisfies the conditions of Theorem 7.4.6 (2). Let $\mathsf{H}_{w_{|S},\theta}(x) = \mathsf{sgn}(\langle w_{|S}, x_{|S} \rangle - \theta)$. Then, from Equation (7.4.2),

$$\begin{aligned} |\mathop{\mathbb{E}}_{x \leftarrow U_n}[H_{w,\theta}(x)] - \mathop{\mathbb{E}}_{x \leftarrow U_n}[\mathsf{H}_{w_{|S},\theta}(x)]| &\leq 2\varepsilon, \\ |\mathop{\mathbb{E}}_{x \leftarrow \mathcal{D}}[H_{w,\theta}(x)] - \mathop{\mathbb{E}}_{x \leftarrow \mathcal{D}}[\mathsf{H}_{w_{|S},\theta}(x)]| &\leq 2\varepsilon. \end{aligned}$$

Moreover, since the distribution of $x_{|S}$ is ε-close to uniform under $\mathcal{D}$ and $\mathsf{H}_{w_{|S},\theta}(x)$ only depends on $x_{|S}$,

$$|\mathop{\mathbb{E}}_{x \leftarrow U_n}[\mathsf{H}_{w_{|S},\theta}(x)] - \mathop{\mathbb{E}}_{x \leftarrow \mathcal{D}}[\mathsf{H}_{w_{|S},\theta}(x)]| \leq \varepsilon.$$

Combining the above three equations, we get that

$$|\mathop{\mathbb{E}}_{x \leftarrow U_n}[H_{w,\theta}(x)] - \mathop{\mathbb{E}}_{x \leftarrow \mathcal{D}}[H_{w,\theta}(x)]| \leq 5\varepsilon,$$

and thus G fools halfspace $H_{w,\theta}$ with error at most 5ε.

Now suppose that condition (1) of Theorem 7.4.6 holds and $w_{\bar{S}} = (w_{K(\varepsilon)+1}, \ldots, w_n)$ is ε-regular. Fix an assignment to the variables $x_{|S} = u_{|S}$ and let $x_{\bar{S}} = (x_{k+1}, \ldots, x_n)$ and $H_u(x_{k+1}, \ldots, x_n) = \mathsf{sgn}(\langle w_{\bar{S}}, x_{\bar{S}} \rangle - \theta_u)$, where $\theta_u = \theta - \langle w_{|S}, x_{|S} \rangle$. We will argue that with probability at least $1 - \varepsilon$, conditioned on the values of $x_{|S}$, the output of G fools the ε-regular halfspace H_u with error $O(\varepsilon)$. Given the last statement it follows that $\mathcal{D}$ fools the halfspace $H_{w,\theta}$ with error $O(\varepsilon)$ since the distribution of $x_{|S}$ under $\mathcal{D}$ is ε-close to uniform.

Since $\mathcal{H}$ is a family of pairwise independent hash functions and a random hash function $h \in_u \mathcal{H}$ is S-good with probability at least $1 - 1/t^2$, even when conditioned on being S-good, a random hash function $h \in_u \mathcal{H}$ is α-pairwise independent for $\alpha = 1$. Further, from Claim 7.4.9, conditioned on the hash function h being S-good, with probability at least $1 - \varepsilon$, even conditioned on $x_{|S}$, the distribution of $x_{|[n]\setminus S}$ is $(\varepsilon^2/4n^5)$-almost 4-wise independent. Thus, we can apply Theorem 7.4.2[1] showing that with probability at least $1 - \varepsilon$, conditioned on the values of $x_{|S}$, the output of G fools H_u with error $O(\varepsilon)$. □

7.4.3 Derandomizing G

We now derandomize the generator from the previous section and prove Theorem 7.1.2. The derandomization is motivated by the fact that for a fixed hash function h and $w \in \mathbb{R}^n, \theta \in \mathbb{R}$, $\mathsf{sgn}(\langle w, G(h, z^1, \ldots, z^t) \rangle - \theta)$ can be computed by a monotone ROBP with t layers. Given this observation, by

[1] Though Theorem 7.4.2 was stated for $t = 1/\varepsilon^2$, the same argument works for all $t \geq 1/\varepsilon^2$.

Theorem 7.2.1, we can use PRGs for small-width ROBP to generate $z^1, \ldots, z^t$ instead of generating them independently as before.

Let $r_0, t, m, \mathcal{H}, G_0$ be set as in the context of Theorem 7.4.8. Let $s_0 = \log(2t/\varepsilon) = O(\log(1/\varepsilon))$ and let $G_{BP} : \{0,1\}^r \to (\{0,1\}^s)^t$ be a PRG fooling (s_0, r_0, t)-branching programs with error δ. Define $G_{\mathsf{D}} : \mathcal{H} \times \{0,1\}^r \to \{1,-1\}^n$ by

$$G_{\mathsf{D}}(h, y) = G(h, G_{BP}(y)). \tag{7.4.4}$$

The randomness used by the above generator is $\log|\mathcal{H}| + r$. We claim that G_{D} fools halfspaces with error at most $O(\varepsilon + \delta)$.

Theorem 7.4.10. *G_{D} fools halfspaces with error $O(\varepsilon + \delta)$.*

Proof. Fix a halfspace $H_{w,\theta}$ and without loss of generality (see [63] for instance) suppose that $w_1, \ldots, w_n, \theta$ are integers. Let $N = \sum_j |w_j| + |\theta|$. Observe that for any $x \in \{1,-1\}^n$, $\langle w, x\rangle - \theta \in \{-N, -N+1, \ldots, 0, \ldots, N\}$. Fix a hash function $h \in \mathcal{H}$. We define a $(\log(2N+1), r_0, t)$-branching program $M_{h,w}$ that for $z = (z^1, \ldots, z^t) \in (\{0,1\}^{r_0})^t$ computes $\langle w, G(h, z)\rangle$.

For $i \in [t]$, let $w^i = w_{|h^{-1}(i)}$. Then, for $z = (z^1, \ldots, z^t) \in (\{0,1\}^{r_0})^t$, by definition of G in Equation 7.3.1,

$$\langle w, G(h, z^1, \ldots, z^t)\rangle = \sum_{i=1}^{t} \langle w^i, G_0(z^i)\rangle.$$

Define a space-bounded machine $M_{h,w}$ as follows. For each $0 \le i \le t$, put N nodes in layer i with labels $1, \ldots, N$. The vertices in layer i correspond to the partial sums $Z_i = \sum_{l=1}^{i} \langle w^l, G_0(z^l)\rangle$. Note that all partial sums Z_i lie in

$\{-N, -N+1, \ldots, N\}$. Now, given the partial sum Z_i there are 2^{r_0} possible values for Z_{i+1} ranging in $\{Z_i + \langle w^{i+1}, G_0(z)\rangle : z \in \{0,1\}^{r_0}\}$. We add 2^{r_0} edges correspondingly. Finally, label all vertices in the final layer corresponding to values less than θ as rejecting and label all other vertices as accepting states.

It follows from the definition of $M_{h,w}$ that $M_{h,w}$ is monotone and for $z = (z^1, \ldots, z^t) \in (\{0,1\}^{r_0})^t$, $M_{h,w}(z)$ is an accepting state if and only if $\mathsf{sgn}(\sum_i \langle w^i, G_0(z^i)\rangle - \theta) = H_{w,\theta}(G(h,z)) = 1$. Thus, from Theorem 7.2.1, for a fixed $h \in \mathcal{H}$,

$$\left| \Pr_{z \in_u (\{0,1\}^{r_0})^t} [H_{w,\theta}(G(h,z)) = 1] - \Pr_{y \in_u \{0,1\}^r} [H_{w,\theta}(G(h, G_{BP}(y))) = 1] \right| \leq \delta + \varepsilon.$$

The theorem now follows from the above equation and Theorem 7.4.8.

□

By choosing the hash family $\mathcal{H}$ from Theorem 7.4.7 and using the PRG of Impagliazzo et al. we get our main result for fooling halfspaces.

Proof of Theorem 7.1.2. Choose G_{BP} in the above theorem to be the PRG of Impagliazzo et al. [46]. To ε-fool (S, D, T)-ROBPs, the generator of Impagliazzo et al., Theorem 2.3.1, has a seed-length of $O(D + (S + \log(1/\varepsilon)) \log T)$. Thus, the seed-length of G_{BP} is $r = O(r_0 + (s_0 + \log(1/\varepsilon)) \log t) = O(\log n + \log^2(1/\varepsilon))$. The theorem follows by choosing the hash family $\mathcal{H}$ as in Theorem 7.4.7. □

7.5 PRGs for Polynomial Threshold Functions

We now extend our results from the previous sections to construct PRGs for degree d PTFs. We set the parameters of G as in Theorem 7.4.8, with the main difference being that we take G_k to generate a k-wise independent space for $k = O(\log^2(1/\varepsilon)/\varepsilon^{O(d)} + 4d)$ instead of $O(\log^2(1/\varepsilon)/\varepsilon^2)$ as was done for fooling halfspaces. The analysis of the construction is, however, more complicated and proceeds as follows.

1. We first use the invariance principle of Mossel et al. [76] to deal with *regular* PTFs.

2. We then use the structural results on random restrictions of PTFs of Diakonikolas et al. [30] and Harsha et al. [41], Theorem 5.3.4, to reduce the case of fooling arbitrary PTFs to that of fooling *regular* PTFs and functions depending only on a few variables.

We carry out the first step above by an extension of the hybrid argument of Mossel et al. where we replace blocks of variables instead of single variables as done by Mossel et al. For this part of the analysis, we also need the *anti-concentration* results of Carbery and Wright [21] for low-degree polynomials over Gaussian distributions.

The second step relies on properties of random restrictions of PTFs similar in spirit to those in Theorem 7.4.6 for halfspaces. Roughly speaking, we use the fact any PTF is a small-depth decision tree with regular PTFs

(or constant functions) at the leaves to show that a generator fooling regular PTFs and having bounded independence also fools arbitrary PTFs.

7.5.1 PRGs for Regular PTFs

Here we extend our result for fooling regular halfspaces, Theorem 7.4.2, to regular PTFs. We recall the definition of regular PTFs from Definition 2.2.2 to set up some notation.

Definition 7.5.1. Let $P(u_1, \ldots, u_n) = \sum_I \alpha_I \prod_{i \in I} u_i$ be a multi-linear polynomial of degree d. Let $\|P\|_2^2 = \sum_I \alpha_I^2$ and the influence of i'th coordinate $\tau_i(P) = \sum_{I \ni i} \alpha_I^2$. We say P is ε-regular if

$$\sum_i \tau_i(P)^2 \leq \varepsilon^2 \|P\|_2^2.$$

We say a polynomial threshold function $f(x) = \mathsf{sgn}(P(x) - \theta)$ is ε-regular if P is ε-regular.

Unless stated otherwise, we will assume throughout that P is normalized with $\|P\|_2^2 = 1$. Fix $d > 0$. Let $t = 1/\varepsilon^2, m = n/t$ and let $\mathcal{H}$ be an α-pairwise independent family as in Theorem 7.4.2. We assume $G_k : \{0,1\}^{r_0} \to \{1,-1\}^m$ generates a $4d$-wise independent space, generalizing the assumption of 4-wise independence used for fooling regular halfspaces.

Theorem 7.5.1. *Let $\mathcal{H}$ be an α-pairwise independent family for $\alpha = O(1)$ and let G_k generate a $4d$-wise independent distribution. Then, G defined by Equation (7.3.1) fools ε-regular PTFs of degree at most d with error at most $O(d\varepsilon^{2/(4d+1)})$.*

We first prove some useful lemmas. The first lemma is simple.

Lemma 7.5.2. *For a multi-linear polynomial P of degree d with $\|P\| = 1$, $\sum_j \tau_j(P) \leq d$.*

The following lemma generalizes Lemma 7.4.5 and says that for pairwise independent hash functions and regular polynomials, the total influence is *almost equidistributed* among the buckets.

Lemma 7.5.3. *Let $\mathcal{H} = \{h : [n] \to [t]\}$ be a α-pairwise independent family of hash functions. Let P be a multi-linear polynomial of degree d with coefficients $(\alpha_J)_{J \subseteq [n]}$ and $\|P\| \leq 1$. For $h \in \mathcal{H}$ let*

$$\tau(h, i) = \sum_{J \cap h^{-1}(i) \neq \emptyset} \alpha_J^2.$$

Then, for $h \in_u \mathcal{H}$

$$\mathop{\mathbb{E}}_{h} \left[\sum_{i=1}^{t} \tau(h, i)^2 \right] \leq (1 + \alpha) \sum_{j=1}^{n} \tau_j(P)^2 + \frac{(1 + \alpha) d^2}{t}. \tag{7.5.1}$$

Proof. Fix $i \in [t]$ and for $1 \leq j \leq n$, let X_j be the indicator variable that is 1 if $h(j) = i$ and 0 otherwise. For brevity, let $\tau_j = \tau_j(P)$ for $j \in [n]$. Now,

$$\begin{aligned}
\tau(h, i) = \sum_{J \cap h^{-1}(i) \neq \emptyset} \alpha_J^2 &= \sum_{J} \alpha_J^2 \left(\vee_{j \in J} X_j \right) \\
&\leq \sum_{J} \alpha_J^2 \left(\sum_{j \in J} X_j \right) \\
&= \sum_{j} X_j \sum_{J : J \ni j} \alpha_J^2 \\
&= \sum_{j} X_j \tau_j.
\end{aligned}$$

Thus,

$$\tau(h,i)^2 \leq \left(\sum_{j=1}^{n} X_j \tau_j\right)^2 = \sum_j X_j^2 \tau_j^2 + \sum_{j \neq k} X_j X_k \tau_j \tau_k.$$

Note that $\mathbb{E}[X_j] \leq (1+\alpha)/t$ and for $j \neq k$, $\mathbb{E}[X_j X_k] \leq (1+\alpha)/t^2$. Thus,

$$\begin{aligned}\mathbb{E}[\,\tau(h,i)^2\,] &\leq \frac{1+\alpha}{t}\sum_j \tau_j^2 + \sum_{j\neq k}\tau_j\tau_k\frac{1+\alpha}{t^2} \\ &\leq \frac{1+\alpha}{t}\sum_j \tau_j^2 + \frac{1+\alpha}{t^2}(\sum_j \tau_j)^2.\end{aligned}$$

The lemma follows by using Lemma 7.5.2 and summing over all $i \in [t]$. □

We use the following structural result of Mossel et al. [76] that reduces the problem of fooling threshold functions to that of fooling certain *nice* functions which are easier to analyze.

Definition 7.5.2. A function $\psi : \mathbb{R} \to \mathbb{R}$ is B-nice, if ψ is smooth and $|\psi''''(t)| \leq B$ for all $t \in \mathbb{R}$.

Lemma 7.5.4 (Mossel et al.)**.** *Let X, Y be two real-valued random variables such that the following hold.*

1. *For any interval $I \subseteq \mathbb{R}$ of length at most α, $\mathsf{Pr}[\,X \in I\,] \leq C\alpha^{1/d}$, where C is a constant independent of α.*

2. *For all 1-nice functions ψ, $|E[\psi(X)] - E[\psi(Y)]| \leq \varepsilon^2$.*

Then, for all $t > 0$, $|\ \mathsf{Pr}[X > t] - \mathsf{Pr}[Y > t]\ | \leq 2C\,\varepsilon^{2/(4d+1)}$.

The following theorem is a restatement of the main result of Mossel et al. who obtain the bound $O(d\,9^d\,\max_i \tau_i(P)))$ instead of the one below. However, their arguments extend straightforwardly to the following.

Theorem 7.5.5 (Mossel et al.)**.** *Let* P *be a multi-linear polynomial of degree at most* d *with* $\|P\| = 1$, $\overline{X} \leftarrow \mathcal{N}(0,1)^n$ *and* $\overline{Y} \in_u \{1,-1\}^n$. *Then, for any* 1*-nice function* ψ,

$$|\ \mathbb{E}[\psi(P(\overline{X}))] - \mathbb{E}[\psi(P(\overline{Y}))]\ | \leq \frac{9^d}{12} \sum_i \tau_i(P)^2.$$

We first prove Theorem 7.5.1, assuming the following lemma which says that the generator G fools nice functions of regular polynomials.

Lemma 7.5.6. *Let* P *be an* ε*-regular multi-linear polynomial of degree at most* d *with* $\|P\| = 1$. *Let* $\overline{Y} \in_u \{1,-1\}^n$ *and* $\overline{Z}$ *be distributed as the output of* G. *Then, for any* 1*-nice function* ψ,

$$|\ E[\psi(P(\overline{Y}))] - E[\psi(P(\overline{Z}))]\ | \leq \frac{1+\alpha}{6}\, d^2\, 9^d\, \varepsilon^2$$

Proof of Theorem 7.5.1. Let P be an ε-regular polynomial of degree at most d and let $\overline{X} \leftarrow \mathcal{N}(0,1)^n$. Let X, Y, Z be real-valued random variables defined by $X = P(\overline{X})$, $Y = P(\overline{Y})$ and $Z = P(\overline{Z})$. Then, by Theorem 7.5.5 and Lemma 7.5.6, for any 1-nice function ψ,

$$|E[\psi(X)] - E[\psi(Y)]| \leq \frac{9^d}{12}\varepsilon^2, \quad |E[\psi(Y)] - E[\psi(Z)]| \leq \frac{(1+\alpha)\, d^2\, 9^d\, \varepsilon^2}{6}.$$

Hence,

$$|E[\psi(X)] - E[\psi(Z)]| = O(d^2\, 9^d\, \varepsilon^2).$$

Further, by Theorem 2.2.6, for any interval $I \subseteq \mathbb{R}$ of length at most α, $\mathsf{Pr}[\, X \in I\,] = O(\, d\, \alpha^{1/d}\,)$. Therefore, we can apply, Lemma 7.5.4 to X, Y and X, Z to get

$$|\,\mathsf{Pr}[X > t] - \mathsf{Pr}[Y > t]| = O(d\, \varepsilon^{2/(4d+1)}), \quad |\,\mathsf{Pr}[X > t] - \mathsf{Pr}[Z > t]| = O(d\, \varepsilon^{2/(4d+1)}).$$

Thus,

$$|\,\mathsf{Pr}[Y > t] - \mathsf{Pr}[Z > t]| = O(d\, \varepsilon^{2/(4d+1)}).$$

□

Proof of Lemma 7.5.6. Fix a hash function $h \in \mathcal{H}$. Let $Z_1, \ldots, Z_t$ be t independent samples generated from the $4d$-wise independent space. Let $Y_1, \ldots, Y_t$ be t independent samples chosen uniformly from $\{1, -1\}^m$. We will prove the claim via a hybrid argument where we replace the blocks $Y_1, \ldots, Y_t$ with $Z_1, \ldots, Z_t$ progressively.

For $0 \leq i \leq t$, let X^i be the distribution with $X^i_{|h^{-1}(j)} = Z_j$ for $1 \leq j \leq i$ and $X^i_{|h^{-1}(j)} = Y_j$ for $i < j \leq t$. Then, for a fixed hash function h, X^0 is uniformly distributed over $\{1, -1\}^n$ and X^t is distributed as the output of the generator. For $i \in [t]$, let $\tau(h, i)$ be the influence of the i'th bucket under h,

$$\tau(h, i) = \sum_{J \cap h^{-1}(i) \neq \emptyset} \alpha_J^2.$$

Claim 7.5.7. *For* $1 \leq i \leq t$,

$$|\,\mathbb{E}[\psi(P(X^i))] - \mathbb{E}[\psi(P(X^{i-1}))]| \leq \frac{9^d}{12}\,\tau(h,i)^2.$$

We will use the following form of the classical Taylor series.

Fact 7.5.8. For any 1-nice function $\psi : \mathbb{R} \to \mathbb{R}$, $\alpha, \beta \in \mathbb{R}$

$$\psi(\alpha+\beta) = \psi(\alpha) + \psi'(\alpha)\beta + \frac{\psi''(\alpha)}{2}\beta^2 + \frac{\psi'''(\alpha)}{6}\beta^3 + err(\alpha,\beta),$$

where $|err(\alpha,\beta)| \leq \beta^4/24$.

Proof. Let $I = h^{-1}(i)$ be the variables that have been changed from X^{i-1} to X^i. Without loss of generality suppose that $I = \{1,\ldots,m\}$. Let

$$P(u_1,\ldots,u_n) = R(u_{m+1},\ldots,u_n) + \sum_{J:J\cap[m]\neq\emptyset} \alpha_J \left(\prod_{j\in J} u_j\right),$$

where $R(\ )$ is a multi-linear polynomial of degree at most d. Let $S(u_1,\ldots,u_m,$ $u_{m+1},\ldots,u_n)$ denote the degree d multi-linear polynomial given by the second term in the above expression.

Observe that X^{i-1}, X^i agree on coordinates not in $[m]$. Let $X^i = (Z_1,\ldots,Z_m,X_{m+1},\ldots,X_n) = (Z,X)$ and $X^{i-1} = (Y_1,\ldots,Y_m,X_{m+1},\ldots,X_n) = (Y,X)$. Then,

$$P(X^i) = R(X) + S(Z,X), \quad P(X^{i-1}) = R(X) + S(Y,X).$$

Now, by using the Taylor series expansion, Fact 7.5.8, for ψ at $R(X)$,

$$\mathbb{E}[\psi(P(X^i))] - \mathbb{E}[\psi(P(X^{i-1}))] = \mathbb{E}[\psi(R + S(Z,X))] - E[\psi(R + S(Y,X))]$$
$$= \mathbb{E}[\,\psi(R) + \psi'(R)S(Z,X) + \frac{\psi''(R)}{2}S(Z,X)^2 + \frac{\psi'''(R)}{6}S(Z,X)^3 \pm \{\leq \frac{1}{24}S(Z,X)^4\}\,] -$$
$$\mathbb{E}[\,\psi(R) + \psi'(R)S(Y,X) + \frac{\psi''(R)}{2}S(Y,X)^2 + \frac{\psi'''(R)}{6}S(Y,X)^3 \pm \{\leq \frac{1}{24}S(Y,X)^4\}\,]$$

Observe that X, Y, Z are independent of one another and are $4d$-wise independent individually. Since $S(\ )$ has degree at most d, it follows that for a fixed assignment of the variables $X_{m+1}, \ldots, X_n$ in X,

$$\mathbb{E}[S(Z,X)] = \mathbb{E}[S(Y,X)], \quad \mathbb{E}[S(Z,X)^2] = \mathbb{E}[S(Y,X)^2],$$

$$\mathbb{E}[S(Z,X)^3] = \mathbb{E}[S(Y,X)^3], \quad \mathbb{E}[S(Z,X)^4] = \mathbb{E}[S(Y,X)^4].$$

Combining the above equations we get

$$|\,\mathbb{E}[\psi(P(X^i))] - \mathbb{E}[\psi(P(X^{i-1}))]| \leq \frac{1}{12}\,\mathbb{E}[\,S(Y,X)^4\,]. \tag{7.5.2}$$

Now, using the fact that $S(\ )$ is a multi-linear polynomial of degree at most d and since (Y,X) is $4d$-wise independent, $\mathbb{E}[\,S(Y,X)^4\,] = \mathbb{E}[\,S(W)^4\,]$, where W is uniformly distributed over $\{1,-1\}^n$. Also note that

$$\begin{aligned}\mathbb{E}[S(W)^2] &= \mathbb{E}\left[\left(\sum_{J:J\cap[m]\neq\emptyset} \alpha_J \left(\prod_{j\in J} W_j\right)\right)^2\right] \\ &= \sum_{J:J\cap I\neq\emptyset} \alpha_J^2 \\ &= \tau(h,i).\end{aligned}$$

Therefore, using the $(2,4)$-hypercontractivity inequality, Lemma 2.2.2, $\mathbb{E}[S(W)^4] \leq 9^d\,\mathbb{E}[S(W)^2]^2$ and Equation (7.5.2),

$$\begin{aligned}|\,\mathbb{E}[\psi(P(X^i))] - \mathbb{E}[\psi(P(X^{i-1}))]| &\leq \frac{1}{12}\,\mathbb{E}[\,S(Y,X)^4\,] = \frac{1}{12}\,\mathbb{E}[\,S(W)^4\,] \\ &\leq \frac{9^d}{12}\,\mathbb{E}[S(W)^2]^2 = \frac{9^d}{12}\,\tau(h,i)^2.\end{aligned}$$

□

□

Proof of Lemma 7.5.6 Continued. From Claim 7.5.7, for a fixed hash function h we have

$$|\,\mathbb{E}[\psi(P(\overline{Y}))] - \mathbb{E}[\psi(P(\overline{Z}))]| \leq \sum_{i=1}^{t} |\,\mathbb{E}[\psi(P(X^i))] - \mathbb{E}[\psi(P(X^{i-1}))]| \leq \frac{9^d}{12}\sum_{i=1}^{t}\tau(h,i)^2.$$

Therefore, for $h \in_u \mathcal{H}$, using Lemma 7.5.3 and $t = 1/\varepsilon^2$,

$$|\,\mathbb{E}[\psi(P(\overline{Y}))] - \mathbb{E}[\psi(P(\overline{Z}))]| \leq \frac{9^d}{12}\,\mathop{\mathbb{E}}_{h}\left[\sum_{i}\tau(h,i)^2\right] = \frac{9^d}{12}\,(1+\alpha)(1+d^2)\varepsilon^2 \leq \frac{(1+\alpha)\,d^2\,9^d\,\varepsilon^2}{6}.$$

□

7.5.2 PRGs for Arbitrary PTFs

We now study the case of arbitrary degree d PTFs. As was done for halfspaces, we will show that the generator G of Equation (7.3.1) fools arbi-

trary PTFs if the family of hash functions $\mathcal{H}$ and generator G_0 satisfy stronger properties. For the case of PTFs we shall use Theorem 5.3.4 from Chapter 5.

Let $t = c_d c'_d \log^2(1/\varepsilon)/\varepsilon^2$, $m = n/t$, where c_d, c'_d are the constants from Theorem 5.3.4. We use a family of hash functions $\mathcal{H} : [n] \to [t]$ that are α-pairwise independent for $\alpha = O(1)$. We choose the generator $G_0 : \{0,1\}^{r_0} \to \{1,-1\}^m$ to generate a $(t+4d)$-wise independent space. Generators G_0 with $r_0 = O(t \log n)$ are known. We claim that with the above setting of parameter the generator G fools all degree d PTFs.

Theorem 7.5.9. *With $\mathcal{H}, G_0$ chosen as above, G defined by Equation (7.3.1) fools degree d PTFs with error at most $O(\varepsilon^{2/(4d+1)})$ and seed length $O_d(\log n \log^4(1/\varepsilon)/\varepsilon^4)$.*

The bound on the seed length of the generator follows directly from the parameter settings. By carefully tracing the constants involved in our calculations and those in the results of Harsha et al. we need, the exact seed length can be shown to be $a^d \log n \log^4(1/\varepsilon)/\varepsilon^4$ for a universal constant a.

Fix a polynomial P of degree d and a PTF $f(x) = \mathsf{sign}(P(x) - \theta)$ and let T denote the block-decision tree computing f as given by Theorem 5.3.4. Let $\mathcal{D}_{PTF}$ denote the output distribution of the generator G with parameters set as above. The intuition behind the proof of the theorem is as follows.

1. As $\mathcal{D}_{PTF}$ has sufficient bounded independence, the distribution on the leaf nodes of T obtained by taking a walk on T according to inputs chosen

from $\mathcal{D}_{PTF}$ is the same as the case when inputs are chosen uniformly. In particular, a random walk on T according to $\mathcal{D}_{PTF}$ leads to a (ε, d)-good leaf node with high probability.

2. As G fools regular PTFs by Theorem 7.5.1, $\mathcal{D}_{PTF}$ will fool the function f_ρ computed at a (ε, d)-good leaf node. We also need to address the subtle issue that we really need $\mathcal{D}_{PTF}$ to fool a regular PTF f_ρ even when conditioned on reaching a particular leaf node ρ.

We first set up some notation. For a leaf node $\rho \in T$, let $U_\rho = [n] \setminus V_\rho$ be the set of variables seen on the path to ρ and let a_ρ be the corresponding assignment of variables in U_ρ that lead to ρ. Further, given an assignment x, let $\mathrm{Leaf}(x)$ denote the leaf node reached by taking a walk according to x on T.

Lemma 7.5.10. *For any leaf node ρ of T,*

$$\Pr_{x \leftarrow \mathcal{D}_{PTF}}[\mathrm{Leaf}(x) = \rho] = \Pr_{x \in_u \{1,-1\}^n}[\mathrm{Leaf}(x) = \rho].$$

Proof. Observe that $\mathcal{D}_{PTF}$ is a t-wise independent distribution and that for any ρ, $|U_\rho| \leq c_d c'_d \log^2(1/\varepsilon)/\varepsilon^2 = t$. Thus,

$$\begin{aligned}\Pr_{x \leftarrow \mathcal{D}_{PTF}}[\mathrm{Leaf}(x) = \rho] &= \Pr_{x \leftarrow \mathcal{D}_{PTF}}[x_{|U_\rho} = a_\rho] = \frac{1}{2^{|U_\rho|}} \\ &= \Pr_{x \in_u \{1,-1\}^n}[x_{|U_\rho} = a_\rho] = \Pr_{x \in_u \{1,-1\}^n}[\mathrm{Leaf}(x) = \rho].\end{aligned}$$

□

Lemma 7.5.11. *Fix an (ε, d)-good leaf node ρ of T. Then,*

$$\left| \Pr_{x \leftarrow \mathcal{D}_{PTF}} [f_\rho(x_{|V_\rho}) = 1 \,|\, x_{|U_\rho} = a_\rho] - \Pr_{y \leftarrow \{1,-1\}^{V_\rho}} [f_\rho(y) = 1] \right| = O(\varepsilon^{2/(4d+1)}).$$

Proof. We consider two cases depending on which of the two conditions of Definition 5.3.2 f_ρ satisfies.

Case (1) - f_ρ has high bias. Note that $\mathcal{D}_{PTF}$ is a $(t + 4d)$-wise independent distribution. Since $|U_\rho| \leq t$, it follows that for $x \leftarrow \mathcal{D}_{PTF}$, even conditioned on $x_{|U_\rho} = a_\rho$, the distribution is $2d$-wise independent. The lemma then follows from the fact that for some $b \in \{1, -1\}$, f_ρ evaluates to b with high probability.

Case (2) - f_ρ is an ε-regular degree d PTF. We deal with this case by using Theorem 7.5.1. Let $x = G(h, z^1, \ldots, z^t)$ for $h \in_u \mathcal{H}$, $z^1, \ldots, z^t \in_u \{0,1\}^{r_0}$, so $x \leftarrow \mathcal{D}_{PTF}$ as in the definition of G. Let $h_\rho : V_\rho \to [t]$ be the restriction of a hash function h to indices in V_ρ. For brevity, let $x(\rho) = x_{|V_\rho}$ and let E_ρ be the event $x_{|U_\rho} = a_\rho$. We show that the distribution of $x(\rho)$, conditioned on E_ρ, satisfies the conditions of Theorem 7.5.1.

Observe that conditioning on E_ρ does not change the distribution of the hash function $h \in_u \mathcal{H}$ because $|U_\rho| \leq t$ and $\mathcal{D}_{PTF}$ is t-wise independent. Thus, even when conditioned on E_ρ, the hash functions h_ρ are almost pairwise independent. For a hash function h, $i \in [t]$, let $B_\rho(h, i) = h^{-1}(i) \setminus V_\rho = h_\rho^{-1}(i)$. Now, since G_0 generates a $(t + 4d)$-wise independent distribution, even conditioned on E_ρ, for a fixed hash function h, the random variables

$x(\rho)_{|B_\rho(h,1)}$, $x(\rho)_{|B_\rho(h,2)}$, $\ldots$, $x(\rho)_{|B_\rho(h,t)}$ are independent of one another. Moreover, each $x(\rho)_{|B_\rho(h,i)}$ is $4d$-wise independent for $i \in [t]$.

Thus, even conditioned on E_ρ, the distribution of $x(\rho)$ satisfies the conditions of Theorem 7.5.1 and hence fools the regular degree d PTF f_ρ with error at most $O(\varepsilon^{2/(4d+1)})$. The lemma now follows. □

Proof of Theorem 7.5.9. Observe that

$$\Pr_{x \leftarrow \{1,-1\}^n}[f(x) = 1] = \sum_{\rho \in Leaves(T)} \Pr_{x \in_u \{1,-1\}^n}[x_{|U_\rho} = a_\rho] \cdot \Pr_{y \leftarrow \{1,-1\}^{V_\rho}}[f_\rho(y) = 1].$$

Similarly,

$$\Pr_{x \leftarrow \mathcal{D}_{PTF}}[f(x) = 1] = \sum_{\rho \in Leaves(T)} \Pr_{x \leftarrow \mathcal{D}_{PTF}}[x_{|U_\rho} = a_\rho] \cdot \Pr_{x \leftarrow \mathcal{D}_{PTF}}[f_\rho(x_{|V_\rho}) = 1 \,|\, x_{|U_\rho} = a_\rho].$$

From the above equations and Lemma 7.5.10 it follows that

$$\begin{aligned} |\Pr_{x \leftarrow \{1,-1\}^n}[f(x) = 1] - \Pr_{x \leftarrow \mathcal{D}_{PTF}}[f(x) = 1]| \leq \\ \sum_{\rho \in Leaves(T)} \Pr_{x \leftarrow \mathcal{D}_{PTF}}[x_{|U_\rho} = a_\rho] \cdot \\ \left| \Pr_{x \leftarrow \mathcal{D}_{PTF}}[f_\rho(x_{|V_\rho}) = 1 \,|\, x_{|U_\rho} = a_\rho] - \Pr_{y \leftarrow \{1,-1\}^{V_\rho}}[f_\rho(y) = 1] \right|. \end{aligned}$$

Now, by Lemma 7.5.11 for any (ε, d)-good leaf ρ the corresponding term on the right hand side of the above equation is $O(\varepsilon^{2/(4d+1)})$. Further, from Theorem 5.3.4 we know that a random walk ends at a good leaf with probability at least $1 - \varepsilon$. It follows that

$$|\Pr_{x \leftarrow \{1,-1\}^n}[f(x) = 1] - \Pr_{x \leftarrow \mathcal{D}_{PTF}}[f(x) = 1]| \leq \varepsilon\, t = O(\varepsilon^{2/(4d+1)}).$$

□

Our main theorem on fooling degree d PTFs, Theorem 7.1.1, follows immediately from the above theorem.

7.6 PRGs for Spherical Caps

We now show how to extend the generator for fooling regular halfspaces and its analysis from Section 7.4.1 to get a PRG for spherical caps and prove Theorem 7.1.3.

Let μ be a discrete distribution (if not, let's suppose we can discretize μ) over a set $U \subseteq \mathbb{R}$. Also, suppose that for $X \leftarrow \mu$, $\mathbb{E}[X] = 0, \mathbb{E}[X^2] = 1, \mathbb{E}[|X|^3] = O(1)$. Given such a distribution μ, a natural approach for extending G to μ is to replace the k-wise independent space generator $G_0 : \{0,1\}^r \to \{1,-1\}^m$ from Equation (7.3.1) with a generator $G_\mu : \{0,1\}^r \to U^m$ that generates a k-wise independent space over U^m. It follows from the analysis of Section 7.4.1 that for G_μ chosen with appropriate parameters, the above generator fools regular halfspaces over μ^n. It then remains to fool non-regular halfspaces over μ^n. It is reasonable to expect that an analysis similar to that in Section 7.4.2 can be applied to μ^n, provided we have analogues of the results of Servedio and Diakonikolas et al., Theorem 7.4.6, for μ^n.

The above ideas can be used to get a PRG for spherical caps by noting that a) the uniform distribution over the sphere is close to a product of Gaussians (when the test functions are halfspaces) and b) analogues of Theorem 7.4.6 for product of Gaussians follow from known *anti-concentration* properties of the univariate Gaussian distribution. Building on the above ar-

gument, Gopalan et al. [37] recently obtained PRGs fooling halfspaces over "reasonable" product distributions. Here we take a different approach and give a simpler, more direct construction for spherical caps based on an idea of Ailon and Chazelle [1] and the invariance of spherical caps with respect to unitary rotations.

Let $\mathcal{S}_{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ denote the n-dimensional sphere. By a *spherical cap* $S_{w,\theta}$ we mean the section of $\mathcal{S}_{n-1}$ cut by a halfspace, i.e., $S_{w,\theta} \stackrel{\text{def}}{=} \{x : x \in \mathcal{S}_{n-1}, H_{w,\theta}(x) = 1\}$.

Definition 7.6.1. A function $G : \{0,1\}^r \to \mathcal{S}_{n-1}$ is said to ε-fool spherical caps if, for all spherical caps $S_{w,\theta}$,

$$|\Pr_{x \in_u \mathcal{S}_{n-1}}[x \in S_{w,\theta}] - \Pr_{y \in_u \{0,1\}^r}[G(y) \in S_{w,\theta}]| \leq \varepsilon.$$

Note that the uniform distribution over $\mathcal{S}_{n-1}$, $\mathcal{U}_{sp}$, is not a product distribution. We first show that $\mathcal{U}_{sp}$ is close to $\mathcal{N}(0, 1/\sqrt{n})^n$ when the test functions are halfspaces.

Lemma 7.6.1. *There exists a universal constant C such that for any halfspace $H_{w,\theta}$,*

$$|\Pr_{x \leftarrow \mathcal{U}_{sp}}[H_{w,\theta}(x) = 1] - \Pr_{x \leftarrow \mathcal{N}(0,1/\sqrt{n})^n}[H_{w,\theta}(x) = 1]| \leq \frac{C \log n}{n^{1/4}}.$$

In particular, for $x \leftarrow \mathcal{U}_{sp}$, the distribution of $\langle w, x \rangle$ is $O(\sqrt{\log n}/n^{1/4})$-close to $\mathcal{N}(0, 1/\sqrt{n})$.

Proof. Observe that for $x \leftarrow \mathcal{N}(0, 1/\sqrt{n})^n$, $x/\|x\|_2$ is distributed uniformly over $\mathcal{S}_{n-1}$. Thus,

$$\Pr_{x \in_u \mathcal{S}_{n-1}}[H_{w,\theta}(x) = 1] = \Pr_{x \leftarrow \mathcal{N}(0,1/\sqrt{n})^n}[H_{w,\theta}\left(\frac{x}{\|x\|_2}\right) = 1].$$

Now, for any $x \in \mathbb{R}^n$,

$$\left|\langle w, x\rangle - \frac{\langle w, x\rangle}{\|x\|_2}\right| = \frac{|\langle w, x\rangle|}{\|x\|_2} \cdot |\|x\|_2 - 1|.$$

Since for $x \leftarrow \mathcal{N}(0, 1/\sqrt{n})$, $\langle w, x\rangle$ is distributed as $\mathcal{N}(0, 1/\sqrt{n})$, for some constant c_1,

$$\Pr_{x \leftarrow \mathcal{N}(0,1/\sqrt{n})^n}\left[|\langle w, x\rangle| \geq \frac{c_1\sqrt{\log n}}{n^{1/2}}\right] \leq \frac{1}{n}.$$

Further, by well-known concentration bounds for the norm of a random Gaussian vector (see [62], for instance), it follows that for some constant $c_2 > 0$,

$$\Pr_{x \leftarrow \mathcal{N}(0,1/\sqrt{n})^n}\left[|\|x\|_2 - 1| \geq \frac{c_2\sqrt{\log n}}{n^{1/4}}\right] \leq \frac{1}{n},$$

Combining the above equations we get

$$\Pr_{x \leftarrow \mathcal{N}(0,1/\sqrt{n})^n}\left[\left|\langle w, x\rangle - \frac{\langle w, x\rangle}{\|x\|_2}\right| \geq \frac{c_1 c_2 \log n}{n^{3/4}}\right] \leq \frac{2}{n}.$$

Therefore, for $C = c_1 c_2$,

$$\Pr_{x \leftarrow \mathcal{N}(0,1/\sqrt{n})^n}\left[H_{w,\theta}\left(\frac{x}{\|x\|_2}\right) \neq H_{w,\theta}(x)\right] \leq$$
$$\Pr_{x \leftarrow \mathcal{N}(0,1/\sqrt{n})^n}\left[|\langle w, x\rangle - \theta| \leq \left|\langle w, x\rangle - \frac{\langle w, x\rangle}{\|x\|_2}\right|\right] \leq$$
$$\Pr_{x \leftarrow \mathcal{N}(0,1/\sqrt{n})^n}\left[|\langle w, x\rangle - \theta| \leq \frac{c_1 c_2 \log n}{n^{3/4}}\right] + \frac{2}{n} \leq \frac{C \log n}{n^{1/4}},$$

where the last inequality follows from the fact that $\langle w, x\rangle$ is distributed as $\mathcal{N}(0, 1/\sqrt{n})$ and for any interval $I \subseteq \mathbb{R}$, $\Pr_{x \leftarrow \mathcal{N}(0,1)}[x \in I] = O(|I|)$. □

Now, by Theorem 7.4.2, for ε-regular w and x generated from G with parameters as in Theorem 7.4.2, the distribution of $\langle w, x/\sqrt{n}\rangle$ is $O(\varepsilon)$-close to $\mathcal{N}(0, 1/\sqrt{n})$. It then follows from the above lemma that G ε-fools spherical caps $S_{w,\theta}$ when w is ε-regular and $\varepsilon \geq C \log n/n^{1/4}$. We now reduce the case of arbitrary spherical caps to *regular spherical caps*.

Observe that the *volume* of a spherical cap $S_{w,\theta}$ is invariant under rotations: for any unitary matrix $A \in \mathbb{R}^{n\times n}$ with $A^T A = I_n$,

$$\Pr_{x \leftarrow \mathcal{U}_{sp}}[x \in S_{w,\theta}] = \Pr_{x \leftarrow \mathcal{U}_{sp}}[Ax \in S_{w,\theta}].$$

We exploit this fact by using a family of rotations $\mathcal{R}$ of Ailon and Chazelle [1] which satisfies the property that for any $w \in \mathbb{R}^n$ and a random rotation $V \in_u \mathcal{R}$, Vw is regular with high probability. Let $H \in \mathbb{R}^{n\times n}$ be the normalized Hadamard matrix such that $H^T H = I_n$ and each entry $H_{ij} \in \{\pm 1/\sqrt{n}\}$. For a vector $x \in \mathbb{R}^n$, let $D(x)$ denote the diagonal matrix with diagonal entries given by x. Observe that for $x \in \{1, -1\}^n$, $HD(x)$ is a unitary matrix. Ailon and Chazelle (essentially) show that for any $w \in \mathbb{R}^n$ and $x \in_u \{1, -1\}^n$, $HD(x)w$ is $O(\sqrt{\log n}/\sqrt{n})$-regular. We derandomize their construction by showing that similar guarantees hold for x chosen from a 8-wise independent distribution.

Lemma 7.6.2. *For all $w \in \mathbb{R}^n$, $\|w\| = 1$, and $x \in \{1, -1\}^n$ chosen from an 8-wise independent distribution the following holds. For $v = HD(x)w$, $\gamma > 0$,*

$$\Pr[\sum_i v_i^4 \geq \frac{\gamma}{n}] = O\left(\frac{1}{\gamma^2}\right).$$

Proof. Let random variable $Z = \sum_i v_i^4$. Observe that each v_i is a linear function of x and

$$\mathbb{E}[v_i^2] = \mathbb{E}[\,(\sum_j H_{ij} x_j w_j\,)^2\,] = \sum_j H_{ij}^2 w_j^2 = \frac{1}{n}.$$

Note that since x is 8-wise independent, we can apply $(2,4)$-hypercontractivity, Lemma 2.2.2, to v_i. Thus,

$$\mathbb{E}[Z] = \sum_i \mathbb{E}[v_i^4] \leq 9 \sum_i \mathbb{E}[v_i^2]^2 \leq \frac{9}{n}.$$

Similarly, by $(2,4)$-hypercontractivity applied to the quadratics v_i^2, v_j^2,

$$\mathbb{E}[Z^2] = \sum_{i,j} \mathbb{E}[v_i^4 v_j^4] \leq \sum_{i,j} 9^2\, \mathbb{E}[v_i^4]\, \mathbb{E}[v_j^4] \leq 9^2\, \mathbb{E}[Z]^2 \leq \frac{9^4}{n^2}.$$

The lemma now follows from the above equation and Markov's inequality applied to Z^2. □

Combining the above lemmas we get the following analogue of Theorem 7.4.2 for spherical caps. Let G be as in Theorem 7.4.2 and let $\mathcal{D}$ be a 8-wise independent distribution over $\{1,-1\}^n$. Define $G_{sph} : \{1,-1\}^n \times \{0,1\}^r \to \mathcal{S}_{n-1}$ by

$$G_{sph}(x,y) = \frac{D(x)H^T G(y)}{\sqrt{n}}.$$

Theorem 7.6.3. *For any spherical cap $S_{w,\theta}$ with $\|w\| = 1$ and $\varepsilon > C \log n / n^{1/4}$,*

$$\left| \Pr_{z \leftarrow \mathcal{U}_{sp}} [\,\langle w, z\rangle \geq \theta\,] - \Pr_{x \leftarrow \mathcal{D}, y \in_u \{0,1\}^r} [\,\langle w, G_{sph}(x,y)\rangle \geq \theta\,] \right| = O(\varepsilon).$$

Proof. By Lemma 7.6.1, for $z \leftarrow \mathcal{U}_{sp}$, $\langle w, z\rangle$ is $O(\varepsilon)$-close to $\mathcal{N}(0, 1/\sqrt{n})$. Further, by applying Lemma 7.6.2 for $\gamma = 1/\sqrt{\varepsilon}$, we get that $v = HD(x)w$ is δ-regular with probability at least $1 - O(\varepsilon)$ for $\delta = 1/(\sqrt{n}\varepsilon^{1/4}) < \varepsilon$. Now, by Theorem 7.4.2 for v ε-regular and $y \in_u \{0,1\}^r$, the distribution of $\langle v, G(y)\rangle$ is $O(\varepsilon)$-close to $\mathcal{N}(0,1)$. The theorem now follows from combining the above claims and noting that $\langle v, G(y)/\sqrt{n}\rangle = \langle w, G_{sph}(x,y)\rangle$. □

Theorem 7.1.3 now follows from the above theorem and derandomizing G as done in Section 7.4.3 for proving Theorem 7.1.2.

7.7 Discussion on Bounded Independence Fooling PTFs

In this section we briefly discuss the approach of fooling PTFs by bounded independence. Here we focus only on the case of regular PTFs as given that bounded independence fools regular PTFs, it is easy to extend the result to arbitrary PTFs by arguments similar to those used in Sections 7.4.2, 7.5.2.

As mentioned in the introduction, Diakonikolas et al. [26] show that $\tilde{O}(1/\varepsilon^2)$-wise independent distributions fool halfspaces and Diakonikolas, Kane and Nelson [28] show that $\tilde{O}(1/\varepsilon^9)$-wise independent distributions fool degree two threshold functions. Unfortunately, we do not know any such results for degrees three and higher. This leads us to the following conjecture:

Conjecture 7.1. *There exists a constant C such that the following holds. For $d > 0, \varepsilon \in [0,1]$, let $\mathcal{D}$ be a $k(d,\varepsilon)$-wise independent distribution for $k(d,\varepsilon) =$*

$(d/\varepsilon)^C$. *Then, for every degree d polynomial $P : \mathbb{R}^n \to \mathbb{R}$ and $x \leftarrow \mathcal{D}$, $y \in_u \{1,-1\}^n$, $\mathsf{d_{cdf}}(P(x), P(y)) < \varepsilon$.*

The bound $k(d,\varepsilon) = (d/\varepsilon)^C$ is probably the best possible: Diakonikolas et al. show a $\Omega(1/\varepsilon^2)$ lower bound to ε-fool halfspaces. We remark that a result like the above with any reasonable function $k(d,\varepsilon)$ that does not depend on n would be interesting. The closest result we have in this direction is that of Kane [53] who shows that k-wise independent Gaussian distributions ε-fool degree d polynomial threshold functions in the Gaussian world for $k = O(\varepsilon^{-2^{O(d)}})$.

In this section we review two old results in probability literature and use them to show that bounded independence suffices to fool degree 1 and 2 threshold functions. For degree 1 we get the optimal bound of $k(\varepsilon) = O(1/\varepsilon^2)$, whereas for degree 2 we get a bound of $k(\varepsilon) = 2^{1/\varepsilon^{O(1)}}$.

7.7.1 Fooling Halfspaces through Characteristic Functions

In this section, we give a Fourier theoretic proof that $O(1/\varepsilon^2)$-wise independence fools ε-regular halfspaces. This bound was first achieved by Diakonikolas, Kane and Nelson [28]. However, our approach is arguably simpler and uses a classical generalization of Esséen's inequality due to Faĭnleĭb [33].

For a real-valued random variable X, define the characteristic function $\varphi_X : \mathbb{R} \to \mathbb{R}$ as follows:

$$\varphi_X(t) = \mathbb{E}[e^{-itX}],$$

where $i = \sqrt{-1}$. Note that if X is symmetric $\varphi_X(t) \in \mathbb{R}$ for every $t \in \mathbb{R}$.

For a real-valued random variable Y, define the anti-concentration function $AC : \mathbb{R}_+ \to \mathbb{R}_+$ by $AC_Y(\alpha) = \sup_{t\in\mathbb{R}}\{\mathsf{Pr}[Y \in [t, t+\alpha]]\}$.

Lemma 7.7.1 (Theorem 1, Faĭnleĭb [33])**.** *There exists a universal constant $C_1 > 0$ such that for any real-valued random variables X, Y and $T > 0$,*

$$\mathsf{d}_{\mathsf{cdf}}(X, Y) < C_1 \left(AC_Y(1/T) + \int_0^T \frac{|\varphi_X(t) - \varphi_Y(t)|}{t} dt \right).$$

Theorem 7.7.2. *There exist universal constants C, C' such that the following holds. Let $\mathcal{D}$ be an m-wise independent distribution over $\{1, -1\}^n$ for $m = C/\varepsilon^2$ even. Then, for every ε-regular $w \in \mathbb{R}^n$ with $\|w\| = 1$ and $x \leftarrow \mathcal{D}$, $y \in_u \{1, -1\}^n$,*

$$\mathsf{d}_{\mathsf{cdf}}(\langle w, x\rangle, \langle w, y\rangle) < C'\varepsilon.$$

In other words, $O(1/\varepsilon^2)$-wise independence $O(\varepsilon)$-fools ε-regular halfspaces.

Proof. Let $X = \langle w, x\rangle$ for $x \leftarrow \mathcal{D}$ and let $Y = \langle w, y\rangle$ for $y \in_u \{1, -1\}^n$. To avoid the minor technicality of dealing with complex numbers, we assume that X is symmetric about 0. We can use the same argument for the general case, incurring only an additional factor of two in the error bound. For any symmetric real-valued random-variable Z with finite moments, by Taylor expansion, for m even,

$$\left| \varphi_Z(t) - \sum_{j=0}^{m-1} \frac{i^r t^r \, \mathbb{E}[Z^r]}{r!} \right| < \frac{|t|^m \, \mathbb{E}[Z^m]}{m!}.$$

Applying the above equation to X, Y and noting that by Khintchine's inequality, Lemma 2.2.3, $\mathbb{E}[X^r] = \mathbb{E}[Y^r] < r^{r/2}$ for $r \leq m$, we get

$$|\varphi_X(t) - \varphi_Y(t)| < \frac{2|t|^m m^{m/2}}{m!}.$$

By Stirling's approximation, $m! > (m/e)^m$. Therefore,

$$\int_0^T \frac{|\varphi_X(t) - \varphi_Y(t)|}{t} dt < \frac{2m^{m/2}}{m!} \int_0^T t^{m-1} dt$$
$$= \frac{2m^{m/2}}{m!} \cdot \frac{T^m}{m} < \frac{2}{m} \cdot \left(\frac{T \cdot e}{\sqrt{m}}\right)^m.$$

Further, by Lemma 7.4.4 and Theorem 2.2.6, $AC_Y(\alpha) = O(\alpha + \varepsilon)$. Thus, by Lemma 7.7.1, and the above equation applied to $T = 1/\varepsilon$, and $m = C/\varepsilon^2$, we get

$$\mathsf{d}_{\mathsf{cdf}}(X, Y) = O(\varepsilon).$$

The statement now follows. □

7.7.2 Relation to the Classical Moment Problem

We first observe that there is a strong connection between the question of fooling PTFs by bounded independence and the classical moment problem in probability. The classical moment problem (or more specifically, the Hamburger moment problem) can be phrased as follows: Given a sequence of numbers $M = (M_1, M_2, \ldots,) \in \mathbb{R}^{\mathbb{N}}$, when is there a unique distribution μ over $\mathbb{R}$ such that the moments of μ match the corresponding entries of M. That is, for every $i \in \mathbb{N}$,

$$M_i = \int_{-\infty}^{\infty} x^i d\mu(x).$$

We refer to the excellent book [2] for a detailed history and results on this problem. Here we only discuss the work of Klebanov and Mkrtchyan [58] who give quantitative bounds for the truncated moment problem where we

only know the first few moments as opposed to knowing all the moments as above.

For two real-valued random variables X, Y define the Lévy distance between them as follows:

$$\mathsf{d_L}(X,Y) = \inf\{\varepsilon > 0 : \mathsf{Pr}[X < t-\varepsilon] - \varepsilon < \mathsf{Pr}[Y < t] < \mathsf{Pr}[X < t+\varepsilon] + \varepsilon, \ \ \forall t \in \mathbb{R}\}.$$

The following result of Klebanov and Mkrtchyan gives quantitative bounds on the Lévy distance between two random variables whose first few moments are identical.

Theorem 7.7.3 (Theorem 1, [58])**.** *Let X, Y be real-valued random variables with $\mathbb{E}[X^i] = \mathbb{E}[Y^i]$ for $1 \leq i \leq 2m$. Then, there exists a constant C_σ that depends only on $\mathbb{E}[X^2] = \sigma^2$ such that*

$$\mathsf{d_L}(X,Y) \leq \frac{C_\sigma \cdot \ln(1+\beta_m)}{\beta_m^{1/4}},$$

where

$$\beta_m = \sum_{i=1}^{m} \frac{1}{(\mathbb{E}[X^{2i}])^{1/2i}}.$$

To go from Lévy distance to cdf distance we use the following simple lemma.

Lemma 7.7.4. *For any two real-valued random variables X, Y,*

$$\mathsf{d_{cdf}}(X,Y) < AC_Y(\mathsf{d_L}(X,Y)) + \mathsf{d_L}(X,Y).$$

Proof. From the definition of $\mathsf{d_L}$, for any $t \in \mathbb{R}$,

$$\mathsf{Pr}[Y < t] < \mathsf{Pr}[X < t + \mathsf{d_L}(X,Y)] + \mathsf{d_L}(X,Y) < \mathsf{Pr}[X < t] + AC_Y(\mathsf{d_L}(X,Y)) + \mathsf{d_L}(X,Y).$$

Similarly, we get that $\mathsf{Pr}[X < t] < \mathsf{Pr}[Y < t] + AC_Y(\mathsf{d_L}(X,Y)) + \mathsf{d_L}(X,Y)$. The lemma follows. □

We now use Theorem 7.7.3 to show that bounded independence fools degree 2 threshold functions. We remark that a similar argument when applied to halfspaces shows that $\tilde{O}(1/\varepsilon^8)$-wise independence fools halfspaces. However, we do not delve on this as we already saw how to get the optimal bound in the previous section.

Note that the bounds implied by Theorem 7.7.3 are worse than those of Diakonikolas et al. [26], Diakonikolas, Kane and Nelson [28]. However, we believe the argument is still interesting as the results of Klebanov and Mkrchtyan use completely different techniques and appeared before the above works.

Corollary 7.7.5. *There exist constants C, C' such that the following holds. Let $\mathcal{D}$ be a m-wise independent distribution over $\{1,-1\}^n$ for $m = 2^{C/\varepsilon^2}$. Then, for every ε-regular degree 2 polynomial $P : \mathbb{R}^n \to \mathbb{R}$, and $x \leftarrow \mathcal{D}$, $y \in_u \{1,-1\}^n$,*

$$\mathsf{d_{cdf}}(P(x), P(y)) < C'\varepsilon^{2/9}.$$

In other words, $(2^{O(1/\varepsilon^2)})$-wise independence $O(\varepsilon^{2/9})$-fools ε-regular degree two threshold functions.

Proof. It suffices to show the statement when $\mathcal{D}$ is $4m$-wise independent for $m = 2^{C/\varepsilon^2}$ for C to be chosen later. Without loss of generality suppose that $\|P\| = 2$. Let random variables $X = P(x)$, for $x \leftarrow \mathcal{D}$ and $Y = P(y)$, for $y \in_u \{1, -1\}^n$. Then, $\mathbb{E}[X^i] = \mathbb{E}[Y^i]$ for $i \leq 2m$ as x is $4m$-wise independent and P is a degree 2 polynomial. Now, for $i \leq m$, by hypercontractivity, Theorem 2.2.1, applied to $q = i$, $d = 2$,

$$\mathbb{E}[X^{2i}] = \mathbb{E}[Y^{2i}] < (2i)^{2i}.$$

Therefore,

$$\beta_m = \sum_{i=1}^{m} \frac{1}{\mathbb{E}[X^{2i}]^{1/2i}} > \sum_{i=1}^{m} \frac{1}{2i} = \Omega(\log m).$$

By Theorem 7.7.3,

$$\mathsf{d}_{\mathsf{L}}(X, Y) = O\left(\frac{\log\log m}{(\log m)^{1/4}}\right).$$

Now, by Theorem 2.2.5 and Theorem 2.2.6 applied to degree $d = 2$, $AC_Y(\alpha) = O(\varepsilon^{2/9} + \sqrt{\alpha})$. Therefore, by Lemma 7.7.4

$$\mathsf{d}_{\mathsf{cdf}}(X, Y) = O\left(\varepsilon^{2/9} + \frac{\sqrt{\log\log m}}{(\log m)^{1/8}}\right).$$

The statement now follows by choosing C to be sufficiently large. □

Remark 7.7.1. Note that the above approach does not work for degrees three and higher. This is because β_m does not grow to infinity as m becomes larger for higher degrees. For instance, in the degree three case, for X, Y as above, using Theorem 2.2.1,

$$\beta_m = \sum_{i=1}^{m} \frac{1}{\mathbb{E}[X^{2i}]^{1/2i}} \sim \sum_{i=1}^{m} \frac{1}{(2i)^{3/2}} = \Theta(1).$$

7.8 Non-Explicit Bounds

It is known ([63], [89]) that the number of distinct halfspaces on n bits is at most 2^{n^2}. One way of extending this bound to degree d PTFs is as follows. It is known that the Fourier coefficients of the first $d+1$ levels of a degree d PTF, also known as the Chow parameters, determine the PTF completely (see [83]). Thus, a PTF f is completely determined by $\mathsf{ChowParam}(f) = (\,\mathbb{E}[f \cdot \chi_I] : I \subseteq [n], |I| \leq d\,)$, where $\chi_I(x) = \prod_{i \in I} x_i$ denotes the parity over the coordinates in I. Observe that for any $I \subseteq [n]$, $\mathbb{E}[\,f \cdot \chi_I\,] \in \{i/2^n : i \in \mathbb{Z}, |i| \leq 2^n\}$. Therefore, the number of distinct degree d PTFs is at most the number of distinct sequences $\mathsf{ChowParam}(\,)$, which in turn is at most $(2^n)^{n^d}$.

The non-explicit bound now follows by observing that any class of Boolean functions $\mathcal{F}$ can be fooled with error at most ε by a set of size at most $O(\log(|\mathcal{F}|)/\varepsilon^2)$. Thus, degree d PTFs can be fooled by a sample space of size at most $O(n^{d+1}/\varepsilon^2)$.

Chapter 8

Pseudorandom Generators for Polytopes

In this chapter we construct pseduorandom generators for polytopes. Our constructions use the invariance principle for polytopes from Chapter 3 and the main generator construction from Section 7.3. Together with the results in Chapter 7, these results illustrate (by examples) a rough framework for combining invariance principles and the construction of Section 7.3 to obtain pseudorandom generators. This theme will be further emphasized in the following chapter.

8.1 Introduction

Recall the definition of pseudorandom generators (PRGs) for polytopes:

Definition 8.1.1. Let μ be a distribution over $\mathbb{R}$. A function $G : \{0,1\}^r \to \{1,-1\}^n$ is said to δ-fool a polytope $\mathcal{K}$ with respect to μ if the following holds.

$$\left| \Pr_{y \in_u \{0,1\}^r}[G(y) \in \mathcal{K}] - \Pr_{X \leftarrow \mu^n}[X \in \mathcal{K}] \right| \leq \delta.$$

Combining our invariance principle with the main generator of Section 7.3 appropriately, we obtain a black-box algorithm for approximately counting the number of $\{-1,1\}^n$ points in polytopes formed by the intersec-

tion of regular halfspaces. Recall the definitions of proper and hypercontractive distributions, Definition 3.2.2, Definition 3.2.3.

Theorem 8.1.1 (PRGs for regular polytopes and approximate counting)**.** *For all $\delta \in (0,1)$, there exists an explicit PRG $G : \{0,1\}^r \to \{1,-1\}^n$ with $r = O((\log n \log k)/\varepsilon)$ that δ-fools all polytopes formed by the intersection of k ε-regular halfspaces with respect to all proper and hypercontractive distributions μ for $\varepsilon = \delta^5/(\log^{8.1} k)(\log(1/\delta))$.*

The constants above depend on the hypercontractivity constants of μ. Note that the uniform distribution over $\{-1,1\}^n$ and the Gaussian distribution are examples of proper and hypercontractive distributions.

Theorem 8.1.1 implies quasi-polynomial time, deterministic, approximate counting algorithms for a broad class of integer programs. For example, dense covering programs such as dense set-cover, and $\{0,1\}$-contingency tables correspond to polytopes formed by the intersection of ε-regular halfspaces. For these types of integer programs, we can deterministically approximate, to within an additive error ε, the number of integer solutions in quasi-polynomial time.

As stated, our invariance principle applies to polytopes whose bounding hyperplanes have coefficients that are sufficiently regular. In some cases, however, we can randomly rotate an arbitrary polytope so that all the bounding hyperplanes become regular. As such, after applying a suitable random transformation (which we derandomize), we can build PRGs for *arbitrary* polytopes

if the underlying distribution is spherically symmetric (e.g., Gaussian):

Theorem 8.1.2 (PRGs for Polytopes in Gaussian space)**.** *For a universal constant $c > 0$ and all $\delta > c \log^2 k / n^{1/11}$, there exists an explicit PRG $G_{\mathcal{N}} : \{0,1\}^r \to \mathbb{R}^n$ with $r = O((\log n)(\log^{9.1} k)/\delta^{5.1})$ that δ-fools all k-polytopes with respect to $\mathcal{N}$.*

Additionally, we prove an invariance principle for polytopes with respect to the uniform distribution over the n-dimensional sphere S^{n-1}. This allows us to easily modify our PRG for polytopes in Gaussian space and build PRGs for intersections of spherical caps:

Theorem 8.1.3 (PRGs for intersections of spherical caps)**.** *For a universal constant $c > 0$ and all $\delta > c \log^2 k / n^{1/11}$, there exists an explicit PRG $G_{sp} : \{0,1\}^r \to S^{n-1}$ with $r = O((\log n)(\log^{9.1} k)/\delta^{5.1})$ that δ-fools all k-polytopes with respect to the uniform distribution over S^{n-1}.*

An immediate consequence of the above PRG construction is a polynomial time derandomization of the Goemans-Williamson approximation algorithm for Max-Cut [35] and other similar hyperplane based randomized rounding schemes. Observe that this derandomization is a *black-box* derandomization as opposed to some earlier derandomizations of the Goemans-Williamson algorithm, which are instance-specific (e.g., [70]).

8.1.1 Related Work

There is a long history of research on approximately counting the number of solutions to integer programs, especially with regard to contingency tables [48, 23]. However, not much is known in terms of *deterministic* algorithms, and we believe that our deterministic quasi-polynomial time algorithms for dense covering problems and dense set cover instances is the first result of its kind.

Regarding contingency tables, Dyer [31] gave a randomized relative-error approximation algorithm for counting solutions to contingency tables that runs in time exponential in the number of rows. In contrast, we obtain an algorithm that runs in quasi-polynomial time in the number of rows (however, we do not give a relative-error approximation). Although not stated explicitly before, it is easy to see that the pseudorandom generator for small space machines of Impagliazzo et al. [46] yields a deterministic algorithm for counting $n \times k$ contingency tables with additive error at most ε and run time $2^{O(\log^2(nk/\varepsilon))}$. This is incomparable to our algorithm for contingency tables which has run time $2^{(\log n)\cdot \mathrm{poly}(\log k, 1/\varepsilon)}$. In our case, we obtain a polynomial-time, black-box derandomization for contingency tables with a constant number of rows (for $\varepsilon = O(1)$).

For PRGs for intersections of halfspaces, recently Gopalan et al. [37] and Diakonikolas et al. [28] gave results incomparable to ours. Gopalan et al. give generators for arbitrary intersections of k halfspaces with seed length linear in k but logarithmic in $1/\delta$. Diakonikolas et al. show that bounded inde-

pendence fools intersections of quadratic threshold functions and in particular, get generators with seed length $O((\log n) \cdot \mathrm{poly}(k, 1/\varepsilon))$ fooling intersections of k halfspaces. Due to the at least linear dependence on k, the results of the above works do not yield good algorithms for counting solutions to integer programs, as in this setting k is typically large (e.g., $\mathrm{poly}(n)$).

8.2 Pseudorandom Generators for Polytopes

We now prove our main theorems for constructing pseudorandom generators for polytopes with respect to a variety of distributions (Theorems 8.1.1, 8.1.2, and 8.1.3).

The results in this section are based on the main generator from Section 7.3. We remark that although the PRG construction is essentially the same, the analysis is not immediate (even given our invariance principle) and requires a careful application of hypercontractivity.

We next describe the parameter settings for the generator from Section 7.3 that we use. We redefine the generator here again to be consistent with the notations of this chapter and that of our invariance principle for polytopes (which used k as the number of halfspaces).

Given $\delta \in (0, 1)$, let $\varepsilon = \Omega(\delta^6 / \log^{9.6} k)$ be such that $\log^{1.6} k (\varepsilon \log(1/\varepsilon))^{1/5} = \delta$. Let $t = 1/\varepsilon$ and let $\mathcal{H} = \{h : h : [n] \to [t]\}$ be a $(2 \log k)$-wise independent family of hash functions. Efficient constructions of hash families $\mathcal{H}$ as above with $|\mathcal{H}| = O(n^{2 \log k})$ are known. To avoid some technical issues that can be

overcome easily, we assume that every hash function $h \in \mathcal{H}$ is equi-distributed in the following sense: for all $j \in [t]$, $|\{i : h(i) = j\}| = n/t$.

Let $m = n/t$ and let $G_0 : \{0,1\}^s \to \{1,-1\}^m$ generate a $(4 \log k)$-wise independent distribution over $\{1,-1\}^m$. Efficient constructions of generators G_0 as above with $s = O(\log k \log n)$ are known [78].

Given a hash family and generator G_0 as above, we consider the following generator. Define $G : \mathcal{H} \times (\{0,1\}^s)^t \to \{1,-1\}^n$ by

$$G(h, z^1, \ldots, z^t) = x, \text{ where } x_{|h^{-1}(i)} = G_0(z^i) \text{ for } i \in [t].$$

8.2.1 Pseudorandom Generators for Regular Polytopes

We now argue that the generator G defined above fools regular polytopes and prove Theorem 8.1.1.

Proof of Theorem 8.1.1. The bound on the seed length of the generator G follows from the construction. The following statement follows from an argument similar to that of the proof of Theorem 3.3.2: for any smooth function $\psi : \mathbb{R}^k \to \mathbb{R}$ and ε-regular W,

$$\left| \mathop{\mathbb{E}}_{y \in_u \{0,1\}^r} [\psi(W^T G(y))] - \mathop{\mathbb{E}}_{Y \leftarrow \mathcal{N}^n} [\psi(W^T Y)] \right| \leq C \log^3 k \, (\varepsilon \log(1/\varepsilon)) \, \|\psi^{(4)}\|_1. \tag{8.2.1}$$

Indeed, to observe that Lemma 3.4.1 holds for any $(2 \log k)$-wise independent family of hash functions and the proof of Lemma 3.4.2 relies only on two key properties of $X \leftarrow \mu^n$: (1) For a fixed hash function h, the blocks $X_{h^{-1}(1)}, X_{h^{-1}(2)}, \ldots, X_{h^{-1}(t)}$ are independent of one another. (2) For a fixed

hash function h, and $j \in [t]$, the distribution of each block $X_{h^{-1}(j)}$ satisfies $(2, 2\log k)$-hypercontractivity for all $j \in [t]$. In other words, we used the property that for all $j \in [t]$, $u \in \mathbb{R}^{|h^{-1}(j)|}$,

$$\mathbb{E}[|\langle u, X_{h^{-1}(j)}\rangle|^{4\log k}] \leq (C\log k)^{2\log k}\,\|u\|^{4\log k}. \tag{8.2.2}$$

Note that X generated according to the generator G satisfies both the above conditions: 1) For a fixed function h, the blocks are independent by definition and 2) the hypercontractivity inequality 8.2.2 only involves the first $(4\log k)$-moments of the distribution of $X_{h^{-1}(j)}$. As a consequence, inequality 8.2.2 holds for any $(4\log k)$-wise independent distribution over $\{1,-1\}^{|h^{-1}(j)|}$.

We can now move from closeness in expectation to closeness in cdf distance by an argument similar to the proof of Theorem 3.3.1, where we use Equation 8.2.1 instead of Theorem 3.3.2, to get

$$|\operatorname*{Pr}_{y\in_u\{0,1\}^r}[G(y)\in\mathcal{K}] - \operatorname*{Pr}_{Y\leftarrow\mathcal{N}^n}[Y\in\mathcal{K}]| \leq \delta.$$

The theorem now follows from the above equation and Theorem 3.3.1. □

8.2.1.1 Approximate Counting for Integer Programs

The PRG from Theorem 8.1.1 coupled with enumeration over all possible seeds immediately implies a quasi-polynomial time, deterministic algorithm for approximately counting, within a small additive error, the number of solutions to "regular" $\{0,1\}$-integer programs. It turns out that "regular" integer programs correspond to a broad class of well-studied combinatorial

problems. For example, we obtain deterministic, approximate counting algorithms for *dense* set cover problems and $\{0,1\}$-contingency tables. We obtain quasi-polynomial time algorithms even when there are a polynomial number of constraints (or polynomial number of rows in the contingency table setting). As far as we know, there is no prior work giving nontrivial *deterministic* algorithms for counting solutions to integer programs with many constraints.

Here we discuss the case of *dense* set cover instances and remark that we get similar results for the special case of counting contingency tables. Covering integer programs are a fundamental class of integer programs and can be formulated as follows.

$$\begin{aligned} &\min \sum_i X_i \\ \text{s.t. } &\sum_i a_{ij} X_i \geq c_j,\ j = 1, \ldots, k, \\ &X \in \{0,1\}^n, \end{aligned} \tag{8.2.3}$$

where the coefficients of the constraints a_{ij} and c_j are all non-negative. An important special class of covering integer programs is set cover, which in turn is a generalization of many important problems in combinatorial optimization such as edge cover and multidimensional $\{0,1\}$-knapsack.

In the standard set cover problem, the input is a family of sets $S_1, \ldots, S_n$ over a universe U of size k and an integer t. The goal is to find a subfamily of sets $\mathcal{C}$ such that $|\mathcal{C}| \leq t$ and the union of all the sets in $\mathcal{C}$ equals U. This corresponds to a covering program as above with k constraints and n unknowns from $\{0,1\}$. Call an instance of set cover ε-dense if each element in

U appears in at least $1/\varepsilon^2$ of the different sets S_i. It is easy to verify that with this restriction, after translating from $\{0,1\}$ to $\{1,-1\}$ and appropriate normalization, all the linear constraints in the corresponding integer program as in Equation 8.2.3 are ε-regular. Thus, using the generator from Theorem 8.1.1 and enumerating over all seeds to the generator, we have the following:

Theorem 8.2.1. *There exists a deterministic algorithm that, given instance of an ε-dense set covering problem with k constraints over a universe of size n, approximates the number of solutions to within an additive factor of δ in time $n^{\mathrm{poly}(\log k, 1/\delta)}$ as long as $\varepsilon \leq \delta^5/(\log^{8.1} k)(\log(1/\delta))$.*

We now briefly elaborate on approximately counting the number of $\{0,1\}$ contingency tables. The problem of counting $\{0,1\}$-contingency tables is the following. Given, positive integers n, k $n > k$, $\mathbf{r} = (r_1, \ldots, r_n) \in \mathbb{Z}^n$, $\mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{Z}^k$ we wish to count the number of solutions, $\mathsf{CT}(\mathbf{r}, \mathbf{c})$, to the following integer program whose solutions are matrices $X \in \{0,1\}^{n \times k}$ with row and column sums given by $\mathbf{r}, \mathbf{c}$.

$$\begin{aligned} &\text{Find } X \in \{0,1\}^{n \times k} \\ \text{s.t. } &\sum_j X_{ij} = r_i, 1 \leq i \leq n, \\ &\sum_i X_{ij} = c_j, 1 \leq j \leq k. \end{aligned}$$

Observe that, after translating from $\{0,1\}$ to $\{1,-1\}$ and appropriately normalizing, solutions to the above integer program correspond to points from $\{1,-1\}^{n \times k}$ that lie in an intersection of $2(n+k)$-halfspaces each of which is

$(1/\sqrt{k})$-regular (recall that the notion of regularity does not depend on the value of the c_i's or r_j's). Thus, as with dense instances of set cover, we can use Theorem 8.1.1 to count the number of $\{0,1\}$-contingency tables:

Theorem 8.2.2. *There exists a deterministic algorithm that on input $\mathbf{r} \in \mathbb{Z}^n$, $\mathbf{c} \in \mathbb{Z}^k$, approximates $\mathsf{CT}(\mathbf{r},\mathbf{c})/2^{nk}$, the fraction of $\{0,1\}$-contingency tables with sums $\mathbf{r},\mathbf{c}$, to within additive error δ, and runs in time $n^{\mathrm{poly}(\log k, 1/\delta)}$.*

We remark that using results of Wolff [99], who shows hypercontractivity for various discrete distributions, we can approximately count number of solutions to dense set cover instances and contingency tables over most natural domains.

8.2.2 Pseudorandom Generators for Polytopes in Gaussian Space

We now prove Theorem 8.1.2. We use an idea of Ailon and Chazelle [1] and the invariance of the Gaussian measure to unitary rotations to obtain PRGs with respect to $\mathcal{N}^n$ for *all* polytopes. The basic idea is similar to that of the PRG for spherical caps from Section 7.6. In the current setting, we must prove that, with respect to a random rotation, *all* of the bounding hyperplanes become regular with high probability. Such a tail bound requires applying hypercontractivity.

Let $H \in \mathbb{R}^{n\times n}$ be the normalized Hadamard matrix with $HH^T = I_n$ and $H_{ij} \in \{1/\sqrt{n}, -1/\sqrt{n}\}$. Ailon and Chazelle show that for any $w \in \mathbb{R}^n$, and a random diagonal matrix D with uniformly random $\{1,-1\}$ entries, the vector

HDw is regular with high probability. We derandomize their observation using hypercontractivity. For a vector $x \in \mathbb{R}^n$, let $D(x) \in \mathbb{R}^{n \times n}$ be the diagonal matrix with diagonal entries x.

Lemma 8.2.3. *There exists a constant $C > 0$ such that the following holds. For any $w \in \mathbb{R}^n$, $\|w\| = 1$, $0 < \delta < 1$ and any $(C \log(k/\delta))$-wise independent distribution $\mathcal{D}$ over $\{1, -1\}^n$,*

$$\Pr_{x \leftarrow \mathcal{D}} \left[\, \|HD(x)w\|_4^4 \geq C \log^2(k/\delta)/n \, \right] \leq \delta/k.$$

Proof. Fix a $w \in \mathbb{R}^n$ and a $C \log(k/\delta)$-wise independent distribution $\mathcal{D}$ for constant C to be chosen later. Let random variable $Z = \|HD(x)w\|_4^4 = \sum_i \left(\sum_l H_{il} x_l w_l\right)^4$ for $x \leftarrow \mathcal{D}$. Note that x satisfies $(2, q)$-hypercontractivity for $q \leq C \log(k/\delta)$. Now,

$$\begin{aligned}
\mathbb{E}[Z^2] &= \sum_{i,j} \mathbb{E}\left[\left(\sum_l H_{il} x_l w_l\right)^4 \left(\sum_{l'} H_{jl'} x_{l'} w_{l'}\right)^4\right] \\
&\leq \sum_{ij} \sqrt{\mathbb{E}\left[\left(\sum_l H_{il} x_l w_l\right)^8\right] \cdot \mathbb{E}\left[\left(\sum_l H_{jl} x_l w_l\right)^8\right]} \\
&\qquad \text{Cauchy-Schwarz inequality} \\
&\leq \sum_{i,j} 8^4 \left(\mathbb{E}\left[\left(\sum_l H_{il} x_l w_l\right)^2\right]\right)^2 \left(\mathbb{E}\left[\left(\sum_l H_{jl} x_l w_l\right)^2\right]\right)^2 \\
&\qquad (2,8)\text{-hypercontractivity} \\
&= 8^4 \sum_{i,j} \frac{1}{n^4} = \frac{c}{n^2}.
\end{aligned}$$

Observe that Z is a degree 4 multilinear polynomial over $x_1, \ldots, x_n$. Therefore, by $(2, q)$-hypercontractivity, Theorem 2.2.1, applied to the random variable Z, for $q \leq C \log(k/\delta)/4$,

$$\mathbb{E}[|Z|^q] \leq q^{2q} (\mathbb{E}[Z^2])^{q/2} \leq \frac{c^{q/2}\, q^{2q}}{n^q}.$$

Hence, by Markov's inequality, for $\gamma > 0$,

$$\mathsf{Pr}[\, |Z| > \gamma \,] = \mathsf{Pr}[\, |Z|^q > \gamma^q \,] \leq \left(\frac{c^{1/2}\, q^2}{\gamma n} \right)^q.$$

The lemma now follows by taking $q = 2 \log(k/\delta)$ and $\gamma = 2\, c^{1/2}\, q^2/n$. □

Let $G : \{0,1\}^r \to \{1,-1\}^n$ be the generator from Theorem 8.1.1 for $r = O((\log n \log k)/\varepsilon)$. Let $G_1 : \{0,1\}^{r_1} \to \{1,-1\}^n$ generate a $C \log(k/\delta)$-wise independent distribution, for constant C as in Lemma 8.2.3. Generators G_1 as above with $r_1 = O(\log(k/\delta) \log n)$ are known. Define $G_{\mathcal{N}} : \{0,1\}^{r_1} \times \{0,1\}^r \to \mathbb{R}^n$ as follows:

$$G_{\mathcal{N}}(x, y) = D(G_1(x)) H G(y).$$

We claim that $G_{\mathcal{N}}$ δ-fools all polytopes with respect to $\mathcal{N}^n$.

Proof of Theorem 8.1.2. Recall that $\varepsilon = \Omega(\delta^{5.1}/\log^{8.1} k) > 1/n^{.51}$. The seed length of $G_{\mathcal{N}}$ is $r_1 + r = O(\log n \log k/\varepsilon)$. Fix $W \in \mathbb{R}^{n \times n}$. Observe that $W^T G_{\mathcal{N}}(x, y) = (HD(G_1(x))W)^T G(y)$. Now, from Lemma 8.2.3 and a union bound it follows that

$$\mathop{\mathsf{Pr}}_{x \in_u \{0,1\}^{r_1}} [\, HD(G_1(x))W \text{ is not } \varepsilon\text{-regular} \,] \leq \delta. \tag{8.2.4}$$

Further, from the invariance of $\mathcal{N}^n$ with respect to unitary rotations, for any $x \in \{0,1\}^{r_1}$,

$$\Pr_{z \leftarrow \mathcal{N}^n}[(HD(G_1(x))W)^T z \in \mathsf{Rect}(\theta)] = \Pr_{z \leftarrow \mathcal{N}^n}[W^T z \in \mathsf{Rect}(\theta)].$$

Thus, from Theorem 8.1.1 applied to $\mathcal{N}$, we get that for $HD(G_1(x))W$ ε-regular,

$$|\Pr_{y \in_u \{0,1\}^r}[\,(HD(G_1(x))W)^T G(y) \in \mathsf{Rect}(\theta)\,] - \Pr_{z \leftarrow \mathcal{N}^n}[W^T z \in \mathsf{Rect}(\theta)]| \leq \delta. \tag{8.2.5}$$

The theorem now follows from Equations (8.2.4), (8.2.5). □

8.2.3 Pseudorandom Generators for Intersections of Spherical Caps

Theorem 8.1.3 follows from Theorem 8.1.2 and the following new invariance principle for polytopes over S^{n-1}:

Lemma 8.2.4. *For any polytope $\mathcal{K}$ with k faces,*

$$\left|\Pr_{X \in_u S^{n-1}}[X \in \mathcal{K}] - \Pr_{Y \leftarrow \mathcal{N}^n}\left[Y/\sqrt{n} \in \mathcal{K}\right]\right| \leq \frac{C \log n \, \log k}{\sqrt{n}}.$$

The proof uses Nazarov's bound on Gaussian surface area and the following classical large deviation bound for the norm of a random Gaussian vector (for a nice exposition of the bound see [96])

Lemma 8.2.5. *For $Y \leftarrow \mathcal{N}^n$,*

$$\mathsf{Pr}[\,|\|Y\| - \sqrt{n}| > t\,] \leq a \exp(-b\,t^2),$$

where $a, b > 0$ are universal constants.

Proof of Lemma 8.2.4. Fix a polytope $\mathcal{K}(W, \theta)$. Let $X \in_u S^{n-1}$ and $Y \leftarrow \mathcal{N}^n$. Note that $Y/\|Y\|$ is uniformly distributed over S^{n-1}. Fix $\delta = c/n^{1/2}$ for a constant c to be chosen later. Observe that for $Y \leftarrow \mathcal{N}^n$, and $u \in \mathbb{R}^n$, $\|u\| = 1$, $\langle u, Y \rangle$ is distributed as $\mathcal{N}$. Hence, for any $u \in \mathbb{R}^n$, $\|u\| = 1$,

$$\mathsf{Pr}[\, |\langle u, Y \rangle| \geq \sqrt{\log(k/\delta)}\,] \leq \frac{\delta}{k}.$$

Therefore, by a union bound,

$$\mathsf{Pr}[\, \|W^T Y\|_\infty / \sqrt{n} > \sqrt{\log(k/\delta)}/\sqrt{n}\,] \leq \delta.$$

Further, by using Lemma 8.2.5 and the fact that $Y/\|Y\|$ is uniformly distributed over S^{n-1},

$$\mathsf{Pr}[\, \|W^T X\|_\infty > \sqrt{C \log(k/\delta)}/\sqrt{n}\,] \leq 2\delta,$$

for a sufficiently large constant C. From the above two equations, it follows that to prove the theorem we can assume that

$$\|\theta\|_\infty < \sqrt{C \log(k/\delta)/n}.$$

Now, from Lemma 8.2.5 and the above equation it follows that

$$\mathsf{Pr}[\, |\|Y\| - \sqrt{n}|\, \|\theta\|_\infty \geq \sqrt{C \log(1/\delta) \log(k/\delta)/n}\,] \leq \delta. \tag{8.2.6}$$

Let $\lambda = \sqrt{C \log(1/\delta) \log(k/\delta)/n}$. Then, since $Y/\|Y\| \in_u S^{n-1}$

$$
\begin{aligned}
&| \Pr[X \in \mathcal{K}] - \Pr[Y/\sqrt{n} \in \mathcal{K}] | \\
&= | \Pr[W^T X \in \mathsf{Rect}(\theta)] - \Pr[W^T Y/\sqrt{n} \in \mathsf{Rect}(\theta)] | \\
&= | \Pr[W^T Y \in \|Y\| \, \mathsf{Rect}(\theta)] - \Pr[W^T Y \in \sqrt{n} \, \mathsf{Rect}(\theta)] | \\
&\leq \Pr[\, | \|Y\| - \sqrt{n} | \, \|\theta\|_\infty \geq \lambda \,] \\
&+ \Pr[\, W^T Y \in \mathsf{Rect}(\sqrt{n}\theta + \lambda 1_k) \setminus \mathsf{Rect}(\sqrt{n}\theta - \lambda 1_k) \,] \\
&\leq \delta + O(\, \lambda \sqrt{\log k} \,). \qquad \text{(Equation 8.2.6, Lemma 3.3.4)}
\end{aligned}
$$

The lemma now follows by choosing $\delta = c/n^{1/2}$ for a sufficiently large constant c. □

Proof of Theorem 8.1.3. Define $G_{sp} : \{0,1\}^{r_1} \times \{0,1\}^r \to S^{n-1}$ by $G_{sp}(x,y) = G_{\mathcal{N}}(x,y)/\sqrt{n}$. It follows from Theorem 8.1.2 and Lemma 8.2.4 that G_{sp} fools polytopes over S^{n-1} as in the theorem. □

Chapter 9

Pseudorandom Generators for Combinatorial Shapes

In this chapter we introduce the class of combinatorial shapes and construct PRGs for the class. In the spirit of the previous two chapters our PRG constructions can be viewed as (implicitly) using the discrete central limit theorems from Chapter 4 in conjunction with a more sophisticated form of the PRG construction from Section 7.3.

9.1 Introduction

The starting point of our results in this chapter are the PRGs for space-bounded computations of Nisan [80] and Impagliazzo, Nisan and Wigderson [46]. are PRGs for space-bounded computations. These PRGs use a seed of length $O(\log^2 n)$ to fool polynomial-width branching programs and have played a central role in studying the relative strength of randomness vs. memory. In particular, reducing their seed length to $O(\log n)$-bit would show that RL=L, namely every randomized algorithm can be derandomized with only a multiplicative constant blow-up in its memory. Improving [80, 46] is a central open question, not only for the possibility of proving RL=L, but also for other

important applications [47, 95, 54, 40]. Despite much effort, the above seed lengths have not been improved in nearly two decades.

While PRGs with logarithmic-seed that fool polynomial-width branching programs are still not known, logarithmic-seed PRGs for weaker classes of distinguishers have been previously constructed and found many applications. In this paper we define a natural common generalization and significant extension of many of these distinguisher classes, which we name *combinatorial shapes*. Combinatorial shapes look at their inputs in consecutive chunks of $\log m$ bits (usually m would be at most polynomial in n). On each chunk of bits the combinatorial shape may apply an arbitrary boolean function. Nevertheless, these Boolean functions are combined into a single output by a symmetric (i.e., order independent) function. Combinatorial shapes generalize combinatorial rectangles, halfspaces with 0/1 coefficients, and modular sums. Our main result is a construction of PRGs with seed length $O(\log n)$ that fools combinatorial shapes.

Definition 9.1.1. A function $f : [m]^n \to \{0,1\}$ is an (m,n)-combinatorial shape if there exist sets $A_1, \ldots, A_n \subseteq [m]$ and a symmetric function $h : \{0,1\}^n \to \{0,1\}$ such that $f(x_1, \ldots, x_n) = h(1_{A_1}(x_1), \ldots, 1_{A_n}(x_n))$. We denote the class of such functions by $\mathsf{CShape}(m,n)$.

We call them *combinatorial shapes* because they generalize combinatorial rectangles, which are simply the subset of $\mathsf{CShape}(m,n)$ where the symmetric function h is the AND function. PRGs for combinatorial rectangles have

received considerable attention [32, 6, 68], and have applications to numerical integration.

The class $\mathsf{CShape}(2,n)$ is interesting in its own right, as it comprises all Boolean functions $f : \{0,1\}^n \to \{0,1\}$ that are symmetric functions of a subset $S \subseteq [n]$ of variables. In order to fool $\mathsf{CShape}(2,n)$, the distribution of $\sum_{i\in S} x_i$ needs to be ε-close to $\mathsf{BIN}(|S|, \frac{1}{2})$ in *statistical* distance for every $S \subseteq [n]$. [1] Prior to our work, the best known generator for this problem was Nisan's generator [80] which gives seed-length $O(\log^2 n)$. Similarly, PRGs for $\mathsf{CShape}(m,n)$ imply generators that can fool such tests under multinomial distributions, by choosing the set A_i so that $1_{A_i}(x_i) = 1$ with probability p_i.

Parities of subsets are a special case of $\mathsf{CShape}(2,n)$; hence PRGs that fool $\mathsf{CShape}(2,n)$ are a strengthening of the ever so versatile ε-biased generators [77]. Recently, a different strengthening of ε-biased generators was considered, where bit-generators were given that fool sums modulo larger primes or even composites [67, 72]. The seed-length of these constructions is super-logarithmic unless the moduli is constant. It is easy to argue that a generator that fools $\mathsf{CShape}(2,n)$ also fools sums modulo an arbitrary moduli, or even non-modular sums.[2]

Note that in the above examples of combinatorial shapes, the symmetric function h could be computed by a constant width branching program. In this

[1] For $n > 0$, $p \in [0,1]$, $\mathsf{BIN}(n,p)$ denotes the binomial distribution of order n and bias p.

[2] Note that [67, 72] gives generators that fool sums with arbitrary coefficients. Generators that fool $\mathsf{CShape}(2,n)$ also fool modular (and non-modular) sums with 0/1 coefficients.

sense, combinatorial shapes seem significantly more powerful. Halfspaces with 0/1 coefficients are also special cases of $\mathsf{CShape}(2, n)$, where the symmetric function cannot be evaluated by a constant width branching program. PRGs which fool halfspaces were first given in [26, 74] and we saw constructions with better seed-length in Chapter 7. Note however that these results only guarantee that $\sum_{i \in S} x_i$ is close to $\mathsf{BIN}(|S|, \frac{1}{2})$ in Kolmogorov distance, whereas our goal is to get closeness in statistical distance.

9.1.1 Main Results

Our main result is a PRG construction which fools $\mathsf{CShape}(m, n)$.

Theorem 9.1.1 (Main)**.** *For every $\varepsilon > 0$, there exists a PRG that ε-fools $\mathsf{CShape}(m, n)$ with seed-length $O(\log m + \log n + \log^2(1/\varepsilon))$.*

When m is polynomial in n, these PRGs have seed length $O(\log n + \log^2(1/\varepsilon))$. Previously, the best known PRGs had seed length $O(\log^2 n)$, even for $m = 2$; these were the PRGs for space-bounded computation by Nisan and Impagliazzo, Nisan and Wigderson.

Along the way we also give a new PRG for combinatorial rectangles with seed-length $O(\log^{3/2} n)$ and error $1/\mathrm{poly}(n)$. This matches the parameters of the previous best generator due to Lu [68] for polynomially small ε.

Theorem 9.1.2. *For every $\varepsilon > 0$, there exists a generator that ε-fools (m, n)-combinatorial rectangles with seed-length $O(\log n \sqrt{\log(1/\varepsilon)})$.*

9.1.2 Outline of Constructions

Our constructions are based on the simple convolution lemma, Lemma 4.3.1, that we used in showing discrete central limit theorems in Chapter 4. (In fact, the PRG construction was the original motivation for the lemma). We restate the lemma below for convenience. Recall that a random variable Y is α-shift invariant if $\mathsf{d_{TV}}(Y, Y+1) \leq \alpha$.

Lemma 9.1.3 (Main Convolution Lemma, cf. Lemma 4.3.1)**.** *Let X be a α-shift invariant distribution and let Y, Z be integer-valued distributions with support contained in $[a, a+b]$ for some $a \in \mathbb{R}, b > 0 \in \mathbb{R}$. Then,*

$$\mathsf{d_{TV}}(X+Y, X+Z) \leq 4\sqrt{\alpha b \mathsf{d_{cdf}}(Y, Z)}.$$

For intuition, it is easier to work with the equivalent goal of fooling *combinatorial sums* in statistical distance.

Definition 9.1.2. A function $f : [m]^n \to [n]$ is an (m, n)-combinatorial sum if there exist sets $A_1, \ldots, A_n \subseteq [m]$ such that $f(x_1, \ldots, x_n) \equiv 1_{A_1}(x_1) + 1_{A_2}(x_2) + \cdots + 1_{A_n}(x_n)$. We denote this class of functions by $\mathsf{CSum}(m, n)$.

It is straightforward to verify that fooling combinatorial shapes is equivalent to fooling combinatorial sums in the stronger, statistical distance.

The basic building block for our constructions is the generator $G_{\mathcal{H},k,t}$ from Section 7.3 that we used to fool polynomial threshold functions. Our high level approach to fooling combinatorial sums is as follows:

1. We first show that $G_{\mathcal{H},k,t}$ fools combinatorial sums with small variance in statistical distance. We show that since the combinatorial sum restricted to each bucket has very small variance, bounded independence fools the sum restricted to a bucket in statistical distance. We then take a union bound across the different buckets. A weak bound for fooling the sum in each bucket is easy; however to apply the union bound requires a much stronger bound, which we prove using the "sandwiching polynomials" technique introduced by Bazzi [10].

2. We then show that $G_{\mathcal{H},k,t}$ fools combinatorial sums with high variance in Kolmogorov distance. We use the pairwise independence of $\mathcal{H}$ to argue that the total variance is well spread among the t buckets and then apply the Berry-Esséen theorem to show that the distribution is close to the right distribution in Kolmogorov distance. The analysis for this case is similar to the argument in Section 7.4.1.

3. We construct a generator $H_{m,n}$ fooling n dimensional combinatorial sums in statistical distance by recursively combining a generator fooling $n/2$ dimensional sums in Kolmogorov distance with a generator fooling $n/2$ dimensional sums in statistical distance. Unfolding this recursion, the generator $H_{m,n}$ hashes variables into $\log n$ buckets of geometrically increasing sizes and applies the generator $G_{\mathcal{H},k,t}$ to each bucket. We analyze this generator by exploiting the recursive construction to apply Lemma 4.3.1 at every step. We view this recursive construction and

analysis of the $H_{m,n}$ as the most novel part of our PRG construction. The analysis, while similar in spirit to our proof of the discrete central limit theorem Theorem 4.1.1 is more involved.

4. Finally, we show that one can generate the seeds for each bucket using the PRGs for small-space sources of [46], [81] rather than independently. This is done by constructing small-width sandwiching branching programs for combinatorial sums.

We obtain our result on fooling combinatorial rectangles by setting the parameters of $G_{\mathcal{H},k,t}$ appropriately and then derandomizing the construction using [80, 46] as above. The analysis however is different and uses a simple application of the principle of inclusion-exclusion and few properties of k-wise independent hash functions.

9.1.3 Related Work

Independently and simultaneously, Watson [98] studied the special case of combinatorial shapes where the symmetric function h is the parity function which are called *combinatorial checkerboards* by Watson. Watson obtains a seed-length of $O(\log m + \log n \log\log n + \log^{3/2}(1/\varepsilon))$ which is better than the seed-length we get for small ε.

As indicated earlier, PRGs for several special cases of combinatorial shapes have been studied previously. There was a lot of classical work on

low-discrepancy sets for axis-parallel rectangles in low dimension; see for example [71]. Even, Goldreich, Luby, Nisan, and Velickovic [32] were the first to give good constructions in high dimension; they gave PRGs for combinatorial rectangles which used an $O(\log^2 n)$ bit seed to achieve error $1/\mathrm{poly}(n)$ when $m = \mathrm{poly}(n)$. Armoni, Saks, Wigderson, and Zhou [6] improved the parameters to achieve a seed of length $O(\log m + \log n + \log^2(1/\varepsilon))$. The best construction is by Lu [68], who achieved a seed length of $O(\log m + \log n + \log^{3/2}(1/\varepsilon))$.

Diakonikolas, Gopalan, Jaiswal, Servedio, and Viola [26] showed that $O(\log^2(1/\varepsilon)/\varepsilon^2)$-wise independence ε-fools halfspaces, which gives a seed of length $O((\log n)\log^2(1/\varepsilon)/\varepsilon^2)$. Our constructions from Chapter 7 gave a different PRG with seed-length $O(\log n + \log^2(1/\varepsilon))$.

The notion of ε-biased spaces was introduced by Naor and Naor [77], who gave a PRG using $O(\log n + \log(1/\varepsilon))$ bits. Alon, Goldreich, Hastad, and Peralta [3] gave alternate constructions matching this bound. Lovett, Reingold, Trevisan, and Vadhan [67] gave a PRG over bits that fools sums modulo m, requiring a seed of length $O(\log n + \log(m/\varepsilon)\log(m\log(1/\varepsilon)))$. A similar, somewhat weaker construction was found independently by Meka and Zuckerman [72].

9.2 PRGs for Combinatorial Shapes

We ask the reader to recall the definition of $G_{\mathcal{H},k,t}$ from Section 7.3. In this chapter we work with the following extension of the generator so as to output values in $[m]^n$ instead of $\{1,-1\}^n$. Fix $k, t > 0$ and let $d = n/t$. Let

$\mathcal{H} = \{h : [n] \to [t]\}$ be a pairwise independent family of hash functions. Let $G_k : \{0,1\}^{r_k} \to [m]^d$ generate a k-wise independent space over $[m]^d$. Define $G_{\mathcal{H},k,t} : \mathcal{H} \times (\{0,1\}^{r_k})^t \to [m]^n$ as follows:

$$G_{\mathcal{H},k,t}(h, z^1, \ldots, z^t) = x, \text{ where } x_{h^{-1}(i)} = G_k(z^i) \text{ for } i = 1, \ldots, t. \tag{9.2.1}$$

As sketched in the introduction we work with fooling combinatorial sums in statistical distance and first study the case of combinatorial sums with small variance.

Definition 9.2.1. A generator $G : \{0,1\}^r \to [m]^n$ ε-fools $\mathsf{CSum}(m,n)$ in statistical distance if for any $f \in \mathsf{CSum}(m,n)$, the random variables $X = f(G(x)), x \in_u \{0,1\}^r$ and $Y = f(y), y \in_u [m]^n$ satisfy $\mathsf{d_{TV}}(X,Y) \leq \varepsilon$. Similarly, we say that G ε-fools $\mathsf{CSum}(m,n)$ in Kolmogorov (cdf) distance if X and Y satisfy $\mathsf{d_{cdf}}(X,Y) \leq \varepsilon$.

We first set up some notation to be used henceforth. Let $f : [m]^n \to [n]$ be an (m,n)-combinatorial sum with $f(x) = \sum_{i=1}^n 1_{A_i}(x_i)$ for $A_i \subseteq [m]$. For $x_i \in_u [m]$, define the indicator variable $X_i = 1_{A_i}(x_i)$. Let

$$p_i = \mathbb{E}[X_i],\ \sigma_i^2 = \mathrm{Var}[X_i] = p_i(1-p_i),\ \mu = \sum_{i=1}^n p_i,\ \sigma^2 = \sum_{i=1}^n \sigma_i^2$$

Let $X = \sum_{i=1}^n X_i$, so $\mathbb{E}[X] = \mu$ and $\sigma^2(X) = \sigma^2$ provided the X_i's are pairwise independent.

9.2.1 Fooling Small Combinatorial Sums

We now study the case of combinatorial sums with small variance. The strategy is as follows: since $\mathsf{Var}[f]$ is small, there is a small set $L \subseteq [n]$ of *large*

variance variables, such that all other indicator random variables $X_i = 1_{A_i}(x_i)$, $i \notin L$, have small variance. To handle variables in L, we argue that they will each be hashed into a different bucket. Thus the distribution on these variables is truly uniform, and moreover, conditioned on their values, the distribution of the output of the generator in each bucket is $(k-1)$-wise independent. We then use the fact that the combinatorial sum restricted to each bucket has very small total variance and show that bounded independence fools the sum restricted to a bucket in statistical distance. Finally we take a union bound across the different buckets to show the desired claim. As mentioned in the introduction, we use the "sandwiching polynomials" technique introduced by Bazzi to show a sufficiently strong bound for fooling the sum in each bucket so as to apply a union bound.

Theorem 9.2.1 (Fooling Small Combinatorial Sums)**.** *Let $f \in \mathsf{CSum}(m,n)$ with $\mathsf{Var}[f] \leq 6/\varepsilon^2$. For $k = 35$ and $t = C/\varepsilon^{15}$, the generator $G_{\mathcal{H},k,t}$ $O(\varepsilon)$-fools f in statistical distance.*

Fix a $f \in \mathsf{CSum}(m,n)$ with $\sigma^2 \leq 6/\varepsilon^2$ and let k, t be as above. Let $L = \{i : \sigma_i^2 \geq \varepsilon^5\}$. Since $\sigma^2 = \sum_i \sigma_i^2 \leq 6/\varepsilon^2$, we have $|L| \leq 6/\varepsilon^7$. For each bucket B_j we define the variable $T_j = \sum_{i \in B_j \setminus L} \sigma_i^2$. We say a hash function $h \in \mathcal{H}$ is *good* if the following conditions hold:

1. All variables in L are mapped to distinct buckets.

2. For every bucket B_j, $T_j \leq \varepsilon$.

Lemma 9.2.2. *A random hash function $h \in_u \mathcal{H}$ is good with probability at least $1 - 2\varepsilon$.*

Proof. By the pairwise independence of $\mathcal{H}$, each pair of variables $i \neq j \in L$ maps to the same bucket with probability $\frac{1}{t}$. By a union bound, the probability that condition (1) fails is at most $|L|^2/2t \leq \varepsilon$.

Fix $j \in [t]$ and for $i \in L^c$, let I_i be the indicator of the event $h(i) = j$. Then $T_j = \sum_{i \in L^c} \sigma_i^2 I_i$,

$$\mathbb{E}[T_j^2] = \mathbb{E}[\,(\sum_{i \in L^c} \sigma_i^2 I_i)^2\,] \leq \sum_{i \in L^c} \frac{\sigma_i^4}{t} + \sum_{i \neq l \in L^c} \frac{\sigma_i^2 \sigma_l^2}{t^2}$$

$$\leq (\max_{i \in L^c} \sigma_i^2) \sum_{i \in L^c} \frac{\sigma_i^2}{t} + \frac{1}{t^2} \left(\sum_{i \in L^c} \sigma_i^2 \right)^2 \leq \frac{\varepsilon^5 \sigma^2}{t} + \frac{\sigma^4}{t^2} \leq \frac{12\varepsilon^3}{t}.$$

Therefore, by Markov's inequality

$$\mathsf{Pr}[T_j > \varepsilon] < \frac{\mathbb{E}[T_j^2]}{\varepsilon^2} \leq \frac{\varepsilon}{t}$$

By a union bound, $T_j \leq \varepsilon$ holds for all $j \in [t]$ except with probability ε.

Thus overall h is good with probability $1 - 2\varepsilon$. □

The above lemma essentially reduces us to the case where all the indicator random variables in each bucket have very small variance, and thus have bias very close to 0 or 1. The following lemma lets us handle such variables.

Lemma 9.2.3. *Let $X = \sum_{i=1}^n X_i$ and $Y = \sum_{j=1}^n Y_j$ be sums of independent indicator random variables such that $\mathbb{E}[X], \mathbb{E}[Y] \leq \varepsilon$. Let D be a $(2d+2)$-wise independent distribution over $\{0,1\}^{2n}$ with the same coordinate-wise*

marginals as $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$. *Then, for* $(X'_1, \ldots, X'_n, Y'_1, \ldots, Y'_n) \leftarrow D$, $(\sum_i X'_i, \sum_i Y'_i)$ *is* $O_d(\varepsilon^d)$*-close in statistical distance to* (X, Y).

We note that a bound of $O(\varepsilon)$ is trivial for the lemma above: each of X and Y are non-zero with probability at most ε under a pairwise independent distribution. However we need a stronger $O(\varepsilon^d)$ bound so that we can use the union bound over all buckets, and this requires more work. We first prove Theorem 9.2.1 assuming the above lemma.

Proof of Theorem 9.2.1. Let $x \in [m]^n$ be the string generated by $G_{\mathcal{H},k,t}$ and let $y \in_u [m]^n$. Let $X_i = 1_{A_i}(x_i)$ and $Y_i = 1_{A_i}(y_i)$ be the indicator variables on each co-ordinate. Assume that the hash function h is good in the sense of Lemma 9.2.2. Then, each variable in L is mapped to a distinct bucket, so the values of $\{x_i\}_{i \in L}$ are uniform and independent. By coupling the variables x_i and y_i for $i \in L$, it suffices to show that $\sum_{i \in L^c} X_i$ and $\sum_{i \in L^c} Y_i$ are close in statistical distance when the distribution within each bucket B_j is $(k-1)$-wise independent, and the buckets are independent. To simplify our notation, we henceforth assume that $L = \varphi$ and $L^c = [n]$.

Fix a bucket B_j. We can partition B_j into $B_j^0 = \{i \in B_j : p_i < \frac{1}{2}\}$ and $B_j^1 = \{i \in B_j : p_i \geq \frac{1}{2}\}$. Let $\bar{X}_i = 1 - X_i$ for $i \in B_j^1$, so that $\Pr[\bar{X}_i = 1] = 1 - p_i$. Define variables $Z_j = \sum_{i \in B_j^0} X_i$ and $Z'_j = \sum_{i \in B_j^1} \bar{X}_i$.

$$\sum_{i \in B_j} X_i = \sum_{i \in B_j^0} X_i + \sum_{i \in B_j^1} (1 - \bar{X}_i) = Z_j - Z'_j + |B_j^1|.$$

Now, since h is good, $T_j \leq \varepsilon$, and $\mathbb{E}[Z_j], \mathbb{E}[Z'_j] \leq 2\varepsilon$. Since the distribution in each bucket is $k-1 \geq 34$-wise independent, we can apply Lemma 9.2.3 to the collections $\{X_i : i \in B_j^0\}$, $\{1 - X_i : i \in B_j^1\}$ with $d = 16$ to conclude that (Z_j, Z'_j) is $O(\varepsilon^{16})$-close in statistical distance to the distribution when the variables $X_i \in B_j$ are truly independent.

This implies that $\sum_{i \in B_j} X_i$ is $O(\varepsilon^{16})$ close in statistical distance to $\sum_{i \in B_j} Y_i$. Since variables across buckets are independent of one another, we conclude by a union bound that $\sum_{i \in [n]} X_i = \sum_{j \in [t]} \sum_{i \in B_j} X_i$ is $O(t\varepsilon^{16}) = O(\varepsilon)$ close in statistical distance to $\sum_{i \in [n]} Y_i$. □

9.2.1.1 Proof of Lemma 9.2.3

We start with a simple concentration bound for k-wise independent variables.

Lemma 9.2.4. *Let $X_1, \ldots, X_n$ be k-wise independent $\{0,1\}$ variables such that $\sum_{i=1}^n \mathbb{E}(X_i) \leq \beta$. Then for all $\ell \geq k$,*

$$\Pr[\sum_{i=1}^{n} X_i \geq \ell\,] \leq \left(\frac{e\beta}{\ell}\right)^k.$$

Proof. Let $S_k(X_1, \ldots, X_n) = \sum_{J \subseteq [n]; |J| = k} \prod_{j \in J} X_j$. By the k-wise independence of $X_1, \ldots, X_n$,

$$\mathbb{E}[S_k(X_1, \ldots, X_n)] = \sum_{J \subseteq [n]; |J| = k} \prod_{j \in n} \mathbb{E}[X_j].$$

But since $\sum_i \mathbb{E}[X_i] \leq \beta$, it follows that

$$\mathbb{E}[S_k(X_1, \ldots, X_n)] \leq \binom{n}{k} \frac{\beta^k}{n^k}.$$

This can be proved by the power-mean inequality, or a weight-shifting argument.

Note that if $\sum_i X_i \geq \ell$, then $S_k(X_1, \ldots, X_n) \geq \binom{\ell}{k}$. Hence by Markov's inequality,

$$\Pr[\sum_i X_i \geq \ell\,] \leq \frac{\mathbb{E}[S_k(X_1, \ldots, X_n)]}{\binom{\ell}{k}} \leq \frac{\binom{n}{k}\beta^k}{n^k\binom{\ell}{k}} \leq \left(\frac{e\beta}{\ell}\right)^k.$$

□

The following easy corollary follows by taking $k = \ell$:

Corollary 9.2.5. *If indicator random variables $X_1, \ldots, X_n$ are (fully) independent with $\sum_i \mathbb{E}[X_i] \leq \varepsilon$, then for $\ell \in [n]$,*

$$\Pr[\sum_{i=1}^{n} X_i \geq \ell] \leq \left(\frac{e\varepsilon}{\ell}\right)^\ell.$$

Let $X = \sum_i X_i$. Let $I_r(X)$ be the indicator random variable for the the event $X = r$. Let U denote the distribution where each X_i is drawn independently with $\mathbb{E}[X_i] = p_i$. We show that there exist constant degree *sandwiching polynomials* for $I_r(X)$.

Lemma 9.2.6. *Let $\mathbb{E}[X] \leq \varepsilon$. For $d \geq 2$ and every $r \leq d$, there exist univariate polynomials $P_r, Q_r : \mathbb{Z} \to \mathbb{Z}$ with $\deg(P_r), deg(Q_r) \leq d+1$ such that $P_r(i) \leq I_r(i) \leq Q_r(i)$, for all $i \in \mathbb{Z}_+$, and*

$$\mathop{\mathbb{E}}_{U}[Q_r(X) - P_r(X)] = O(\varepsilon^d).$$

Proof. Assume that $d-r$ is even. Let

$$Q_r(x) = \frac{1}{r!(d-r)!} \prod_{i \in \{0,\ldots,d\} \setminus \{r\}} (x-i), \quad P_r(x) = Q_r(x) \cdot \frac{d+1-x}{d+1-r}.$$

Clearly $P_r(\ell) = I_r(\ell) = Q_r(\ell) = 0$ for $\ell \in \{0,\ldots,d\} \setminus \{r\}$. Further, since $d-r$ is even, we have

$$P_r(r) = Q_r(r) = \frac{1}{r!(d-r)!} \prod_{i \in \{0,\ldots,d\} \setminus \{r\}} (r-i) = (-1)^{d-r} = 1.$$

Thus $P_r(\ell) = I_r(\ell) = Q_r(\ell)$ for $\ell \in \{0,\ldots,d\}$. For $\ell \geq d+1$ we have $I_r(\ell) = 0$ whereas

$$\frac{-\ell^d}{r!(d-r)!} \leq P_r(\ell) \leq 0, \quad \leq 0 \leq Q_r(\ell) \leq \frac{\ell^d}{r!(d-r)!}.$$

Hence $P_r(\ell) \leq I_r(\ell) \leq Q_r(\ell)$ as claimed. Further, using Corollary 9.2.5

$$\begin{aligned}
\mathop{\mathbb{E}}_{U}[Q_r(X) - P_r(X)] &\leq \sum_{\ell \geq d+1} (Q_r(\ell) - P_r(\ell)) \Pr[X \geq \ell] \leq \sum_{\ell \geq d+1} \frac{2\ell^d}{r!(d-r)!} \left(\frac{e\varepsilon}{\ell}\right)^\ell \\
&\leq \frac{2(e\varepsilon)^{d+1}}{r!(d-r)!} \sum_{\ell \geq d+1} \frac{1}{\ell^{\ell-d}} \\
&= O(\varepsilon^{d+1}).
\end{aligned}$$

In the case when $d-r$ is odd, it holds that $r \leq d-1$ and $d-1-r$ is even. So we repeat the above argument with d replaced by $d-1$ to get an error bound of $O(\varepsilon^d)$. □

Next we consider the setting where we have two sets $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_n$ of $\{0,1\}$ variables. Let $X = \sum_i X_i$ and $Y = \sum_j Y_j$ and $\mathbb{E}[X], \mathbb{E}[Y] \leq \varepsilon$. Let U^2 denote the distribution where all $2n$ variables are independent.

Corollary 9.2.7. *For any $d \geq 2$ and $r, s \in \{0, \ldots, d\}$ there are polynomials $P_{r,s}(X,Y)$ and $Q_{r,s}(X,Y)$ where $\deg(P_{r,s}), \deg(Q_{r,s}) \leq 2d+2$, $P_{r,s}(X,Y) \leq I_r(X)I_s(Y) \leq Q_{r,s}(X,Y)$ and*

$$\mathop{\mathbb{E}}_{U^2}[Q_{r,s}(X,Y) - P_{r,s}(X,Y)] = O(\varepsilon^d).$$

Proof. Let $P_{r,s}(X,Y) = P_r(X)Q_s(Y)$ and $Q_{r,s}(X,Y) = Q_r(X)Q_s(Y)$. Then, it follows from the calculations of the previous lemma that $P_{r,s}(X,Y) \leq I_r(X)I_s(Y) \leq Q_{r,s}(X,Y)$. Further,

$$\mathbb{E}[Q_{r,s}(X,Y) - P_{r,s}(X,Y)] = \mathop{\mathbb{E}}_{U}[Q_r(X) - P_r(X)] \cdot \mathop{\mathbb{E}}_{U}[Q_s(Y)] \leq O(\varepsilon^d(1+\varepsilon^d)) = O(\varepsilon^d),$$

where we used Lemma 9.2.6 to bound the error between $P_r(X), Q_r(X)$ and also to bound $\mathbb{E}[Q_s(Y)$ using

$$\mathop{\mathbb{E}}_{U}[Q_s(Y)] \leq \mathop{\mathbb{E}}_{U}[I_s(Y)] + \mathop{\mathbb{E}}_{U}[Q_s(Y) - I_s(Y)] \leq 1 + O(\varepsilon^d).$$

□

We now show that $(2d+2)$-wise independence on $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ suffices to fool (X,Y) in statistical distance. To do this we shall use the following observation due to Bazzi.

Lemma 9.2.8. *Let $f, g, h : V \to \{0,1\}$ be functions on a universe V such that $f \leq g \leq h$. Further, let D, D' be two distributions on V such that $\mathbb{E}_{u \leftarrow D}[h(u) - f(u)] \leq \varepsilon$ and*

$$|\mathop{\mathbb{E}}_{v \leftarrow D'}[f(v)] - \mathop{\mathbb{E}}_{u \leftarrow D}[f(u)]| \leq \delta, \quad |\mathop{\mathbb{E}}_{v \leftarrow D'}[h(v)] - \mathop{\mathbb{E}}_{u \leftarrow D}[h(u)]| \leq \delta.$$

Then,

$$|\mathop{\mathbb{E}}_{v\leftarrow D'}[g(v)] - \mathop{\mathbb{E}}_{v\leftarrow D}[g(v)]| \leq \varepsilon + \delta.$$

Proof. Let $u \leftarrow D, v \leftarrow D'$. Then,

$$\mathbb{E}[g(v)] \leq \mathbb{E}[h(v)] \leq \mathbb{E}[h(u)] + \delta \leq \mathbb{E}[f(u)] + \varepsilon + \delta \leq \mathbb{E}[g(u)] + \varepsilon + \delta.$$

A similar chain starting with f instead of h shows the lower bound and the lemma. □

Proof of Lemma 9.2.3. Let $X' = \sum_i X'_i, Y' = \sum_i Y'_i$. Fix $r, s \in \{0, 1 \ldots, d\}$. Then, as $(2d+2)$-wise independence fools degree $(2d+2)$ polynomials, by Corollary 9.2.7 and Lemma 9.2.8 we get that

$$|\,\mathsf{Pr}[(X,Y) = (r,s)] - \mathsf{Pr}[(X',Y') = (r,s)]| = \\ |\,\mathbb{E}[I_r(X)I_s(Y) = 1] - \mathbb{E}[I_r(X')I_s(Y')] = 1| = O(\varepsilon^d).$$

Further, by Lemma 9.2.4, $\mathsf{Pr}[X' \geq d+1 \vee Y' \geq d+1] \leq O(\varepsilon^{d+1})$. Therefore,

$$\begin{aligned}\mathsf{d_{TV}}((X,Y),(X',Y')) &= \sum_{0\leq r,s\leq n} |\,\mathsf{Pr}[(X,Y) = (r,s)] - \mathsf{Pr}[(X',Y') = (r,s)]| \\ &\leq \sum_{0\leq r,s\leq d} |\,\mathsf{Pr}[(X,Y) = (r,s)] - \mathsf{Pr}[(X',Y') = (r,s)]| + \\ &\mathsf{Pr}[X \geq d+1 \vee Y \geq d+1] + \mathsf{Pr}[X' \geq d+1 \vee Y' \geq d+1] \\ &\leq d^2 O(\varepsilon^d) + O(\varepsilon^d) = O(d^2\varepsilon^d).\end{aligned}$$

□

9.2.2 Fooling Large Combinatorial Sums in Kolmogorov Distance

We next show that the generator $G_{\mathcal{H},k,t}$ fools combinatorial sums in Kolmogorov distance when the variance σ^2 of the sum is large.

Theorem 9.2.9 (Fooling Large Combinatorial Sums)**.** *Let $f \in \mathsf{CSum}(m,n)$ with $\mathsf{Var}[f] \geq 1/\varepsilon^2$. Then for $k \geq 4$ and $t \geq 1/\varepsilon^2$, the generator $G_{\mathcal{H},k,t}$ $O(\varepsilon)$-fools f in Kolmogorov distance.*

We use the following property of pairwise independent hash functions. For a hash function $h \in_u \mathcal{H}$, Let $B_j = \{i : h(i) = j\}$ denote the j^{th} bucket of variables. Let $P_j = \sum_{i \in B_j} p_i$ and $S_j = \sum_{i \in B_j} \sigma_i^2$. Finally, let $S_h = (\sum_{j=1}^{t} S_j^2)^{\frac{1}{2}}$.

Lemma 9.2.10. *We have $\mathbb{E}_h[S_h] \leq \sigma + \sigma^2/\sqrt{t}$.*

Proof of Lemma 9.2.10. Fix $j \in [t]$. For each $i \in [n]$, let I_i be the indicator of the event $h(i) = j$ where $h \in_R \mathcal{H}$. Then, $\mathbb{E}_h[I_i] = 1/t$ and for $l \neq i$, $\mathbb{E}_h[I_i I_l] = 1/t^2$ by pairwise independence. As $S_j = \sum_{i=1}^{n} I_i \sigma_i^2$,

$$\begin{aligned}
\mathop{\mathbb{E}}_h[S_j^2] &= \sum_{i=1}^{n} \sigma_i^4 \mathop{\mathbb{E}}_h[I_i] + 2 \sum_{i \neq j} \sigma_i^2 \sigma_j^2 \, \mathbb{E}[I_i I_j] \\
&\leq \frac{1}{t} \sum_{i=1}^{n} \sigma_i^2 + \frac{2}{t^2} \sum_{i \neq j} \sigma_i^2 \sigma_j^2 \qquad \text{since } \sigma_i^4 \leq \sigma_i^2 \\
&\leq \frac{\sigma^2}{t} + \frac{\sigma^4}{t^2}.
\end{aligned}$$

Since $S_h^2 = \sum_{j=1}^{t} S_j^2$, using linearity of expectation we get

$$\mathop{\mathbb{E}}_h[S_h^2] \leq \sum_{j=1}^{t} \mathop{\mathbb{E}}_h[S_j^2] \leq \sigma^2 + \frac{\sigma^4}{t}.$$

The claim now follows using $\mathbb{E}_h[S_h] \leq \sqrt{\mathbb{E}_h[S_h^2]}$. □

Proof of Theorem 9.2.9. Let random variable $Y = f(y)$ for $y \in_u [m]^n$. Then, Y has a multinomial distribution with variance $\sigma^2 = \sum_i p_i(1-p_i) > 1/\varepsilon^2$. Therefore, by Corollary 4.2.1,

$$\mathsf{d}_{\mathsf{cdf}}\left(\frac{Y-\mu}{\sigma}, \mathbb{N}(0,1)\right) \leq \frac{1}{\sigma} = \varepsilon. \tag{9.2.2}$$

Let $x \in [m]^n$ be generated according to the generator $G_{\mathcal{H},k,t}$ with parameters as in the theorem and let indicator random variables $X_i = 1_{A_i}(x_i)$ and let $X = \sum_i X_i$. We shall show that $(X-\mu)/\sigma$ is also close to $\mathbb{N}(0,1)$. Fix a hash function $h \in \mathcal{H}$. Let $Z_j = \sum_{i \in B(j)} X_i$. Since the X_is are 4-wise independent, $\mathbb{E}[Z_j] = P_j$, $\mathrm{Var}[Z_j] = \sum_{i \in B_j} \sigma_i^2 = S_j$. Further, we have

$$\begin{aligned}
\mathbb{E}[(Z_j - P_j)^4] &= \mathbb{E}[(\sum_{i \in B_j}(X_i - p_i))^4] \\
&= \sum_{i \in B_j} \mathbb{E}[(X_i - p_i)^4] + 3 \sum_{i \neq l \in B_j} \mathbb{E}[(X_i - p_i)^2]\,\mathbb{E}[(X_l - p_l)^2] \\
&\leq \sum_{i \in B_j} \sigma_i^2 + 3 \sum_{i \neq l \in B_j} \sigma_i^2 \sigma_l^2 \qquad \text{since} (X_i - p_i)^4 \leq (X_i - p_i)^2 \\
&= S_j + 3S_j^2.
\end{aligned}$$

Therefore, summing over all j we get

$$\sum_{j=1}^{t} \mathbb{E}[(Z_j - P_j)^4] \leq \sum_{j=1}^{t} S_j + 3\sum_{j=1}^{t} S_j^2 = \sigma^2 + 3S_h^2.$$

Using the Berry-Esséen theorem applied to independent random variables $Z_1, \ldots, Z_t$, for a fixed hash function h,

$$\mathsf{d}_{\mathsf{cdf}}\left(\frac{X-\mu}{\sigma}, \mathbb{N}(0,1)\right) \leq \frac{(\sigma^2 + 3S_h^2)^{1/2}}{\sigma^2} \leq 2\left(\frac{1}{\sigma} + \frac{S_h}{\sigma^2}\right).$$

Further, as $\mathsf{d}_{\mathsf{cdf}}$ is a convex function, using Lemma 9.2.10,

$$\mathsf{d}_{\mathsf{cdf}}\left(\frac{X-\mu}{\sigma}, \mathbb{N}(0,1)\right) \leq 2\left(\frac{1}{\sigma} + \frac{\mathbb{E}_h[S_h]}{\sigma^2}\right) \leq 2\left(\frac{2}{\sigma} + \frac{1}{\sqrt{t}}\right) \leq 6\varepsilon.$$

By Equation (9.2.2) we get $\mathsf{d}_{\mathsf{cdf}}((X-\mu)/\sigma, (Y-\mu)/\sigma) = O(\varepsilon)$ which implies $\mathsf{d}_{\mathsf{cdf}}(X,Y) = O(\varepsilon)$. □

9.2.3 Reducing the seed-length via INW

We now derandomize $G_{\mathcal{H},k,t}$ using PRGs for small space sources of Impagliazzo, Nisan, and Wigderson [46] (Theorem 2.3.1), which we call INW PRG, as was done in Section 7.4.3. The derandomization follows from Theorems 9.2.1, 9.2.9 and replacing the independent seeds $z^1, \ldots, z^t$ in Equation 9.2.1 with the output of the INW PRG.

Theorem 9.2.11 (Derandomizing $G_{\mathcal{H},k,t}$)**.** *There exists a generator* $G \equiv G_{m,n,\varepsilon} : \{0,1\}^{r_{m,n}} \to [m]^n$ *with seed-length* $r_{m,n} = O(\log m + \log n + \log^2(1/\varepsilon))$ *with the following properties:*

1. *G $O(\varepsilon)$-fools all $f \in \mathsf{CSum}(m,n)$ with $\mathsf{Var}[f] < 6/\varepsilon^2$ in statistical distance.*
2. *G $O(\varepsilon)$-fools all $f \in \mathsf{CSum}(m,n)$ with $\mathsf{Var}[f] > 1/\varepsilon^2$ in Kolmogorov distance.*

Consider $G_{\mathcal{H},k,t}$ with parameters set so as to satisfy the conditions of Theorems 9.2.1, 9.2.9. Note that the seed length of $G_{\mathcal{H},k,t}$ is $O((\log n)\mathrm{poly}(1/\varepsilon))$. We will reduce the seed length by choosing the seeds $z^1, \ldots, z^t$ from the output

of the INW PRG (instead of independently as before). The analysis proceeds roughly by arguing that for any (m, n)-combinatorial sum f and hash function $h \in \mathcal{H}$, $f(G_{\mathcal{H},k,t}(h, z^1, \ldots, z^t)) \equiv g_h(z^1, \ldots, z^t)$ is computable by a small-space machine when viewed as a function of $z^1, \ldots, z^t$.

Let $\mathsf{INW} : \{0,1\}^r \to (\{0,1\}^{r_k})^t$ be the INW generator that ε-fools $(10 \log(1/\varepsilon), r_k, t)$, read-once branching programs. Define

$$G : \mathcal{H} \times \{0,1\}^r \to [m]^n \text{ by } G(h, y) = G_{\mathcal{H},k,t}(h, \mathsf{INW}(y)).$$

We claim that G satisfies the conditions of Theorem 9.2.11.

Proof of Theorem 9.2.11. The claim on the seed length of G follows from the seed length of the INW generator, Theorem 2.3.1, which uses $r = O(r_k + (\log(1/\varepsilon) + \log(t/\varepsilon)) \log t) = O(\log m + \log n + \log^2(1/\varepsilon))$ bits. We next show that G satisfies properties (1), (2).

Fix an (m, n)-combinatorial sum f and let x be the output of generator $G_{\mathcal{H},k,t}$ with parameters as above. Fix a hash function $h \in \mathcal{H}$ and define $g_h : (\{0,1\}^{r_k})^t \to [n]$ by $g_h(z^1, \ldots, z^t) = f(G_{\mathcal{H},k,t}(h, z^1, \ldots, z^t))$. For $\ell \in [t]$, let $B_\ell = \{i : h(i) = \ell\}$ and let random variable $Y_\ell = \sum_{j : j \in B_\ell} 1_{A_j}(x_j)$. Then, Y_ℓ depends only on z^ℓ and $g_h(z^1, \ldots, z^t) = \sum_\ell Y_\ell$.

There is a natural $(\log n, r_k, t)$-ROBP M for computing g_h: the vertices of M are labeled $\{1, \ldots, n\}$ with states in layer ℓ corresponding to the possible values of the partial sum $\sum_{i \le \ell} Y_i$ and the edges out of layer ℓ are drawn according to the change in the value of the partial sum. However, using M directly

to do the derandomization is problematic as G_S only fools $O(\log(1/\varepsilon))$ space ROBPs. We get over this hurdle by appropriately sandwiching M between smaller-width branching programs.

Case 1: $\mathsf{Var}[f] < 6/\varepsilon^2$. Observe that $x_1, \ldots, x_n$ are k-wise independent. Therefore, by an argument similar to that of Lemma 9.2.4, it follows that for $\ell \in [t]$,

$$\mathsf{Pr}[\,|\sum_{j\leq\ell}(Y_j - \mu(Y_j))| > 6e/\varepsilon^4\,] \leq \varepsilon^{2k}. \tag{9.2.3}$$

We exploit this fact by ignoring all states of M corresponding to partial sums not in $I = [-6e/\varepsilon^4, 6e/\varepsilon^4]$.

Fix a statistical test function $F : [n] \to \{0,1\}$. Let $\bar{z} = (z^1, \ldots, z^t) \in_u (\{0,1\}^{r_k})^t$. Observe that $F(\bar{z}) \equiv F(g_h(\bar{z})) = F(M(z))$ is computable by a $(\log n, r_k, t)$-ROBP, say M'. We now sandwich F between two small-width branching programs. Let M_u be a ROBP that works the same as M' except that it accepts all strings $\bar{z}$ that lead to a partial sum $\sum_{i\leq\ell}(Y_i - \mu(Y_i)) \notin I$. Similarly, let M_l be a machine a ROBP that works the same as M' except that it rejects all strings $\bar{z}$ that lead to a partial sum $\sum_{i\leq\ell}(Y_i - \mu(Y_i)) \notin I$. Then, $M_l \leq M' \leq M_u$ and M_l, M_u are computable by $((\log|I|)+1, r_k, t)$-ROBPs. Further, from Equation 9.2.3 and a union bound over $\ell \in [t]$,

$$\mathsf{Pr}[M_u(z) = 1] - \mathsf{Pr}[M_l(z) = 1] \leq t\varepsilon^{2k} = O(\varepsilon).$$

Now, as G_S fools M_u, M_l with error at most ε, it follows from the above equation and the sandwiching property (Lemma 9.2.8) that G_S fools M' with error

at most $O(\varepsilon)$. The theorem now follows from the above fact and Theorem 9.2.1.

Case 2: $\mathsf{Var}[f] > 1/\varepsilon^2$. This case follows straightforwardly from Theorem 9.2.9 and the monotone trick argument we used in Section 7.4.3. We skip the details to avoid repetition. □

9.2.4 Fooling Combinatorial Sums

We now combine the generators from the previous section to get our final generator fooling combinatorial sums in statistical distance. The basic idea is as follows: we partition the n variables into two subsets L, R with $|L| \sim n/2$, and then use $G_{m,n/2}$ for the variables in L and an independent $G_{m,n/2}$ on the variables in R. We analyze the construction by induction and considering two cases. If the variance of the combinatorial sum is small, we invoke Theorem 9.2.11 (1). So now assume that the variance is large.

Let f be a combinatorial sum with $\mathsf{Var}[f] > 6/\varepsilon^2$ and write $f = f_L + f_R$, where f_L, f_R are the combinatorial sums obtained by restricting to variables in L, R respectively. We use the induction hypothesis to get a statistical distance guarantee for f_L and use Theorem 9.2.11 (2) to get a Kolmogorov distance guarantee for f_R. We then argue that the combinatorial sum f_L has high variance and hence is shift invariant. We then apply Lemma 4.3.1 and get a statistical distance guarantee for $f = f_L + f_R$.

Fix $\varepsilon \in [1/\sqrt{n}, 1/\log n]$ and let $s = \log(n+1)$. Let $\mathcal{H}_1 = \{\pi : [n] \to [n]\}$ be a family of pairwise independent permutations. Efficient constructions of

$\mathcal{H}_1$ with $\mathcal{H}_1 = \text{poly}(n)$ are known. We pick $\pi \in_u \mathcal{H}_1$ and use it to partition $[n]$ into s buckets of geometrically increasing sizes. We define sets $B_1, \ldots, B_s$ where $B_j = \{\pi(2^{j-1}), \ldots, \pi(2^j - 1)\}$, thus $|B_j| = 2^{j-1}$. Let r_j be the seed-length of the generator $G_{m,2^{j-1},\varepsilon}$ from Theorem 9.2.11. Our main generator $G : \mathcal{H}_1 \times \{0,1\}^{r_1} \times \cdots \times \{0,1\}^{r_s} \to [m]^n$ uses an independent sample from $G_{m,2^{j-1},\varepsilon}$ for each bucket B_j:

$$G(\pi, z^1, \ldots, z^s) = x, \text{ where } x_{B_j} = G_{m,2^{j-1},\varepsilon}(z^j). \tag{9.2.4}$$

As before, let $f(x_1, \ldots, x_n) = \sum_{i=1}^n X_i$ where $X_i = 1_{A_i}(x_i)$ has mean p_i and variance σ_i^2. For each bucket B_j, let $S_j = \sum_{i \in B_j} \sigma_i^2$. Let $q \in \{1, \ldots, s\}$ be the least index such that $\mathbb{E}[S_q] > 3/\varepsilon^2$.

Call a permutation π *bad* if one of the following conditions holds and *good* otherwise:

1. There exists an index $j \in \{q, \ldots, s\}$ such that $S_j \notin [0.5\,\mathbb{E}[S_j], 1.5\,\mathbb{E}[S_j]]$.
2. There exists $j \in \{1, \ldots, q-1\}$ such that $S_j \geq 6/\varepsilon^2$.

Note that the sequence $\{\mathbb{E}[S_j]\}_{j=1}^s$ is in geometric progression. If π is good, then $\{S_j\}_{j=q}^s$ is roughly geometric, and none of $\{S_j\}_{j \leq q}$ are too large.

Claim 9.2.12. $\Pr_{\pi \in_u \mathcal{H}_1}[\pi \text{ is bad}] \leq 2\varepsilon$.

Proof. Fix $j \in \{q, \ldots, s\}$. Let Z_i be the indicator of the event $\pi^{-1}(i) \in \{2^{j-1}, \ldots, 2^j - 1\}$ and hence $i \in B_j$. Then

$$S_j = \sum_{i=1}^n \sigma_i^2 Z_j \;\Rightarrow\; \mathbb{E}[S_j] = \frac{\sigma^2 2^{j-1}}{n}.$$

By the pairwise-independence of π,

$$\begin{aligned}\mathbb{E}[S_j^2] &= \sum_i \sigma_i^2 \,\mathbb{E}[Z_i] + \sum_{i \neq l} 2\sigma_i^2 \sigma_l^2 \,\mathbb{E}[Z_i Z_l] \leq \frac{\sigma^2 2^{j-1}}{n} + \frac{\sigma^2 2^{j-1}(2^{j-1}-1)}{n(n-1)} \\ &\leq \frac{\sigma^2 2^{j-1}}{n} + \frac{\sigma^4 2^{2(j-1)}}{n^2},\end{aligned}$$

hence, $\mathsf{Var}[S_j] \leq \mathbb{E}[S_j^2] - \mathbb{E}[S_j]^2 \leq \sigma^2 2^{j-1}/n = \mathbb{E}[S_j]$.

We now bound the probability of bad event (1). Fix $j \in \{q, \ldots, s\}$ so that $\mathbb{E}[S_j] \geq \frac{3}{\varepsilon^2}$. By Chebychev's inequality

$$\mathsf{Pr}\left[|S_j - \mathbb{E}[S_j]| > \frac{\mathbb{E}[S_j]}{2}\right] \leq \frac{4\,\mathrm{Var}[S_j]}{(\mathbb{E}[S_j]^2)} \leq \frac{4}{\mathbb{E}[S_j]} \leq 2\varepsilon^2.$$

Similarly, to bound bad event (2), we observe that $\mathbb{E}[S_j] \leq 3/\varepsilon^2$ for $j \leq q-1$, hence

$$\mathsf{Pr}[S_j \geq 6/\varepsilon^2] \leq \Pr[|S_j - \mathbb{E}[S_j]| > 3/\varepsilon^2] \leq \varepsilon^4 \,\mathrm{Var}[S_j]/9 \leq \varepsilon^2.$$

Since $\varepsilon < 1/\log n$, the claim follows by a union bound over $i \in \{1, \ldots \log n\}$.

□

Theorem 9.2.13. *The Generator G fools* $\mathsf{CSum}(m,n)$ *with error* $O(\log n \sqrt{\varepsilon \log(1/\varepsilon)})$.

Proof. Let $x \in [m]^n$ be sampled from G, while $y \in_u [m]^n$. Let $X_i = 1_{A_i}(x_i), Y_i = 1_{A_i}(y_i)$ and

$$X^j = \sum_{i \in B_j} X_i, \quad Y^j = \sum_{i \in B_j} Y_i, \qquad X^{\leq j} = \sum_{l \leq j} X^l, \quad Y^{\leq j} = \sum_{l \leq j} Y^l.$$

We assume from now on we condition on the chosen permutation π being good. Observe that $\mathbb{E}[X^j] = \mathbb{E}[Y^j]$ and

$$\mathrm{Var}[X^j] = \mathrm{Var}[Y^j] = \sum_{i \in B_j} \mathsf{Var}[X_i] = \sum_{i \in B_j} \sigma_i^2 = S_j.$$

We claim that there is a constant C such that for $j \in [s]$,

$$\mathsf{d_{TV}}(X^{\le j}, Y^{\le j}) \le Cj\sqrt{\varepsilon(\log(1/\varepsilon))}. \tag{9.2.5}$$

The proof is by induction on j. It is easy to prove for $j \le q$. Since $\mathsf{Var}[X^l] = \mathsf{Var}[Y^l] = S_l < 6/\varepsilon^2$ for all $l \le j$, by Theorem 9.2.11 (1), $\mathsf{d_{TV}}(X^l, Y^l) \le \varepsilon$. As $X^1, \ldots, X^j$ are independent of one another, we have $\mathsf{d_{TV}}(X^{\le j}, Y^{\le j}) \le j\varepsilon$. Now consider $j \in \{q+1, \ldots, s\}$. We have

$$\begin{aligned}&\mathsf{d_{TV}}(X^{\le j-1} + X^j, Y^{\le j-1} + Y^j) \le \\ &\quad \mathsf{d_{TV}}(X^{\le j-1} + X^j, Y^{\le j-1} + X^j) + \mathsf{d_{TV}}(Y^{\le j-1} + X^j, Y^{\le j-1} + Y^j).\end{aligned} \tag{9.2.6}$$

The first term can be bounded using the induction hypothesis:

$$\mathsf{d_{TV}}(X^{\le j-1} + X^j, Y^{\le j-1} + X^j) \le \mathsf{d_{TV}}(X^{\le j-1}, Y^{\le j-1}) \le C(j-1)\sqrt{\varepsilon(\log(1/\varepsilon))}. \tag{9.2.7}$$

To bound the second term, we will apply Corollary 4.3.2. As π is good and $j > q$, $\mathsf{Var}[X^j] = \mathsf{Var}[Y^j] = S_j \ge \mathbb{E}[S_j]/2 > 1/\varepsilon^2$. Thus the variance is sufficiently large to apply Theorem 9.2.11 (2), which gives $\mathsf{d_{cdf}}(X^j, Y^j) < \varepsilon$. Moreover, by Fact 4.2.4,

$$\mathsf{Pr}\left[|Y^j - \mathbb{E}[Y^j]| > 3\sqrt{S_j \log(1/\varepsilon)}\right] \le \varepsilon.$$

Since X^j and Y^j have the same mean and $\mathsf{d_{cdf}}(X^j, Y^j) < \varepsilon$, we get similar concentration for X^j:

$$\mathsf{Pr}\left[|X^j - \mathbb{E}[X^j]| > 3\sqrt{S_j \log(1/\varepsilon)}\right] \leq 3\varepsilon.$$

Thus, with probability $1 - 4\varepsilon$, we have $X^j, Y^j \in [\mathbb{E}[X^j] - b, \mathbb{E}[X^j] + b]$, where $b = 3\sqrt{S_j \log(1/\varepsilon)}$. Further, since π is good, we have

$$\mathrm{Var}[Y^{\leq j-1}] \geq \mathrm{Var}[Y^{j-1}] = S_{j-1} > \mathbb{E}[S_{j-1}]/2 \geq \mathbb{E}[S_j]/4 > S_j/6.$$

Hence by Fact 4.2.3, $Y^{\leq j-1}$ is $\alpha = (6/\sqrt{S_j})$-shift invariant.

We can now apply Corollary 4.3.2 with $\alpha = 6/\sqrt{S_j}$ and $b = 6\sqrt{S_j \log(1/\varepsilon)}$ to get

$$\mathsf{d_{TV}}(Y^{\leq j-1} + X^j, Y^{\leq j-1} + Y^j) \leq 24\sqrt{\varepsilon \log(1/\varepsilon)} + 4\varepsilon. \tag{9.2.8}$$

Substituting the bounds from Equations (9.2.7) and (9.2.8) back into Equation (9.2.6) gives

$$\mathsf{d_{TV}}(X^{\leq j}, Y^{\leq j}) \leq C(j-1)\sqrt{\varepsilon \log(1/\varepsilon)} + 24\sqrt{\varepsilon \log(1/\varepsilon)} + 4\varepsilon \leq Cj\sqrt{\varepsilon \log(1/\varepsilon)},$$

where $C = 30$. □

We now derandomize the generator of Theorem 9.2.13 to get our main result for fooling combinatorial shapes.

Proof of Theorem 9.1.1. We derandomize the generator G of Equation 9.2.4 as was done in Theorem 9.2.11 by choosing the seeds $z^1, \ldots, z^s$ from the output of

PRGs for ROBPs. Fix $\delta > 0$ and set the parameters of G as in Theorem 9.2.13 with $\varepsilon = \delta/(\log(1/\delta) \cdot \log n)$. Fix a (m, n)-combinatorial shape f and note that for a hash function $g \in \mathcal{H}_1$, $f(G(g, z^1, \ldots, z^s))$ when viewed as a function of $z^1, \ldots, z^s$ is computable by a (S, D, T)-ROBP, where $S = \log n$, $D = O(\log m + \log n + \log^2(1/\varepsilon))$, and $T = s = O(\log n)$. Further, as $T = O(S + D)$, such ROBPs can be fooled with error ε and seed length $O(\log m + \log n + \log^2(1/\varepsilon))$ by using the PRG of [81].

Let G be the generator obtained from G by using the PRG of [81] with parameters as above to generate the seeds $z^1, \ldots, z^s$ of Equation 9.2.4 instead of independently as before. Then, by Theorem 9.2.13, G $O(\delta)$-fools (m, n)-combinatorial sums with seed length $O(\log m + \log n + \log^2(1/\varepsilon)) = O(\log m + \log n + \log^2(1/\delta))$. □

9.3 PRGs for Combinatorial Rectangles

We prove that the generator $G_{\mathcal{H},k,t}$ of Section 7.3 with $k = O(\sqrt{\log(1/\varepsilon)})$ and $t = \exp(O(\sqrt{\log n}))$ and $\mathcal{H}$ k-wise independent fools combinatorial rectangles. We then derandomize the generator using the INW generator as in the proofs of Theorems 9.2.11 and 9.1.1 to get our final PRG for combinatorial rectangles. As mentioned before, though our result is weaker than Lu's generator, our construction is perhaps simpler than Lu's and our analysis is different from Lu's. Moreover, we match Lu's parameters for the important case when the desired error $\varepsilon = \text{poly}(n)$.

Theorem 9.3.1. *The generator $G_{\mathcal{H},k,t}$ with $k = 5\sqrt{\log(1/\varepsilon)}$, $t = \exp(5\sqrt{\log(1/\varepsilon)})$*

and $\mathcal{H}$ a k-wise independent family of hash functions, fools combinatorial rectangles with error at most $O(\varepsilon)$.

We use the following properties of a k-wise independent family of hash functions.

Lemma 9.3.2. *For $\mathcal{H} = \{h : [n] \to [t]\}$, k-wise independent, the following properties hold.*

1. *For any $L \subseteq [n]$, $|L| \leq r$, $\mathsf{Pr}[\ \exists \ell,\ |h^{-1}(\ell) \cap L| \ \geq k/2\ ] \leq t \cdot (2re/kt)^{k/2}$.*

2. *Let $q_1, \ldots, q_n \in [0,1]$, $\sum_i q_i = Q$ and $\max_i q_i \leq \beta Q$. Then, for any $\ell \in [t]$,*

$$\mathsf{Pr}[\ \sum_{i:h(i)=\ell} q_i \ \geq Q/t + \beta^{1/4} Q\ ] \leq 2(k\beta^{1/2} \log(1/\beta))^{k/2}.$$

Proof. (1). Without loss of generality, let $L = \{1, \ldots, r\}$. Fix $\ell \in [t]$ and let $X_1, \ldots, X_n$ be indicator random variables with $X_i = 1$ if $h(i) = \ell$ and 0 else. Then, $X_1, \ldots, X_r$ are k-wise independent and

$$\mathsf{Pr}[\sum_i X_i \geq k/2\,] \leq \mathbb{E}[\ \sum_{J \subseteq [r], |J| = k/2} \ \prod_{j \in J} X_j\ ] = \binom{r}{k/2} \frac{1}{t^{k/2}} \leq \left(\frac{2re}{kt}\right)^{k/2}.$$

The claim now follows by taking a union bound over $\ell \in [t]$.

(2). Fix $\ell \in [t]$ and let $X_1, \ldots, X_n$ be as above. Then, $X = \sum_{i:h(i)=\ell} q_i = \sum_i q_i X_i$, where the X_i are k-wise independent with $\mathsf{Pr}[X_i = 1] = 1/t$. Let

$Y_1, \ldots, Y_n$ be independent random variables with $\mathsf{Pr}[Y_i = 1] = 1/t$ and $Y = \sum_i q_i Y_i$. Then, by Hoeffding's inequality, for all $\gamma > 0$,

$$\mathsf{Pr}[\,|Y - Q/t| \geq \gamma\,] \leq 2\exp(-2\gamma^2/\sum_i q_i^2) \leq 2\exp(-2\gamma^2/\beta Q^2).$$

Let k be even and fix $\gamma > 0$ to be chosen later. Then, as $Y \leq Q$,

$$\mathbb{E}[\,(Y - Q/t)^k\,] \leq \gamma^k + Q^k\,\mathsf{Pr}[\,|Y - Q/t| \geq \gamma\,] \leq \gamma^k + Q^k 2\exp(-2\gamma^2/\beta Q^2).$$

Since $\mathbb{E}[(X - Q/t)^k] = \mathbb{E}[(Y - Q/t)^k]$, it follows from Markov's inequality that for any $\theta > 0$,

$$\mathsf{Pr}[\,|X - Q/t| > \theta\,] \leq \frac{\gamma^k + Q^k 2\exp(-2\gamma^2/\beta Q^2)}{\theta^k}.$$

Setting $\theta = \beta^{1/4} \cdot Q$, $\gamma = (2k\beta\log(1/\beta))^{1/2}Q$, we get

$$\mathsf{Pr}[\,|X - Q/t| > \beta^{1/4}Q\,] \leq 2(k\beta^{1/2}\log(1/\beta))^{k/2}.$$

□

Proof of Theorem 9.3.1. Fix an (m, n)-combinatorial rectangle $f : [m]^n \to \{0, 1\}$ with $f(x_1, \ldots, x_n) = 1_{A_1}(x_1) \wedge 1_{A_2}(x_2) \cdots 1_{A_n}(x_n)$. Let $y \in_u [m]^n$ and $Y_i = 1_{A_i}(y_i)$, $q_i = 1 - \mathbb{E}[Y_i]$. Let x be the output of the generator with parameters as in the statement. Let $X_i = 1_{A_i}(x_i)$ and $X = \sum_i X_i$. Note that

$$\mathsf{Pr}[f(y) = 1] = (1 - q_1)(1 - q_2)\cdots(1 - q)_n \leq \exp(-\sum_i q_i).$$

Therefore, if $\sum_i q_i > \log(1/\varepsilon)$, then $\mathsf{Pr}[f(y) = 1] < \varepsilon$. We accordingly consider two cases to analyze our generator.

Case 1: $Q = \sum_i q_i \leq 3\log(1/\varepsilon)$. Let $L = \{i : q_i > Q/\sqrt{t}\}$, $L^c = [n]/L$. Then, $|L| < \sqrt{t}$ and by Lemma 9.3.2 (1) it follows that for $h \in_u \mathcal{H}$, $\max_\ell |h^{-1}(\ell) \cap L| \leq k/2$ with probability at least $1 - 1/t^{\Omega(k)} = 1 - \varepsilon$. Consequently, for a random h we can assume that the variables in L are truly independent of one another. Moreover, when conditioned on the variables in L, the variables from L^c in each bucket, $\{x_i : i \in B_\ell = h^{-1}(\ell), \ \wedge i \notin L^c\}$ for $\ell \in [t]$, are $(k/2)$-wise independent. To simplify notation we assume that $L = \emptyset$ and analyze the case where the X_i's in a single bucket are $(k/2)$-wise independent.

Now, for $\beta = 1/\sqrt{t}$, $\max_i q_i < \beta Q$. Therefore, by Lemma 9.3.2 (2), for $h \in_u \mathcal{H}$ with probability at least $1 - \varepsilon$, $Q^\ell = \sum_{i:h(i)=\ell} q^i < 6\log(1/\varepsilon)/t^{1/8}$ for all $\ell \in [t]$. Further, by the principle of inclusion-exclusion and $(k/2)$-wise independence of $X_i, i \in B_\ell$,

$$
\begin{aligned}
|\,\mathsf{Pr}[\wedge_{i\in B_\ell} X_i] - \mathsf{Pr}[\wedge_{i\in B_\ell} Y_i]| &\leq \sum_{J\subseteq B_\ell, |J|=k/2} \mathsf{Pr}[\wedge_{i\in J} X_i] \\
&\leq \binom{|B_\ell|}{k/2}\left(\frac{Q^\ell}{|B_\ell|}\right)^{k/2} \qquad \text{(power-mean inequality)} \\
&\leq \left(\frac{2eQ^\ell}{k}\right)^{k/2} \\
&= \left(\frac{O(\sqrt{\log(1/\varepsilon)})}{t^{1/8}}\right)^{k/2} = O(\varepsilon/t).
\end{aligned}
$$

Therefore, as the X_i's in different buckets are independent of one another, by a union bound over $\ell \in [t]$ it follows that $|\,\mathsf{Pr}[\wedge_i X_i = 1] - \mathsf{Pr}[\wedge_i Y_i = 1]| = O(\varepsilon)$.

Case 2: $\sum_i q_i > 3\log(1/\varepsilon)$. Let $j \in [n]$ be the maximum index such that $\sum_i q_i \leq 3\log(1/\varepsilon)$. Then, $\sum_{i \leq j} q_i \geq 3\log(1/\varepsilon) - 1 > 2\log(1/\varepsilon)$. Therefore, $\mathsf{Pr}[\wedge_{i \leq j} Y_i = 1] \leq \exp(-\sum_{i \leq j} q_i) \leq \varepsilon$. Now, by applying the argument of the previous case to the collection of variables $X_1, \ldots, X_j$ it follows that $\mathsf{Pr}[\wedge_{i \leq j} X_i = 1] = O(\varepsilon)$. Therefore, $\mathsf{Pr}[\wedge_i X_i = 1] = O(\varepsilon)$ from which the claim follows. □

Proof of Theorem 9.1.2. The theorem follows by derandomizing $G_{\mathcal{H},k,t}$ with parameters as above by using the INW PRG to generate $z^1, \ldots, z^t$ of Equation 7.3.1 instead of independently as before. □

Chapter 10

Open Problems

10.1 Invariance Principles

An obvious question arising from our invariance principle for polytopes, Theorem 3.3.1, is to improve the error estimate $\log^{O(1)}(k) \cdot \varepsilon^{\Omega(1)}$. To this end, we conjecture that the right bound should be of the same order as Nazarov's bound on the Gaussian noise-sensitivity of polytopes:

Conjecture 10.1. *For any ε-regular polytope $\mathcal{K} \subseteq \mathbb{R}^n$,*

$$\left| \Pr_{X \in_u \{1,-1\}^n} [X \in \mathcal{K}] - \Pr_{Y \leftarrow \mathcal{N}^n} [Y \in \mathcal{K}] \right| = O(\varepsilon \sqrt{\log k}).$$

In contrast to the above question, we could also ask how good an invariance principle we can get for polytopes. For a single halfspace it is easy to see that the right bound is $\Theta(\varepsilon)$. Embarrassingly, we do not know of any better lowerbounds for polytopes.

Question 10.2. For every sufficiently small $0 < \varepsilon < 1$, does there exist an ε-regular polytope $\mathcal{K} \subseteq \mathbb{R}^n$, with $k = \text{poly}(n)$ faces such that

$$\left| \Pr_{X \in_u \{1,-1\}^n} [X \in \mathcal{K}] - \Pr_{Y \leftarrow \mathcal{N}^n} [Y \in \mathcal{K}] \right| = \omega(\varepsilon)?$$

10.2 Sensitivity of Boolean Functions

The foremost question here is of course to prove the Gotsman-Linial conjecture. We restate their conjecture below:

Conjecture 10.3 (Gotsman and Linial, [39]). *For any degree d PTF $f : \{1,-1\}^n \to \{1,-1\}$, $\mathbb{NS}_\delta(f) = O(d\sqrt{\delta})$.*

In this regard, it would be of great interest to even obtain a bound on noise sensitivity of the form $O_d(\delta^{\Omega(1)})$, as our techniques do not seem capable of avoiding a $\delta^{\Omega(1/d)}$ dependency due to a similar loss in the invariance principle, Theorem 2.2.5.

One obvious weakness of our sensitivity bounds for polytopes, Chapter 6, is the regularity requirement. A natural approach to remove the restriction would be to use a suitable *regularity lemma* to "reduce" the problem for arbitrary polytopes to the regular case as we did for the case of PTFs. Unfortunately, applying the reductions to the regular case as in the case of PTFs leads to bounds that are at least linear in k, even when using our stronger bounds for the regular case. We (optimistically) believe that the above difficulty could be overcome and a better reduction to the regular case can be achieved.

10.3 Pseudorandom Generators

The main open question here is to get better PRGs for low-degree PTFs. It would be interesting to even obtain improvements over our results

in any of the following directions:

1. A PRG for degree d PTFs with error ε and seed-length $O_d(\log n \cdot \mathrm{poly}(\log(1/\varepsilon)))$.
2. A PRG for degree d PTFs with error ε and seed-length $O(\log n \cdot \mathrm{poly}(d) \cdot \mathrm{poly}(1/\varepsilon))$.
3. A PRG for halfspaces with error $1/\mathrm{poly}(n)$ and seed-length $o(\log^2 n)$ (say, $\log^{3/2} n$).

As we explain below, the first two questions seem particularly challenging for our current techniques.

The issue with (1) is that the main technique leading to our improvement for halfspaces - PRGs for ROBPs of Nisan [80], and Impagliazzo, Nisan and Wigderson [46] seems inapplicable here as allowing even degree $d = 2$ spoils the read once structure. One plausible approach is to try an idea similar to that of Bogdanov, Viola [20] and Lovett [66] and Viola [97] for constructing PRGs fooling low-degree polynomials over finite fields:

Question 10.4. Fix $d > 1$. Do there exist functions $f, g : \mathbb{N} \to \mathbb{N}$ such that the XOR (sum modulo 2) of $f(d)$ pseudorandom generators for halfspaces with error ε fools degree d PTFs with error $O_d(\varepsilon^{1/g(d)})$?

The problem with (2) is that our current analysis, as well as the invariance principle for PTFs of Mossel et al. [76] uses hypercontractivity for degree d polynomials critically, and any use of hypercontractivity seems to incur a loss of at least $2^{O(d)}$. To this end, the first more principled question is to ask

for an improvement to the invariance principle, Theorem 2.2.5, to get an error bound that is not exponential in d.

Finally, question (3) seems more tractable to us. In particular, we believe that the generator $G_{\mathcal{H},k,t}$ from Section 7.3 with $k = O(\sqrt{\log n})$, $t = 1/\exp(O(\sqrt{\log n}))$ and $\mathcal{H}$ a k-wise independent hash family would fool halfspaces with error $1/\mathrm{poly}(n)$. Combined with the monotone trick as in Section 7.4.3, such a result would lead to a PRG for halfspaces with seed-length $O(\log^{3/2} n)$ and error $1/\mathrm{poly}(n)$.

Bibliography

[1] Nir Ailon and Bernard Chazelle. The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM J. Computing*, 39(1):302–322, 2009.

[2] N.I Akhiezer. *The Classical Moment Problem and Some Related Questions in Analysis*. Hanfer Publishing Co, 1 edition, 1965.

[3] N. Alon, O. Goldreich, J. Håstad, and R. Peralta. Simple constructions of almost k–wise independent random variables. *Random Structures and Algorithms*, 3(3):289–303, 1992.

[4] Noga Alon, Oded Goldreich, Johan Håstad, and René Peralta. Simple construction of almost k-wise independent random variables. *Random Struct. Algorithms*, 3(3):289–304, 1992.

[5] Noga Alon, Gregory Gutin, and Michael Krivelevich. Algorithms with large domination ratio. *J. Algorithms*, 50(1):118–131, 2004.

[6] R. Armoni, M. Saks, A. Wigderson, and S. Zhou. Discrepancy sets and pseudorandom generators for combinatorial rectangles. In *Proc. 37th IEEE Symp. on Foundations of Comp. Science (FOCS)*, pages 412–421, 1996.

[7] A. D. Barbour and V. Cekanavičius. Total variation asymptotics for sums of independent integer random variables. *The Annals of Probability*, 30(2):509–545, 2002.

[8] Andrew D. Barbour and Aihua Xia. Poisson perturbations. *ESAIM*, 3:131–150, October 1999.

[9] Alexander Barvinok and Ellen Veomett. The computational complexity of convex bodies. In Jacob E. Goodman, János Pach, and Richard Pollack, editors, *Surveys on Discrete and Computational Geometry: Twenty Years Later*, volume 453 of *Contemporary Mathematics*, pages 117–137. AMS, 2008.

[10] Louay M. J. Bazzi. Polylogarithmic independence can fool DNF formulas. *SIAM J. Comput.*, 38(6):2220–2272, 2009.

[11] Robert Beals, Harry Buhrman, Richard Cleve, Michele Mosca, and Ronald de Wolf. Quantum lower bounds by polynomials. *J. ACM*, 48(4):778–797, 2001.

[12] Matthias Beck and Sinai Robins. *Computing the Continuous Discretely: Integer-point Enumeration in Polyhedra*. Undergraduate Texts in Mathematics. Springer, 1st edition, 2007.

[13] Richard Beigel. The polynomial method in circuit complexity. In *Proc. 8th Annual Structure in Complexity Theory Conference*, pages 82–95, 1993.

[14] Ido Ben-Eliezer, Shachar Lovett, and Ariel Yadin. Polynomial threshold functions: Structure, approximation and pseudorandomness, 2009.

[15] Michael Ben-Or and Nathan Linial. Collective coin flipping, robust voting schemes and minima of Banzhaf values. In *Proc. 26th IEEE Symp. on Foundations of Comp. Science (FOCS)*, pages 408–416. IEEE, 1985.

[16] Itai Benjamini, Gil Kalai, and Oded Schramm. Noise sensitivity of Boolean functions and applications to percolation. *Inst. Hautes Études Sci. Publ. Math.*, 90(1):5–43, 1999.

[17] Vidmantas K. Bentkus. Smooth approximations of the norm and differentiable functions with bounded support in Banach space l_∞^k. *Lithuanian Mathematical Journal*, 30(3):223–230, July 1990.

[18] Vidmantas K. Bentkus. On the dependence of the Berryv-Esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385–402, May 2003.

[19] Eric Blais, Ryan O'Donnell, and Karl Wimmer. Polynomial regression under arbitrary product distributions. In *Proc. 21st Annual Conference on Learning Theory (COLT)*, pages 193–204, 2008.

[20] Andrej Bogdanov and Emanuele Viola. Pseudorandom bits for polynomials. In *Proc. 48th IEEE Symp. on Foundations of Comp. Science (FOCS)*, pages 41–51, 2007.

[21] Anthony Carbery and James Wright. Distributional and l^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical Research Letters*, 8(3):233–248, 2001.

[22] Larry Carter and Mark N. Wegman. Universal classes of hash functions. In *Proc. 9th ACM Symp. on Theory of Computing (STOC)*, pages 106–112, 1977.

[23] Mary Cryan and Martin E. Dyer. A polynomial-time algorithm to approximately count contingency tables when the number of rows is constant. *J. Computer and System Sciences*, 67(2):291–310, 2003.

[24] Constantinos Daskalakis and Christos H. Papadimitriou. Computing equilibria in anonymous games. In *Proc. 48th IEEE Symp. on Foundations of Comp. Science (FOCS)*, 2007.

[25] Constantinos Daskalakis and Christos H. Papadimitriou. Discretized multinomial distributions and nash equilibria in anonymous games. In *Proc. 49th IEEE Symp. on Foundations of Comp. Science (FOCS)*, 2008.

[26] Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco A. Servedio, and Emanuele Viola. Bounded independence fools halfspaces. *SIAM J. Computing*, 39(8):3441–3462, 2010.

[27] Ilias Diakonikolas, Prahladh Harsha, Adam Klivans, Raghu Meka, Prasad Raghavendra, Rocco A. Servedio, and Li-Yang Tan. Bounding the average sensitivity and noise sensitivity of polynomial threshold functions.

In *Proc. 42nd ACM Symp. on Theory of Computing (STOC)*, pages 533–542, 2010.

[28] Ilias Diakonikolas, Daniel Kane, and Jelani Nelson. Bounded independence fools degree-2 threshold functions. In *Proc. 51st IEEE Symp. on Foundations of Comp. Science (FOCS)*, pages 11–20, 2010.

[29] Ilias Diakonikolas, Prasad Raghavendra, Rocco Servedio, and Li-Yang Tan. Average sensitivity and noise sensitivity of polynomial threshold functions, 2009.

[30] Ilias Diakonikolas, Rocco A. Servedio, Li-Yang Tan, and Andrew Wan. A regularity lemma, and low-weight approximators, for low-degree polynomial threshold functions. In *IEEE Conference on Computational Complexity*, pages 211–222, 2010.

[31] Martin E. Dyer. Approximate counting by dynamic programming. In *Proc. 35th ACM Symp. on Theory of Computing (STOC)*, pages 693–699, 2003.

[32] G. Even, O. Goldreich, M. Luby, N. Nisan, and B. Veličković. Approximations of general independent distributions. In *Proc. 24th ACM Symp. on Theory of Computing (STOC)*, pages 10–16, 1992.

[33] A S Faĭnleĭb. A generalization of Esséen's inequality and its application in probabilistic number theory. *Mathematics of the USSR-Izvestiya*, 2(4):821, 1968.

[34] William Feller. *An Introduction to Probability Theory and Its Applications, Volume 2*. Wiley, 2 edition, January 1971.

[35] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145, 1995.

[36] Parikshit Gopalan, Raghu Meka, Omer Reingold, and David Zuckerman. Pseudorandom generators for combinatorial shapes. In *Proc. 43rd ACM Symp. on Theory of Computing (STOC)*, pages 253–262, 2011.

[37] Parikshit Gopalan, Ryan O'Donnell, Yi Wu, and David Zuckerman. Fooling functions of halfspaces under product distributions. In *IEEE Conference on Computational Complexity*, pages 223–234, 2010.

[38] Parikshit Gopalan and Jaikumar Radhakrishnan. Finding duplicates in a data stream. In *SODA*, pages 402–411, 2009.

[39] Craig Gotsman and Nathan Linial. Spectral properties of threshold functions. *Combinatorica*, 14(1):35–50, 1994.

[40] Iftach Haitner, Danny Harnik, and Omer Reingold. On the power of the randomized iterate. In *CRYPTO*, pages 22–40, 2006.

[41] Prahladh Harsha, Adam Klivans, and Raghu Meka. Bounding the sensitivity of polynomial threshold functions, 2009. arXiv: 0909.5175.

[42] Prahladh Harsha, Adam Klivans, and Raghu Meka. An invariance principle for polytopes. In *Proc. 42nd ACM Symp. on Theory of Computing (STOC)*, pages 543–552, 2010.

[43] Johan Hastad. On the size of weights for threshold gates. *SIAM J. Discret. Math.*, 7(3):484–492, 1994.

[44] Johan Håstad. Some optimal inapproximability results. *J. ACM*, 48(4):798–859, July 2001.

[45] David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.*, 100(1):78–150, 1992.

[46] Russell Impagliazzo, Noam Nisan, and Avi Wigderson. Pseudorandomness for network algorithms. In *Proc. 26th ACM Symp. on Theory of Computing (STOC)*, pages 356–364, 1994.

[47] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM*, 53(3):307–323, 2006.

[48] Mark Jerrum and Alistair Sinclair. The Markov chain Monte Carlo method: An approach to approximate counting and integration. In Dorit S. Hochbaum, editor, *Approximation Algorithms for NP-hard Problems*. PWS Publishing Company, 1997.

[49] Jeff Kahn, Gil Kalai, and Nathan Linial. The influence of variables on boolean functions (extended abstract). In *Proc. 29th IEEE Symp. on Foundations of Comp. Science (FOCS)*, pages 68–80, 1988.

[50] Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. *SIAM J. Computing*, 37(6):1777–1805, 2008.

[51] Gil Kalai. Noise sensitivity and chaos in social choice theory. Technical Report 399, Center for Rationality and Interactive Decision Theory, Hebrew University of Jerusalem, 2005.

[52] Daniel M. Kane. The gaussian surface area and noise sensitivity of degree-d polynomial threshold functions. In *IEEE Conference on Computational Complexity*, pages 205–210, 2010.

[53] Daniel M. Kane. k-independent gaussians fool polynomial threshold functions. In *IEEE Conference on Computational Complexity*, pages 252–262, 2011.

[54] Eyal Kaplan, Moni Naor, and Omer Reingold. Derandomized constructions of k-wise (almost) independent permutations. In *Proceedings of the 8th International Workshop on Randomization and Computation (RANDOM '05)*, number 3624 in Lecture Notes in Computer Science, pages 354 – 365, Berkeley, CA, August 2005. Springer.

[55] Zohar Shay Karnin, Yuval Rabani, and Amir Shpilka. Explicit dimension reduction and its applications. In *IEEE Conference on Computational Complexity*, pages 262–273, 2011.

[56] Michael J. Kearns, Robert E. Schapire, and Linda Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2–3):115–141, 1994.

[57] Subhash Khot, Guy Kindler, Elchanan Mossel, and Ryan O'Donnell. Optimal inapproximability results for max-cut and other 2-variable csps? *SIAM J. Comput.*, 37(1):319–357, 2007.

[58] L. B. Klebanov and S. T. Mkrtchyan. Estimation of the closeness of distributions in terms of identical moments. *Journal of Mathematical Sciences*, 32:54–60, 1986. 10.1007/BF01084500.

[59] Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning intersections and thresholds of halfspaces. *J. Comput. Syst. Sci.*, 68(4):808–840, 2004.

[60] Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning geometric concepts via gaussian surface area. In *Proc. 49th IEEE Symp. on Foundations of Comp. Science (FOCS)*, pages 541–550. IEEE, 2008.

[61] Adam R. Klivans and Rocco A. Servedio. Learning DNF in time $2^{O(n^{1/3})}$. *J. Computer and System Sciences*, 68(2):303–318, 2004.

[62] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*. Springer Verlag, 1991.

[63] P. M. Lewis and C. L. Coates. *Threshold Logic*. John Wiley, New York, 1967.

[64] J. W. Lindeberg. Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15:211–225, 1922.

[65] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, fourier transform, and learnability. *J. ACM*, 40(3):607–620, 1993.

[66] Shachar Lovett. Unconditional pseudorandom generators for low degree polynomials. In *Proc. 40th ACM Symp. on Theory of Computing (STOC)*, pages 557–562, 2008.

[67] Shachar Lovett, Omer Reingold, Luca Trevisan, and Salil P. Vadhan. Pseudorandom bit generators that fool modular sums. In *APPROX-RANDOM*, pages 615–630, 2009.

[68] Chi-Jen Lu. Improved pseudorandom generators for combinatorial rectangles. *Combinatorica*, 22:417–434, 2002.

[69] Wolfgang Maass and György Turán. How fast can a threshold gate learn? In *Proceedings of a workshop on Computational learning theory and natural learning systems (vol. 1) : constraints and prospects*, pages 381–414, Cambridge, MA, USA, 1994. MIT Press.

[70] Sanjeev Mahajan and Ramesh Hariharan. Derandomizing approximation algorithms based on semidefinite programming. *SIAM J. Computing*, 28(5):1641–1663, 1999.

[71] J. Matousek. *Geometric Discrepancy*. Springer, 1999.

[72] R. Meka and D. Zuckerman. Small-bias spaces over finite groups. In *APPROX-RANDOM*, 2009.

[73] Raghu Meka and David Zuckerman. Small-bias spaces for group products. In *APPROX-RANDOM*, pages 658–672, 2009.

[74] Raghu Meka and David Zuckerman. Pseudorandom generators for polynomial threshold functions. In *Proc. 42nd ACM Symp. on Theory of Computing (STOC)*, pages 427–436, 2010.

[75] Elchanan Mossel. Gaussian bounds for noise correlation of functions and tight analysis of long codes. In *Proc. 49th IEEE Symp. on Foundations of Comp. Science (FOCS)*, pages 156–165, 2008.

[76] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences invariance and optimality. *Annals of Mathematics*, 171(1):295–341, 2010.

[77] J. Naor and M. Naor. Small–bias probability spaces: Efficient constructions and applications. *SIAM J. Computing*, 22(4):838–856, 1993.

[78] Joseph Naor and Moni Naor. Small-bias probability spaces: Efficient constructions and applications. *SIAM J. Computing*, 22(4):838–856, August 1993.

[79] Fedor Nazarov. On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In *Geometric Aspects of Functional Analysis (Israel Seminar 2001–2002)*, volume 1807/2003 of *Lecture Notes in Mathematics*, pages 169–187. Springer, 2003.

[80] Noam Nisan. Pseudorandom generators for space-bounded computation. *Combinatorica*, 12(4):449–461, 1992.

[81] Noam Nisan and David Zuckerman. Randomness is linear in space. *J. Comput. Syst. Sci.*, 52(1):43–52, 1996.

[82] Ryan O'Donnell. Hardness amplification within NP. *J. Computer and System Sciences*, 69(1):68–94, 2004.

[83] Ryan O'Donnell and Rocco A. Servedio. The Chow parameters problem. In *Proc. 40th ACM Symp. on Theory of Computing (STOC)*, pages 517–526, 2008.

[84] Vygantas Paulauskas and Alfredas Račkauskas. *Approximation Theory in the Central Limit Theorem: Exact Results in Banach Spaces*. Kluwer Academic Publishers, 1989. (Translated from Russian).

[85] Yuval Peres. Noise stability of weighted majority, 2004. arXiv: math/0412377.

[86] Iosif Pinelis. Extremal probabilistic problems and hotellings T^2 test under a symmetry condition. *Ann. Statist.*, 22(1):357–368, 1994.

[87] Yuval Rabani and Amir Shpilka. Explicit construction of a small epsilon-net for linear threshold functions. *SIAM J. Computing*, 39(8):3501–3520, 2010.

[88] Prasad Raghavendra. Optimal algorithms and inapproximability results for every csp? In *Proc. 40th ACM Symp. on Theory of Computing (STOC)*, pages 245–254, 2008.

[89] V. Roychowdhury, K. Y. Siu, A. Orlitsky, and T. Kailath. A geometric approach to threshold circuit complexity. In *COLT '91: Proceedings of the fourth annual workshop on Computational learning theory*, pages 97–111, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc.

[90] Eyal Rozenman and Salil P. Vadhan. Derandomized squaring of graphs. In *APPROX-RANDOM*, pages 436–447, 2005.

[91] Rocco A. Servedio. Every linear threshold function has a low-weight approximator. In *IEEE Conference on Computational Complexity*, pages 18–32, 2006.

[92] Rocco A. Servedio. Every linear threshold function has a low-weight approximator. *Computational Complexity*, 16(2):180–209, 2007. (Pre-

liminary version in *21st IEEE Conference on Computational Complexity*, 2006).

[93] I. G. Shevtsova. Sharpening of the upper bound of the absolute constant in the berry–esseen inequality. *Theory of Probability and its Applications*, 51(3):549–553, 2007.

[94] Yaoyun Shi. Lower bounds of quantum black-box complexity and degree of approximating polynomials by influence of Boolean variables. *Inf. Process. Lett.*, 75(1-2):79–83, 2000.

[95] D. Sivakumar. Algorithmic derandomization via complexity theory. In *Proc. 32nd ACM Symp. on Theory of Computing (STOC)*, pages 619–626, 2002.

[96] Terry Tao. Talagrand's concentration inequality, 2009. (Post in Blog "What's new").

[97] Emanuele Viola. The sum of d small-bias generators fools polynomials of degree d. In *IEEE Conference on Computational Complexity*, pages 124–127, 2008.

[98] Thomas Watson. Pseudorandom generators for combinatorial checkerboards. In *IEEE Conference on Computational Complexity*, 2011. To Appear.

[99] Pawel Wolff. Hypercontractivity of simple random variables. *Studia Math*, 180(3):219–236, 2007.

[100] Günter M. Ziegler. *Lectures on polytopes*, volume 152 of *Graduate texts in Mathematics*. Springer, 1995.

Vita

Raghu Meka was born in Hyderabad, India, the son of Latha Meka and Sudhakar Meka. After completing high school in Hyderabad, he joined the Computer Science department at Indian Institute of Technology, Madras in 2001. He received a Bachelor of Technology degree from Indian Institute of Technology, Madras in 2005. He entered the graduate program at University of Texas at Austin in September 2005.

Permanent address: 5 Dartmouth Drive
Northborough, MA 01532
raghuvardhan@gmail.com

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.