Fooling Functions of Halfspaces under Product Distributions

Parikshit Gopalan
Microsoft Research SVC
parik@microsoft.com

Ryan O'Donnell*
Carnegie Mellon University
odonnell@cs.cmu.edu

Yi Wu†
Carnegie Mellon University
yiwu@cs.cmu.edu

David Zuckerman‡
UT Austin
diz@cs.utexas.edu

December 15, 2009

Abstract

We construct pseudorandom generators that fool functions of halfspaces (threshold functions) under a very broad class of product distributions. This class includes not only familiar cases such as the uniform distribution on the discrete cube, the uniform distribution on the solid cube, and the multivariate Gaussian distribution, but also includes any product of discrete distributions with probabilities bounded away from 0.

Our first main result shows that a recent pseudorandom generator construction of Meka and Zuckerman [MZ09], when suitably modified, can fool arbitrary functions of d halfspaces under product distributions where each coordinate has bounded fourth moment. To ϵ-fool any size-s, depth-d decision tree of halfspaces, our pseudorandom generator uses seed length $O((d\log(ds/\epsilon)+\log n)\cdot\log(ds/\epsilon))$. For monotone functions of d halfspaces, the seed length can be improved to $O((d\log(d/\epsilon)+\log n)\cdot\log(d/\epsilon))$. We get better bounds for larger ϵ; for example, to $1/\mathrm{polylog}(n)$-fool all monotone functions of $(\log n)/\log\log n$ halfspaces, our generator requires a seed of length just $O(\log n)$.

Our second main result generalizes the work of Diakonikolas et al. [DGJ+09] to show that bounded independence suffices to fool functions of halfspaces under product distributions. Assuming each coordinate satisfies a certain stronger moment condition, we show that any function computable by a size-s, depth-d decision tree of halfspaces is ϵ-fooled by $\tilde{O}(d^4s^2/\epsilon^2)$-wise independence.

Our technical contributions include: a new multidimensional version of the classical Berry-Esseen theorem; a derandomization thereof; a generalization of Servedio [Ser07]'s regularity lemma for halfspaces which works under any product distribution with bounded fourth moments; an extension of this regularity lemma to functions of many halfspaces; and, new analysis of the sandwiching polynomials technique of Bazzi [Baz09] for arbitrary product distributions.

*Work was partially done while the author consulted at Microsoft Research SVC. Supported by NSF grants CCF-0747250 and CCF-0915893, BSF grant 2008477, and Sloan and Okawa fellowships.

†Work done while an intern at Microsoft Research SVC.

‡Work was partially done while the author consulted at Microsoft Research SVC. Partially supported by NSF Grants CCF-0634811 and CCF-0916160 and THECB ARP Grant 003658-0113-2007.

1 Introduction

Halfspaces, or threshold functions, are a central class of Boolean-valued functions. A halfspace is a function $h : \mathbb{R}^n \to \{0,1\}$ of the form $h(x_1,\ldots,x_n) = \mathbf{1}[w_1x_1 + \cdots + w_nx_n \geq \theta]$ where the weights $w_1,\ldots,w_n$ and the threshold θ are arbitrary real numbers. These functions have been studied extensively in theoretical computer science, social choice theory, and machine learning. In computer science, they were first studied in the context of switching circuits; see for instance [Der65, Hu65, LC67, She69, Mur71]. Halfspaces (with non-negative weights) have also been studied extensively in game theory and social choice theory as models for voting; see e.g. [Pen46, Isb69, DS79, TZ92]. Halfspaces are also ubiquitous in machine learning contexts, playing a key role in many important algorithmic techniques, such as Perceptron , Support Vector Machine, Neural Networks, and AdaBoost. One of the outstanding open problems in circuit lower bounds is to find an explicit function that cannot be computed by a depth two circuit ("neural network") of threshold gates [HMP$^+$93, Kra91, KW91, FKL$^+$01].

In this work we investigate the problem of constructing explicit *pseudorandom generators* for functions of halfspaces.

Definition 1.1. *A function $G : \{0,1\}^s \to B$ is a pseudorandom generator (PRG) with seed length s and error ϵ for a class $\mathcal{F}$ of functions from B to $\{0,1\}$ under distribution $\mathcal{D}$ on B — or more succinctly, G ϵ-fools $\mathcal{F}$ under $\mathcal{D}$ with seed length s — if for all $f \in \mathcal{F}$,*

$$\left| \mathop{\mathbf{Pr}}_{\boldsymbol{X}\sim\mathcal{D}}[f(\boldsymbol{X}) = 1] - \mathop{\mathbf{Pr}}_{\boldsymbol{Y}\sim\{0,1\}^s}[f(G(\boldsymbol{Y})) = 1] \right| \leq \epsilon.$$

Under the widely-believed complexity-theoretic assumption $\mathsf{BPP} = \mathsf{P}$, there must be a deterministic algorithm that can approximate the fraction of satisfying assignments to any polynomial-size circuit of threshold gates. Finding such an algorithm even for simple functions of halfspaces has proven to be a difficult derandomization problem. Very recently, however, there has been a burst of progress on constructing PRGs for halfspaces [RS08, DGJ$^+$09, MZ09]. The present paper makes progress on this problem in several different directions, as do several concurrent and independent works [HKM09, DKN09, BELY09].

This flurry of work on PRGs for functions of halfspaces has several motivations beyond its status as a fundamental derandomization task. For one, it can be seen as a natural geometric problem, with connections to deterministic integration; for instance, the problem of constructing PRGs for halfspaces under the uniform distribution on the n-dimensional sphere amounts to constructing a poly(n)-sized set that hits every spherical cap with roughly the right frequency [RS08]. Second, PRGs for halfspaces have applications in streaming algorithms [GR09], while PRGs for functions of halfspaces can be used to derandomize the Goemans-Williamson Max-Cut algorithm, algorithms for approximate counting, algorithms for dimension reduction and intractability results in computational learning [KS08]. Finally, proving lower bounds for the class TC^0 of small depth threshold circuits is an outstanding open problem in circuit complexity. An explicit PRG for a class is easily seen to imply lower bounds against that class. Constructions of explicit PRGs might shed light on structural properties of threshold circuits and the lower bound problem.

1.1 Previous Work

The work of Rabani and Shpilka [RS08] constructed a hitting set generator for halfspaces under the uniform distribution on the sphere. Diakonikolas et al. [DGJ$^+$09] constructed the first PRG for halfspaces over bits; i.e., the uniform distribution on $\{-1,1\}^n$. They showed that any k-wise

independent distribution ϵ-fools halfspaces with respect to the uniform distribution for $k = \tilde{O}(1/\epsilon^2)$, giving PRGs with seed length $(\log n) \cdot \tilde{O}(1/\epsilon^2)$.

Meka and Zuckerman constructed a pseudorandom generator that ϵ-fools degree-d polynomial threshold functions ("PTFs", a generalization of halfspaces) over uniformly random bits with seed length $(\log n)/\epsilon^{O(d)}$ [MZ09]. Their generator is a simplified version of Rabani and Shpilka's hitting set generator. In the case of halfspaces, they combine their generator with generators for small-width branching programs due to Nisan and Nisan-Zuckerman [Nis92, NZ96] to bring the seed length down to $O((\log n)\log(1/\epsilon))$. This is the only previous or independent work where the seed length depends logarithmically on $1/\epsilon$.

1.2 Independent Concurrent Work

Independently and concurrently, a number of other researchers have extended some of the aforementioned results, mostly to intersections of halfspaces and polynomial threshold functions over the hypercube or Gaussian space.

Diakonikolas et al. [DKN09] showed that $O(1/\epsilon^9)$-wise independence suffices to fool degree-2 PTFs under the uniform distribution on the hypercube and under the Gaussian distribution. They also prove that $\mathrm{poly}(d, 1/\epsilon)$-wise independence suffices to fool intersections of d degree-2 PTFs in these settings.

Harsha et al. [HKM09] obtain a PRG that fools intersections of d halfspaces under the Gaussian distribution with seed length $O((\log n) \cdot \mathrm{poly}(\log d, 1/\epsilon))$. They obtain similar parameters for intersections of d "regular" halfspaces under the uniform distribution on $\{-1, 1\}^n$ (a halfspace is regular if all of its coefficients have small magnitude compared to their sum of squares).

Ben-Eliezer et al. [BELY09] showed that roughly $\exp((d/\epsilon)^d)$-wise independence ϵ-fools degree-d PTFs which depend on a small number of linear functions.

1.3 Our Results

In this work, we construct pseudorandom generators for arbitrary functions of halfspaces under (almost) arbitrary product distributions. Our work diverges from previous work in making minimal assumptions about the distribution we are interested in, and in allowing general functions of halfspaces. For both of our main results, we only assume that the distribution is a product distribution where each coordinate satisfies some mild conditions on its moments. These conditions include most distributions of interest, such as the Gaussian distribution, the uniform distribution on the hypercube, the uniform distribution on the solid cube, and discrete distributions with probabilities bounded away from 0. Our results can also be used to fool the uniform distribution on the sphere, even though it is not a product distribution. This allows us to derandomize the hardness result of Khot and Saket [KS08] for learning intersections of halfspaces.

We also allow for arbitrary functions of d halfspaces, although the seed length improves significantly if we consider monotone functions or small decision trees. In particular, we get strong results for intersections of halfspaces.

1.3.1 The Meka-Zuckerman Generator

We show that a suitable modification of the Meka-Zuckerman (MZ) generator can fool arbitrary functions of d halfspaces under any product distribution, where the distribution on each coordinate has bounded fourth moments. More precisely, we consider product distributions on $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ where for every $i \in [n]$, $\mathbf{E}[\boldsymbol{x}_i] = 0, \mathbf{E}[\boldsymbol{x}_i^2] = 1$, $\mathbf{E}[\boldsymbol{x}_i^4] \leq C$ where $C \geq 1$ is a parameter of the generator G. We say that the distribution $\boldsymbol{X}$ has C-bounded fourth moments.

We get our best results for monotone functions of d halfspaces, such as intersections of d halfspaces. For distributions with polynomially bounded fourth moments, our modified MZ PRG fools the intersection of d halfspaces with polynomially small error using a seed of length $O(d \log^2 n)$. Many natural distributions have $O(1)$-bounded fourth moments. Even for polylog(n)-bounded fourth moments, our PRG fools the intersection of $(\log n)/\log\log n$ halfspaces with error $1/\text{polylog}(n)$ using a seed of length just $O(\log n)$. Both of these cases are captured in the following theorem.

Theorem 1.2. *Let $\boldsymbol{X}$ be sampled from a product distribution on $\mathbb{R}^n$ with C-bounded fourth moments. The modified MZ generator ϵ-fools any monotone function of d halfspaces with seed length $O((d\log(Cd/\epsilon) + \log n)\log(Cd/\epsilon))$. When $Cd/\epsilon \geq \log^{-c} n$ for any $c > 0$, the seed length becomes $O(d\log(Cd/\epsilon) + \log n)$.*

As a corollary, we get small seed length for functions of halfspaces that have small decision tree complexity. In the theorem below we could even take s to be the minimum of the number of 0-leaves and 1-leaves.

Theorem 1.3. *Let $\boldsymbol{X}$ be as in Theorem 1.2. The modified MZ generator ϵ-fools any size-s, depth-d function of halfspaces, using a seed of length $O((d\log(Cds/\epsilon) + \log n)\log(Cds/\epsilon))$. When $Cds/\epsilon \geq \log^{-c} n$ for any $c > 0$, the seed length becomes $O(d\log(Cds/\epsilon) + \log n)$.*

Since the decision tree complexity is at most 2^d, we deduce the following.

Corollary 1.4. *Let $\boldsymbol{X}$ be as in theorem 1.2. The modified MZ generator ϵ-fools any function of d halfspaces, using a seed of length $O((d^2 + d\log(Cd/\epsilon) + \log n)(d + \log(Cd/\epsilon)))$. When $Cd2^d/\epsilon \geq \log^{-c} n$ for any $c > 0$, the seed length becomes $O(d^2 + d\log(Cd/\epsilon) + \log n)$.*

1.3.2 Bounded Independence Fools Functions of Halfspaces

We prove that under a large class of product distributions, bounded independence suffices to fool functions of d halfspaces. This significantly generalizes the result of Diakonikolas et al. [DGJ$^+$09] who proved that bounded independence fools halfspaces under the uniform distribution on $\{-1, 1\}^n$. The condition necessary on the product distributions is unfortunately somewhat technical; we state here a theorem that covers the main cases of interest:

Theorem 1.5. *Suppose f is computable as a size-s, depth-d function of halfspaces over the independent random variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. If we assume the $\boldsymbol{x}_j$'s are discrete, then k-wise independence suffices to ϵ-fool f, where*

$$k = \widetilde{O}(d^4 s^2/\epsilon^2) \cdot \text{poly}(1/\alpha).$$

Here $0 < \alpha \leq 1$ is the least nonzero probability of any outcome for an $\boldsymbol{x}_j$. Moreover, the same result holds with $\alpha = 1$ for certain continuous random variables $\boldsymbol{x}_j$, including Gaussians (possibly of different variance) and random variables which are uniform on (possibly different) intervals.

For example, whenever $\alpha \geq 1/\text{polylog}(d/\epsilon)$ it holds that $\widetilde{O}(d^6/\epsilon^2)$-wise independence suffices to ϵ-fool intersections of m halfspaces. For random variables that do not satisfy the hypotheses of Theorem 1.5, it may still be possible to extract a similar statement from our techniques. Roughly speaking, the essential requirement is that the random variables $\boldsymbol{x}_j$ be "$(p, 2, p^{-c})$-hypercontractive" for large values of p and some constant $c < 1$.

Notation: Throughout, all random variables take values in $\mathbb{R}$ or $\mathbb{R}^d$. Random variables will be in boldface. Real scalars will be lower-case letters; real vectors will be upper-case letters. If X is a d-dimensional vector, we will write $X[1], X[2], \ldots, X[d]$ for its coordinates values and $\|X\| = \sqrt{\sum_{i=1}^{d} X[i]^2}$ for its Euclidean length. When C is a matrix, we also use the notation $C[i,j]$ for its (i,j) entry. If $\boldsymbol{X}$ is a vector-valued random variable, we write $\|\boldsymbol{X}\|_p = \mathbf{E}[\|\boldsymbol{X}\|^p]^{1/p}$. We typically use i to index dimensions and j to index sequences. Given $x \in \mathbb{R}$ we define $\mathrm{sgn}(x) = 1$ if $x \geq 0$ and $\mathrm{sgn}(x) = -1$ if $x < 0$. If X is a d-dimensional vector, then $\overrightarrow{\mathrm{sgn}}(X)$ denotes the vector in $\{-1,1\}^d$ with $\overrightarrow{\mathrm{sgn}}(X)[i] = \mathrm{sgn}(X[i])$.

Our results concern arbitrary functions of d halfspaces. Thus we have vectors $W_1, \ldots, W_n, \Theta \in \mathbb{R}^d$, and we're interested in functions $f : \{-1,1\}^d \to \{0,1\}$ of the vector $\overrightarrow{\mathrm{sgn}}(x_1 W_1 + \ldots + x_n W_n - \Theta)$, which we abbreviate to $\overrightarrow{\mathrm{sgn}}(W \cdot X - \Theta)$ where $W = (W_1, \ldots, W_n)$ and $X = (x_1, \ldots, x_n)$.

Organization: We give an overview of our results and their proofs in 2. We state the multi-dimensional Berry-Esseen type theorems in Section 4. In Section 5, we prove a regularity lemma for multiple halfspaces in the general setting of hypercontractive variables. We state modified MZ generator in Section C, and analyze it using the machinery above in Section D. In Section E, we show how to combine it with PRGs for branching programs to get our Theorems 1.2 and 1.3. We prove Theorem 1.5 in Section F. In Section G, we show how our results apply to fooling the uniform distribution on the sphere, and use it to derandomize the hardness result of [KS08].

2 Overview of the Main Results

2.1 The Meka-Zuckerman Generator

There are five steps in the analysis:

1. Discretize the distribution $\boldsymbol{X}$ so that it is the product of discrete distributions whose moments nearly match those of $\boldsymbol{X}$.
2. Prove a multidimensional version of the classical Berry-Esseen theorem, and a derandomization thereof. This allows us to handle functions of regular halfspaces. See Subsection 2.1.1.
3. Generalize the regularity lemma/critical index lemma (see [Ser07, DGJ$^+$09]) to d halfspaces. This gives a small set of variables such that after conditioning on these variables, each halfspace becomes either regular or close to a constant function. See Subsection 2.1.2.
4. Use the regularity lemma to reduce analyzing functions of d arbitrary halfspaces to analyzing functions of d (or fewer) regular halfspaces.
5. Finally, generalize the monotone trick from [MZ09], which previously worked only for a single "monotone" branching program, to monotone functions of monotone branching programs. This enables us to get seed length logarithmic in $1/\epsilon$. See Subsection 2.1.3.

2.1.1 Multi-Dimensional Berry-Esseen Theorem

The classic Berry-Esseen Theorem is a quantitative version of the Central Limit Theorem. This theorem is essential in the analyses of [MZ09] and [DGJ$^+$09] for one halfspace. Since we seek to fool functions of several halfspaces, we prove a multi-dimensional version of the Berry-Esseen theorem, which approximates the distribution of $\sum_i \boldsymbol{x}_i W_i$. The error of the approximation is small when all the halfspaces are regular (no coefficient is too large). While there are multi-dimensional versions known, we were unable to find in the literature any theorems which we could use in a "black-box" fashion. The reason for this is twofold: known results tend to focus on measuring the

difference between probability distributions vis-a-vis convex sets; whereas, we are interested in more specialized sets, unions of orthants. Second, results in the literature tend to assume a nonsingular covariance matrix and/or have a dependence in the error bound on its least eigenvalue; whereas, we need to work with potentially singular covariance matrices. We believe this theorem could be of independent interest.

Next we show how this theorem can be derandomized in a certain sense. This derandomization enables us to show that our modified MZ PRG fools *regular* halfspaces.

2.1.2 Multi-Dimensional Critical Index

The concept of critical index was introduced in the work of Servedio [Ser07]. It is used to prove a regularity lemma for halfspaces, which asserts that every halfspace contains a *head* consisting of constantly many variables, such that once these variables are set randomly, the resulting function is either close to constant, or close to a regular halfspace. This lemma has found numerous applications in complexity and learning theoretic questions related to halfspaces [Ser07, OS08, FGRW09, DGJ$^+$09, MZ09].

The obvious generalization of the one-dimensional theorem to multiple halfspaces would be to take the union of the heads of each halfspace. This does not work, since setting variables in a regular halfspace can make it irregular. We prove a multidimensional version of this lemma, which moreover holds in the setting of product distributions with bounded fourth moments. Our analysis shows that the lemma only requires some basic concentration and anti-concentration properties, which are enjoyed by any random variable with bounded fourth moments.

2.1.3 Monotone Branching Programs

The only known method to get logarithmic dependence on $1/\epsilon$ for PRGs for halfspaces, due to Meka and Zuckerman, considers the natural branching program accepting a halfspace. This branching program is "monotone," in the sense that in every layer the set of accepting suffixes forms a total order under inclusion. Meka and Zuckerman showed that any monotone branching program of arbitrary width can be sandwiched between two small-width monotone branching programs. Therefore, PRGs for small-width branching programs, such as those by Nisan [Nis92] can be used.

Since we deal with several halfspaces, we get several monotone branching programs. We consider monotone functions of monotone branching programs, to encompass intersections of halfspaces. However, such functions are not necessarily computable by monotone branching programs. Nevertheless, we show how to sandwich such functions between two small-width branching programs, and thus can use the PRGs like Nisan's.

2.2 Bounded Independence Fools Functions of Halfspaces

2.2.1 Sandwiching "polynomials"

To prove Theorem 1.5 we use the "sandwiching polynomials" method as introduced by Bazzi [Baz09] and used by [DGJ$^+$09]. However in our setting of general random variables it is not appropriate to use polynomials per se. The essence of the sandwiching polynomial method is showing that only groups of d random variables need to be "simultaneously controlled'. When the random variables are ± 1-valued, controlling sub-functions of at most d random variables is equivalent to controlling polynomials of degree at most d. But for random variables with more than two outcomes, a function of d random variables requires degree higher than d in general, a price we should not be forced to pay. We instead introduce the following notions:

Definition 2.1. *Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ be a product set. We say that $p : \Omega \to \mathbb{R}$ is a* k-junta *if $f(x_1, \ldots, x_n)$ depends on at most k of the x_j's. We say that p is a* generalized polynomial of order (at most) k *if it is expressible as a sum of simple functions of order at most k. In the remainder of this section we typically drop the word "generalized" from "generalized polynomial", and add the modifier "ordinary" when referring to "ordinary polynomials".*

We now give the simple connection to fooling functions with bounded independence:

Definition 2.2. *Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be a vector of independent random variables, where $\boldsymbol{x}_j$ has range Ω_j. Let $f : \Omega \to \mathbb{R}$, where $\Omega = \Omega_1 \times \cdots \times \Omega_n$. We say that polynomials $p_l, p_u : \Omega \to \mathbb{R}$ are* ϵ-sandwiching *for f if*

$$p_l(X) \leq f(X) \leq p_u(X) \text{ for all } X \in \Omega, \text{ and } \mathbf{E}[p_u(\boldsymbol{X})] - \epsilon \leq \mathbf{E}[f(\boldsymbol{X})] \leq \mathbf{E}[p_l(\boldsymbol{X})] + \epsilon.$$

Proposition 2.3. *Suppose p_l, p_u are ϵ-sandwiching for f as in Definition 2.2 and have order at most k. Then f is ϵ-fooled by k-wise independence. I.e., if $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ is a vector of random variables such that each marginal of the form $(\boldsymbol{y}_{j_1}, \ldots, \boldsymbol{y}_{j_k})$ matches the corresponding marginal $(\boldsymbol{x}_{j_1}, \ldots, \boldsymbol{x}_{j_k})$, then*

$$|\mathbf{E}[f(\boldsymbol{X})] - \mathbf{E}[f(\boldsymbol{Y})]| \leq \epsilon.$$

Proof. Write $p_u = \sum_t q_t$, where each q_t is a k-junta. Then

$$\mathbf{E}[f(\boldsymbol{Y})] \leq \mathbf{E}[p_u(\boldsymbol{Y})] = \mathbf{E}[\sum_t q_t(\boldsymbol{Y})] = \sum_t \mathbf{E}[q_t(\boldsymbol{Y})] = \sum_t \mathbf{E}[q_t(\boldsymbol{X})] = \mathbf{E}[p_u(\boldsymbol{X})] \leq \mathbf{E}[f(\boldsymbol{X})] + \epsilon,$$

where in addition to the sandwiching properties of p_u we used the fact that q_t is a k-junta to deduce $\mathbf{E}[q_t(\boldsymbol{Y})] = \mathbf{E}[q_t(\boldsymbol{X})]$. We obtain the bound $\mathbf{E}[f(\boldsymbol{Y})] \geq \mathbf{E}[f(bX)] - \epsilon$ similarly, using p_l. □

2.2.2 Upper polynomials for intersections suffice

We begin with a trivial observation:

Proposition 2.4. *Let $\mathcal{C}$ be a class of functions $\Omega \to \{0, 1\}$, and suppose that for every $f \in \mathcal{C}$ we have just the "upper sandwiching polynomial", p_u, of an ϵ-sandwiching pair for f. Then if $\mathcal{C}$ is closed under Boolean negation, we obtain a matching "lower polynomial" p_l of the same order as p_u automatically.*

This is simply because given p_u for f, we may take $p_l = 1 - p_u$. Since the Boolean negation of a halfspace is a halfspace, this observation could have been used for slight simplification in [DGJ+09].

Our Theorem 1.5 is concerned with the class of 0-1 functions f computable as size-s, depth-d functions of halfspaces. This class is closed under Boolean negation; hence it suffices for us to obtain upper sandwiching polynomials. Furthermore, every such f can be written as $f = \sum_{t=1}^{s'} H_t$, where $s' \leq s$ and H_t is an intersection (AND) of up to d halfspaces. To see this, simply sum the indicator function for each root-to-leaf path in the decision tree (this again uses the fact that the negation of a halfspace is a halfspace). Thus if we have (ϵ/s)-sandwiching upper polynomials of order k for each H_t, by summing them we obtain an ϵ-sandwiching upper polynomial for f of the same order. Hence to prove our main Theorem 1.5, it suffices to prove the following:

Theorem 2.5. *Suppose f is the intersection of d halfspaces $h_1, \ldots, h_d$ over the independent random variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Suppose α is as in Theorem 1.5. Then there exists an ϵ-sandwiching upper polynomial for f of order $k \leq \widetilde{O}(d^4/\epsilon^2) \cdot \mathrm{poly}(1/\alpha)$.*

2.2.3 Polynomial construction techniques

Suppose for simplicity we are only concerned with the intersection f of d halfspaces $h_1, \ldots, h_d$ over uniform random ± 1 bits $\boldsymbol{x}_j$. The work of Diakonikolas et al. [DGJ$^+$09] implies that there are is an ϵ_0-sandwiching upper polynomial p_i of order $\widetilde{O}(1/\epsilon_0^2)$ for each h_i. To obtain an ϵ-sandwiching upper polynomial for the intersection $h_1 h_2 \cdots h_d$, a natural first idea is simply to try $p = p_1 p_2 \cdots p_d$. This is certainly an upper-bounding polynomial; however the ϵ-sandwiching aspect is unclear. We can begin the analysis as follows. Let $\boldsymbol{h}_i = h_i(\boldsymbol{X})$ and $\boldsymbol{p}_i = p_i(\boldsymbol{X})$. By telescoping,

$$\begin{aligned} \mathbf{E}[\boldsymbol{p}_1 \cdots \boldsymbol{p}_d] - \mathbf{E}[\boldsymbol{h}_1 \cdots \boldsymbol{h}_d] &= \mathbf{E}[(\boldsymbol{p}_1 - \boldsymbol{h}_1)\boldsymbol{p}_2 \cdots \boldsymbol{p}_d] \quad + \quad \cdots \\ &\ldots \quad + \quad \mathbf{E}[\boldsymbol{h}_1 \cdots \boldsymbol{h}_{i-1}(\boldsymbol{p}_i - \boldsymbol{h}_i)\boldsymbol{p}_{i+1} \cdots \boldsymbol{p}_d] \quad + \quad \cdots \\ &\ldots \quad + \quad \mathbf{E}[\boldsymbol{h}_1 \cdots \boldsymbol{h}_{d-1}(\boldsymbol{p}_d - \boldsymbol{h}_d)]. \end{aligned} \tag{1}$$

Now the last term here could be upper-bounded as

$$\mathbf{E}[\boldsymbol{h}_1 \cdots \boldsymbol{h}_{d-1}(\boldsymbol{p}_d - \boldsymbol{h}_d)] \leq \mathbf{E}[\boldsymbol{p}_d - \boldsymbol{h}_d] \leq \epsilon_0,$$

since each $0 \leq \boldsymbol{h}_i \leq 1$ with probability 1. But we cannot make an analogous bound for the remaining terms because we have no a priori control over the values of the $\boldsymbol{p}_i$'s beyond the individual sandwiching inequalities

$$\mathbf{E}[\boldsymbol{p}_i - \boldsymbol{h}_i] \leq \epsilon_0.$$

Nevertheless, we will be able to make this strategy work by establishing *additional boundedness conditions* on the polynomials p_i; specifically, that each $\boldsymbol{p}_i$ exceeds $1 + 1/d^2$ extremely rarely, and that even the high $2d$-norm of $\boldsymbol{p}_i$ is not much more than 1.

Establishing these extra properties requires significant reworking the construction in [DGJ$^+$09]. Even in the case of uniform random ± 1 bits, the calculations are not straightforward, since the upper sandwiching polynomials implied by [DGJ$^+$09] are only fully explicit in the case of regular halfspaces. And to handle general random variables $\boldsymbol{x}_j$, we need more than just our new Regularity Lemma 5.3 for halfspaces. We also need to assume a stronger hypercontractivity property of the random variables to ensure they have rapidly decaying tails.

3 Hypercontractivity

The notion of of *hypercontractive random variables* was introduced in [KS88] and developed by Krakowiak, Kwapień, and Szulga:

Definition 3.1. *We say that a real random variable $\boldsymbol{x}$ is* (p, q, η)-hypercontractive *for $1 \leq q \leq p < \infty$ and $0 < \eta < 1$ if $\|\boldsymbol{x}\|_p < \infty$, and for all $a \in \mathbb{R}$, $\|a + \eta\boldsymbol{x}\|_p \leq \|a + \boldsymbol{x}\|_q$.*

In this paper we will be almost exclusively concerned with the simplest case, $p = 4$, $q = 2$. Let us abbreviate the definition in this case (and also exclude constantly-0 random variables):

Definition 3.2. *A real random variable $\boldsymbol{x}$ is* η-HC *for $0 < \eta < 1$ if $0 < \|\boldsymbol{x}\|_4 < \infty$ and for all $a \in \mathbb{R}$, $\|a + \eta\boldsymbol{x}\|_4 \leq \|a + \boldsymbol{x}\|_2$, i.e. $\mathbf{E}[(a + \eta\boldsymbol{x})^4] \leq \mathbf{E}[(a + \boldsymbol{x})^2]^2$.*

Essentially, a mean 0 real random variable is η-HC with large η if and only if it has a small 4th moment (compared to its 2nd moment). Random variables with small 4th moment are known to enjoy some basic concentration and anti-concentration properties. We work with hypercontractivity

rather than 4^{th} moments because it tends to slightly shorten proofs and improve constants; the main convenience is that a linear combination of η-HC random variables is also η-HC.

In Appendix I we list some basic properties of η-HC random variables which we will use. The notion of hypercontractivity can be extended to $\mathbb{R}^d$-valued random variables:

Definition 3.3. *An $\mathbb{R}^d$-random variable $\boldsymbol{X}$ is* η-HC *for $0 < \eta < 1$ if $\|\boldsymbol{X}\|_4 < \infty$ and for all $A \in \mathbb{R}^d$, $\|A + \eta\boldsymbol{X}\|_4 \leq \|A + \boldsymbol{X}\|_2$.*

We require the following facts about vector-valued hypercontractivity:

Fact 3.4. *[Szu90]*

1. *If $W \in \mathbb{R}^d$ is a fixed vector and $\boldsymbol{x}$ is an η-HC real random variable, then $\boldsymbol{X} = \boldsymbol{x}W$ is an η-HC.*
2. *If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent η-HC random vectors, then so is $c_1\boldsymbol{X}_1 + \cdots c_n\boldsymbol{X}_n$ for any real constants $c_1, \ldots, c_n$. (Again, 4-wise independence also suffices.)*

Hypercontractive real random variables possess the following good concentration and anti-concentration properties. The proofs are in Appendix I.

Proposition 3.5. *If $\boldsymbol{x}$ is η-HC then for all $t > 0$, $\mathbf{Pr}[|\boldsymbol{x}| \geq t\|\boldsymbol{x}\|_2] \leq \frac{1}{\eta^4 t^4}$.*

Proposition 3.6. *If $\boldsymbol{x}$ is η-HC then for all $\theta \in \mathbb{R}$ and $0 < t < 1$, $\mathbf{Pr}[|\boldsymbol{x} - \theta| > t\|\boldsymbol{x}\|_2] \geq \eta^4(1-t^2)^2$.*

4 The Multi-Dimensional Berry-Esseen Theorem

In this section we state a Berry-Esseen-style result in the setting of multidimensional random variables, and a derandomization of it. The proofs are deferred to Appendix A.

We assume the following setup: $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent $\mathbb{R}^d$-valued η-HC random variables, not necessarily identically distributed, satisfying $\mathbf{E}[\boldsymbol{X}_j] = 0$ for all $j \in [n]$. We let $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$. We write $C_j = \mathbf{Cov}[\boldsymbol{X}_j] \in \mathbb{R}^{d \times d}$ for the covariance matrix of $\boldsymbol{X}_j$, which is positive semidefinite. We also write $C = \mathbf{Cov}[\boldsymbol{S}]$ for the covariance matrix of $\boldsymbol{S}$; by the independence and mean-zero assumptions we have $C = C_1 + \cdots + C_n$. We will also assume that

$$C[i,i] = \sum_{j=1}^{n} \mathbf{E}\left[\boldsymbol{X}_j[i]^2\right] = 1 \qquad \text{for all } i \in [d].$$

If we write $\sigma_j^2 = \|\boldsymbol{X}_j\|^2$, it follows that $\sum_{j=1}^n \sigma_j^2 = d$. We introduce new independent random variables $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_n$, where $\boldsymbol{G}_j$ is a d-dimensional Gaussian random variable with covariance matrix C_j; we also write also $\boldsymbol{G} = \boldsymbol{G}_1 + \cdots + \boldsymbol{G}_n$. We say that $A \subseteq \mathbb{R}^d$ is a translate of a union of orthants if there exists some vector $\Theta \in \mathbb{R}^d$ such that $X \in A$ depends only on $\overrightarrow{\mathrm{sgn}}(X - \Theta)$.

Theorem 4.1. *Let $\boldsymbol{S}$ and $\boldsymbol{G}$ be as above. Let $A \subseteq \mathbb{R}^d$ be a translate of a union of orthants. Then*

$$|\mathbf{Pr}[\boldsymbol{S} \in A] - \mathbf{Pr}[\boldsymbol{G} \in A]| \leq O(\eta^{-1/2} d^{13/8}) \cdot \Big(\sum_{j=1}^{n} \sigma_j^4\Big)^{1/8}.$$

We now show that this result can be "derandomized" using the output of the MZ generator $\boldsymbol{Y}$ in place of $\boldsymbol{X}$. We describe here a simplified version of the output of their generator.

Definition 4.2. *A family $\mathcal{H} = \{h : [n] \to [t]\}$ of hash functions is* C-collision preserving *if*

1. *For all $i \in [n], \ell \in [t]$, $\mathbf{Pr}_{h \in_u \mathcal{H}}[h(i) = \ell] \leq C/t$.*

2. For all $i \neq j \in [n]$, $\mathbf{Pr}_{h \in_u \mathcal{H}}[h(i) = h(j)] \leq C/t$.

We choose a partition $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_t$ of $[n]$ into t buckets using a C-collision preserving family of hash functions (where $C \leq 2$). The vector of variables $\{\boldsymbol{Y}_j\}_{j \in \boldsymbol{H}_\ell}$ is generated 4-wise independently. There is full independence across different buckets. Let $\boldsymbol{T} = \boldsymbol{Y}_1 + \cdots + \boldsymbol{Y}_n$.

Theorem 4.3. *Let $\boldsymbol{T}$ and $\boldsymbol{G}$ be as above. Let $A \subseteq \mathbb{R}^d$ be a translate of a union of orthants. Then*

$$|\mathbf{Pr}[\boldsymbol{T} \in A] - \mathbf{Pr}[\boldsymbol{G} \in A]| \leq O(\eta^{-1/2} d^{13/8}) \cdot \Big(\frac{d^2}{t} + \sum_{j=1}^{n} \sigma_j^4\Big)^{1/8}.$$

Putting these two theorems together, we have shown the following statement

Theorem 4.4. *Let $\boldsymbol{S}$ and $\boldsymbol{T}$ be as above. Let $A \subseteq \mathbb{R}^d$ be a translate of a union of orthants. Then*

$$|\mathbf{Pr}[\boldsymbol{S} \in A] - \mathbf{Pr}[\boldsymbol{T} \in A]| \leq O(\eta^{-1/2} d^{13/8}) \cdot \Big(\frac{d^2}{t} + \sum_{j=1}^{n} \sigma_j^4\Big)^{1/8}.$$

5 Critical Index for Hypercontractive Random Variables

In this section we will consider η-HC random variables $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n$ which are at least pairwise independent. Write $\sigma_j^2 = \|\boldsymbol{x}_j\|_2^2$, and note that pairwise independence implies $\|\boldsymbol{x}_0 + \cdots + \boldsymbol{x}_n\|_2^2 = \sigma_0^2 + \cdots + \sigma_n^2$. We also write $\tau_i^2 = \|\boldsymbol{x}_i + \boldsymbol{x}_{i+1} + \cdots + \boldsymbol{x}_n\|_2^2 = \sum_{j \geq i} \sigma_j^2$.

Definition 5.1. *For $0 < \delta < 1$, we say that the collection of random variables $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n$ is* δ-regular *if $\sum_{j=0}^n \|\boldsymbol{x}_j\|_4^4 \leq \delta\Big(\sum_{j=0}^n \|\boldsymbol{x}_j\|_2^2\Big)^2 = \delta\tau_0^4$.*

Definition 5.2. *Suppose the sequence $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n$ is* ordered*, meaning that $\sigma_0^2 \geq \sigma_1^2 \geq \sigma_2^2 \geq \cdots$. Then for $0 < \delta < 1$, the* δ-critical index *is defined to be the smallest index ℓ such that the sequence $\boldsymbol{x}_\ell, \boldsymbol{x}_{\ell+1}, \ldots, \boldsymbol{x}_n$ is δ-regular, or $\ell = \infty$ no such index exists.*

Theorem 5.3. *Let $0 < \delta < 1$, $0 < \epsilon < 1/2$, and $s > 1$ be parameters. Let $L = br$, where $b = \lceil (2/\eta^4) \ln(1/\epsilon) \rceil$ and $r = \lceil (1/\eta^4 \delta) \ln(1 + 16s^2) \rceil$; note that*

$$L \leq O\left(\frac{\log(s)\log(1/\epsilon)}{\eta^8}\right) \cdot \frac{1}{\delta}.$$

Assume the sequence $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n$ is ordered, that $n \geq L$, and that $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{L-1}$ are independent. Then if ℓ is the δ-critical index for the sequence, and $\ell \geq L$, then for all $\theta \in \mathbb{R}$,

$$\mathbf{Pr}\left[|\boldsymbol{x}_0 + \cdots + \boldsymbol{x}_{L-1} - \theta| \leq s \cdot \tau_L\right] \leq \epsilon + \frac{O(\ln(1/\epsilon))}{\eta^8 s^4}.$$

Proof. For any $0 \leq j < L$, since the critical index ℓ is at least j we have

$$\delta\tau_j^4 < \sum_{i \geq j} \|\boldsymbol{x}_i\|_4^4 \leq (1/\eta^4) \sum_{i \geq j} \sigma_i^4 (\text{since each } \boldsymbol{x}_i \text{ is } \eta\text{-HC}) \leq (\sigma_j^2/\eta^4) \sum_{i \geq j} \sigma_i^2 = (\sigma_j^2/\eta^4)\tau_j^2.$$

where we used hypercontractivity and the fact that σ_is are ordered. Hence for all $0 \leq j < L$,

$$\eta^4 \delta \tau_j^2 < \sigma_j^2 = \tau_j^2 - \tau_{j+1}^2 \quad \Rightarrow \quad \tau_{j+1}^2 < (1 - \eta^4\delta)\tau_j^2.$$

It follows that for all $0 \le k < b$,

$$\tau^2_{(k+1)r} < (1-\eta^4\delta)^r \tau^2_{kr} < \frac{1}{1+16s^2}\tau^2_{kr}, \tag{2}$$

where we used the definition of r.

Now for each $0 \le k < b$ define $\boldsymbol{y}_k = \boldsymbol{x}_{kr} + \boldsymbol{x}_{kr+1} + \boldsymbol{x}_{kr+2} + \cdots + \boldsymbol{x}_{(k+1)r-1}$ and $\upsilon_k^2 = \|\boldsymbol{y}_k\|_2^2 = \tau^2_{kr} - \tau^2_{(k+1)r}$. Using (2) we have immediately conclude

$$\upsilon_k^2 > 16s^2\tau^2_{(k+1)r} \qquad \Rightarrow \qquad \upsilon_k > 4s\tau_{(k+1)r}. \tag{3}$$

Since all of $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{L-1}$ are independent and η-HC, we have that $\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{b-1}$ are independent η-HC random variables. For $0 \le k < b$, define the event $A_k =$ "$|\boldsymbol{y}_0 + \boldsymbol{y}_1 + \cdots + \boldsymbol{y}_k - \theta| \le (1/2)\upsilon_k$",. We claim that for any $0 \le k < b$,

$$\mathbf{Pr}[A_k \mid A_0 \wedge A_1 \wedge \cdots \wedge A_{k-1}] < 1 - \eta^4/2.$$

To see this, note that conditioning only affects the values of random variables $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{k-1}$, of which $\boldsymbol{y}_k$ is independent. Further, for *every* choice of values for $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{k-1}$, the event A_k is an anti-concentration event of the type in Proposition 3.6, with some shifted θ. Hence the claim follows from this Proposition, as $(1-(1/2)^2)^2 > 1/2$. Having established the claim, we conclude

$$\mathbf{Pr}[A_0 \wedge A_1 \wedge \cdots \wedge A_{b-1}] < (1-\eta^4/2)^b \le \epsilon. \tag{4}$$

Let us now define, for each $1 \le k < b$, random variables $\boldsymbol{z}_k = \boldsymbol{y}_k + \boldsymbol{y}_{k+1} + \cdots + \boldsymbol{y}_{b-1}$. These random variables are also η-HC, and they satisfy $\|\boldsymbol{z}_k\|_2^2 \le \tau^2_{kr}$. If we define the events $B_k =$ "$|\boldsymbol{z}_k| \ge s\tau_{kr}$", then Proposition 3.5 implies $\mathbf{Pr}[B_k] \le 1/\eta^4 s^4$. Hence

$$\mathbf{Pr}[B_1 \vee B_2 \vee \cdots \vee B_{b-1}] \le (b-1)/\eta^4 s^4 < b/\eta^4 s^4. \tag{5}$$

Combining (4) and (5) we see that except with probability less than $\epsilon + b/\eta^4 s^4 \le \epsilon + \frac{O(\ln(1/\epsilon))}{\eta^8 s^4}$, at least one event $\overline{A_k}$ occurs, and none of the events B_k occurs. Since this is the error bound in the Theorem, it remains to show that in this case, the desired result "$|\boldsymbol{x}_0 + \cdots + \boldsymbol{x}_{L-1} - \theta| > s \cdot \tau_L$" occurs. Assume then that $\overline{A_m}$ occurs and B_{m+1} does not occur, $0 \le m < b$. (For $m = b-1$ we need not make the latter assumption.) Thus

$$|\boldsymbol{y}_0 + \boldsymbol{y}_1 + \cdots + \boldsymbol{y}_m - \theta| > (1/2)\upsilon_m \quad \text{and} \quad |\boldsymbol{z}_{m+1}| \le s\tau_{(m+1)r} < (1/4)\upsilon_m,$$

where we used (3). (This makes sense also in the case $m = b-1$ if we naturally define $\boldsymbol{z}_b \equiv 0$.) By definition of $\boldsymbol{z}_{m+1}$, we therefore obtain

$$|\boldsymbol{y}_0 + \boldsymbol{y}_1 + \cdots + \boldsymbol{y}_{b-1} - \theta| = |\boldsymbol{x}_0 + \cdots + \boldsymbol{x}_{L-1} - \theta| > (1/4)\upsilon_m \ge (1/4)\upsilon_{b-1} \ge s\tau_{br} = s\tau_L,$$

as desired, where we used (3). □

We state the high-dimensional generalization of Theorem 5.3, the proof is in Appendix B. Assume $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are η-HC real random variables which are at least pairwise independent. Assume also that $W_1, \ldots, W_n$ are arbitrary fixed vectors in $\mathbb{R}^d$, and write $\boldsymbol{X}_j = \boldsymbol{x}_j W_j$.

Theorem 5.4. *Let δ, ϵ, s, L be as in Theorem 5.3. Then there exists a set of coordinates $H_0 \subseteq [n]$, $|H_0| \le dL$, with the following property. Assuming the collection of random variables $\{\boldsymbol{x}_j : j \in H_0\}$ is independent, for each coordinate $i \in [d]$ we have either:*

1. the sequence of real random variables $\{\boldsymbol{X}_j[i] : j \notin H_0\}$ is δ-regular; or,
2. for all $\theta \in \mathbb{R}$,

$$\mathbf{Pr}\left[\left|\sum_{j\in H_0} \boldsymbol{X}_j[i] - \theta\right| \leq s \cdot \sqrt{\sum_{j\notin H_0} \|\boldsymbol{X}_j[i]\|_2^2}\right] \leq \epsilon + \frac{O(\ln(1/\epsilon))}{\eta^8 s^4}.$$

References

[Baz09] L. Bazzi. Polylogarithmic independence can fool DNF formulas. *SIAM Journal on Computing*, 38:2220–2272, 2009.

[BELY09] Ido Ben-Eliezer, Shachar Lovett, and Ariel Yadin. Polynomial threshold functions: Structure, approximation and pseudorandomness. In *Submitted*, 2009.

[Ben04] Vidmantas Bentkus. A Lyapunov type bound in $\mathbb{R}^d$. *Theory of Probability and its Applications*, 49(2):311–322, 2004.

[Der65] M. Dertouzos. *Threshold logic: a synthesis approach.* MIT Press, Cambridge, MA, 1965.

[DGJ+09] Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco A. Servedio, and Emanuele Viola. Bounded independence fools halfspaces. In *Proceedings of the 50th IEEE Symposium on Foundations of Computer Science*, 2009.

[DKN09] I. Diakonikolas, D. Kane, and J. Nelson. Bounded independence fools degree-2 threshold functions. In *Submitted*, 2009.

[DS79] P. Dubey and L.S. Shapley. Mathematical properties of the banzhaf power index. *Mathematics of Operations Research*, 4:99–131, 1979.

[FGRW09] V. Feldman, V. Guruswami, P. Raghavendra, and Y. Wu. Agnostic learning of monomials by halfspaces is hard. In *FOCS*, 2009.

[FKL+01] J. Forster, M. Krause, S.V. Lokam, R. Mubarakzjanov, N. Schmitt, and H.-U. Simon. Relations between communication complexity, linear arrangements, and computational complexity. In *FSTTCS*, pages 171–182, 2001.

[GR09] P. Gopalan and J. Radhakrishnan. Finding duplicates in a data stream. In *Proc. 20th Annual Symposium on Discrete Algorithms (SODA'09)*, pages 402–411, 2009.

[HKM09] Prahladh Harsha, Adam Klivans, and Raghu Meka. An invariance principle for polytopes. In *Submitted*, 2009.

[HMP+93] A. Hajnal, W. Maass, P. Pudlak, M. Szegedy, and G. Turan. Threshold circuits of bounded depth. *Journal of Computer and System Sciences*, 46:129–154, 1993.

[Hu65] S.T. Hu. *Threshold Logic.* University of California Press, 1965.

[INW94] Russell Impagliazzo, Noam Nisan, and Avi Wigderson. Pseudorandomness for network algorithms. In *STOC*, pages 356–364, 1994.

[Isb69] J.R. Isbell. A Counterexample in Weighted Majority Games. *Proceedings of the AMS*, 20(2):590–592, 1969.

[Kra91] M. Krause. Geometric arguments yield better bounds for threshold circuits and distributed computing. In *Proc. 6th Structure in Complexity Theory Conference*, pages 314–322, 1991.

[KS88] Wiesław Krakowiak and Jerzy Szulga. Hypercontraction principle and random multilinear forms. *Probability Theory and Related Fields*, 77(3):325–342, 1988.

[KS08] S. Khot and R. Saket. On hardness of learning intersection of two halfspaces. In *STOC*, 2008.

[KW91] M. Krause and S. Waack. Variation ranks of communication matrices and lower bounds for depth two circuits having symmetric gates with unbounded fanin. In *Proc. 32nd IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 777–782, 1991.

[LC67] P.M. Lewis and C.L. Coates. *Threshold Logic*. New York, Wiley, 1967.

[MOO05] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. In *Proceedings of the 46th IEEE Symposium on Foundations of Computer Science*, pages 21–30, 2005. To appear, *Annals of Mathematics* 2010.

[Mos08] Elchanan Mossel. Gaussian bounds for noise correlation of functions and tight analysis of long codes. In *Proceedings of the 49th IEEE Symposium on Foundations of Computer Science*, pages 156–165, 2008.

[Mur71] S. Muroga. *Threshold logic and its applications*. Wiley-Interscience, New York, 1971.

[MZ09] Raghu Meka and David Zuckerman. Pseudorandom generators for polynomial threshold functions, 2009. arXiv:0910.4122 [cs.CC].

[Nis92] Noam Nisan. Pseudorandom generators for space-bounded computation. *Combinatorica*, 12(4):449–461, 1992.

[NZ96] Noam Nisan and David Zuckerman. Randomness is linear in space. *Journal of Computer and System Sciences*, 52(1):43–52, 1996.

[OS08] R. O'Donnell and R. Servedio. The Chow Parameters Problem. In *STOC*, pages 517–526, 2008.

[Pen46] L.S. Penrose. The elementary statistics of majority voting. *Journal of the Royal Statistical Society*, 109(1):53–57, 1946.

[RS08] Y. Rabani and A. Shpilka. Explicit construction of a small epsilon-net for linear threshold functions. In *STOC*, 2008.

[Ser07] R. Servedio. Every linear threshold function has a low-weight approximator. *Computational Complexity*, 16(2):180–209, 2007.

[She69] Q. Sheng. *Threshold Logic*. London, New York, Academic Press, 1969.

[Szu90] Jerzy Szulga. A note on hypercontractivity of stable random variables. *The Annals of Probability*, 18(4):1746–1758, 1990.

[TZ92] A. Taylor and W. Zwicker. A Characterization of Weighted Voting. *Proceedings of the AMS*, 115(4):1089–1094, 1992.

[Wol06a] Paweł Wolff. Hypercontractivity of random variables and geometry of linear normed spaces, 2006. Unpublished.

[Wol06b] Paweł Wolff. Hypercontractivity of simple random variables. *Studia Mathematica*, 180(3):219–236, 2006.

A Proof of the Multi-Dimensional Berry-Esseen Theorem

In this section we prove a Berry-Esseen-style result in the setting of multidimensional random variables. Our aim is not to get the best bounds possible (for which one might pursue the methods of Bentkus [Ben04]). Rather, we aim to provide a simple method which achieves a reasonable bound, and thus use the Lindeberg method, following [MOO05, Mos08] very closely.

A.1 Setup

We assume the following setup: $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent $\mathbb{R}^d$-valued random variables, not necessarily identically distributed, satisfying $\mathbf{E}[\boldsymbol{X}_j] = 0$ for all $j \in [n]$ and possessing fourth moments (only third moments are needed for the first part of the argument). We let $\boldsymbol{S} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$. We write $C_j = \mathbf{Cov}[\boldsymbol{X}_j] \in \mathbb{R}^{d \times d}$ for the covariance matrix of $\boldsymbol{X}_j$, which is positive semidefinite. We also write $C = \mathbf{Cov}[\boldsymbol{S}]$ for the covariance matrix of $\boldsymbol{S}$; by the independence and mean-zero assumptions we have $C = C_1 + \cdots + C_n$. We will also assume that

$$C[i,i] = \sum_{j=1}^{n} \mathbf{E}\left[\boldsymbol{X}_j[i]^2\right] = 1 \qquad \text{for all } i \in [d].$$

If we write $\sigma_j^2 = \|\boldsymbol{X}_j\|^2$, it follows that $\sum_{j=1}^n \sigma_j^2 = d$. Finally, we introduce new independent random variables $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_n$, where $\boldsymbol{G}_j$ is a d-dimensional Gaussian random variable with covariance matrix C_j; we also write also $\boldsymbol{G} = \boldsymbol{G}_1 + \cdots + \boldsymbol{G}_n$.

A.2 The basic lemma

In what follows, K will denote a d-dimensional multi-index $(k_1, \ldots, k_d) \in \mathbb{N}^d$, with $|K|$ denoting $j_1 + \cdots + j_d$ and $K!$ denoting $k_1!k_2! \cdots k_d!$. Given a vector $H \in \mathbb{R}^d$, the expression H^K denotes $\prod_{i=1}^d H[i]^{k_i}$. Given a function $\psi : \mathbb{R}^d \to \mathbb{R}$, the expression $\psi^{(K)}$ denotes the mixed partial derivative taken k_i times in the ith coordinate; we will always assume ψ is smooth enough that the order of the derivatives does not matter.

The following lemma is essentially proven in, e.g., [Mos08, Theorem 4.1]. To obtain it, simply repeat Mossel's proof in the degree 1 case, until equation (31). (Although Mossel assumes that the covariance matrices C_j are identity matrices, this is not actually necessary; it suffices that $\mathbf{Cov}[\boldsymbol{X}_j] = \mathbf{Cov}[\boldsymbol{G}_j]$.) Then instead of using hypercontractivity, skip directly to summing the error terms over all coordinates.

Lemma A.1. *Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be a $\mathcal{C}^3$ function with $\left|\psi^{(K)}\right| \le b$ for all $|K| = 3$. Then*

$$|\mathbf{E}[\psi(\boldsymbol{S})] - \mathbf{E}[\psi(\boldsymbol{G})]| \le b \sum_{|K|=3} \frac{1}{K!} \sum_{j=1}^{n} \left(\mathbf{E}\left[\left|\boldsymbol{X}_j^K\right|\right] + \mathbf{E}\left[\left|\boldsymbol{G}_j^K\right|\right] \right). \tag{6}$$

We further deduce:

Corollary A.2. *In the setting of Lemma A.1,*

$$|\mathbf{E}[\psi(\boldsymbol{S})] - \mathbf{E}[\psi(\boldsymbol{G})]| \le 2bd^3 \sum_{j=1}^{n} \|\boldsymbol{X}_j\|_3^3.$$

Proof. Fix a multi-index K with $|K| = 3$ and also an index j. We will show that

$$\mathbf{E}\left[\left|\boldsymbol{X}_j^K\right|\right] + \mathbf{E}\left[\left|\boldsymbol{G}_j^K\right|\right] \le 2.6 \|\boldsymbol{X}_j\|_3^3. \tag{7}$$

Substituting this into (6) completes the proof, since

$$b \sum_{|K|=3} \frac{2.6}{K!} \le 2bd^3.$$

Let the nonzero coordinates in K be $i_1, i_2, i_3 \in [d]$, written with multiplicity. Write also

$$\sigma_i^2 = C_j[i,i] = \mathbf{E}\left[\boldsymbol{G}_j[i]^2\right] = \mathbf{E}\left[\boldsymbol{X}_j[i]^2\right].$$

On one hand, by Hölder we have

$$\mathbf{E}\left[\left|\boldsymbol{G}_j^K\right|\right] = \mathbf{E}\left[|\boldsymbol{G}_j[i_1]\boldsymbol{G}_j[i_2]\boldsymbol{G}_j[i_3]|\right] \le \sqrt[3]{\mathbf{E}\left[\left|\boldsymbol{G}_j[i_1]\right|^3\right] \mathbf{E}\left[\left|\boldsymbol{G}_j[i_2]\right|^3\right] \mathbf{E}\left[\left|\boldsymbol{G}_j[i_3]\right|^3\right]}.$$

Note that the distribution of $\boldsymbol{G}_j[i_1]$ is $N(0, \sigma_{i_1}^2)$. It is elementary that such a random variable has third absolute moment equal to $2\sqrt{2/\pi} \cdot \sigma_{i_1}^3 \le 2.6\sigma_{i_1}^3$. As the same is true for i_2 and i_3, we conclude that

$$\mathbf{E}\left[\left|\boldsymbol{G}_j^K\right|\right] \le 1.6\sigma_{i_1}\sigma_{i_2}\sigma_{i_3}. \tag{8}$$

On the other hand, we can similarly upper-bound

$$\mathbf{E}\left[\left|\boldsymbol{X}_j^K\right|\right] \le \sqrt[3]{\mathbf{E}\left[\left|\boldsymbol{X}_j[i_1]\right|^3\right] \mathbf{E}\left[\left|\boldsymbol{X}_j[i_2]\right|^3\right] \mathbf{E}\left[\left|\boldsymbol{X}_j[i_3]\right|^3\right]} \tag{9}$$

But

$$\sqrt[3]{\mathbf{E}\left[\left|\boldsymbol{X}_j[i_1]\right|^3\right] \mathbf{E}\left[\left|\boldsymbol{X}_j[i_2]\right|^3\right] \mathbf{E}\left[\left|\boldsymbol{X}_j[i_3]\right|^3\right]} \ge \sqrt[3]{\mathbf{E}\left[\left|\boldsymbol{X}_j[i_1]\right|^2\right]^{3/2} \mathbf{E}\left[\left|\boldsymbol{X}_j[i_2]\right|^2\right]^{3/2} \mathbf{E}\left[\left|\boldsymbol{X}_j[i_3]\right|^2\right]^{3/2}} = \sigma_{i_1}\sigma_{i_2}\sigma_{i_3},$$

and hence from (8) and (9) we conclude

$$\mathbf{E}\left[\left|\boldsymbol{X}_j^K\right|\right] + \mathbf{E}\left[\left|\boldsymbol{G}_j^K\right|\right] \le 2.6 \sqrt[3]{\mathbf{E}\left[\left|\boldsymbol{X}_j[i_1]\right|^3\right] \mathbf{E}\left[\left|\boldsymbol{X}_j[i_2]\right|^3\right] \mathbf{E}\left[\left|\boldsymbol{X}_j[i_3]\right|^3\right]}.$$

Finally, we clearly have $|\boldsymbol{X}_j[i_1]| \le \|\boldsymbol{X}_j\|$ always, and similarly for j_2, j_3. Hence

$$\mathbf{E}\left[\left|\boldsymbol{X}_j^K\right|\right] + \mathbf{E}\left[\left|\boldsymbol{G}_j^K\right|\right] \le 2.6 \sqrt[3]{\mathbf{E}\left[\|\boldsymbol{X}_j\|^3\right] \mathbf{E}\left[\|\boldsymbol{X}_j\|^3\right] \mathbf{E}\left[\|\boldsymbol{X}_j\|^3\right]} = 2.6 \|\boldsymbol{X}_j\|_3^3,$$

confirming (7). □

Corollary A.3. *In the setting of Lemma A.1,*

$$|\mathbf{E}[\psi(\boldsymbol{S})] - \mathbf{E}[\psi(\boldsymbol{G})]| \le 2bd^{7/2} \sqrt{\sum_{j=1}^{n} \|\boldsymbol{X}_j\|_4^4}.$$

Proof. Using Cauchy-Schwarz twice,

$$\sum_{j=1}^{n} \|\boldsymbol{X}_j\|_3^3 = \sum_{j=1}^{n} \mathbf{E}\left[\|\boldsymbol{X}_j\|^3\right] = \sum_{j=1}^{n} \mathbf{E}\left[\|\boldsymbol{X}_j\| \, \|\boldsymbol{X}_j\|^2\right] \le \sum_{j=1}^{n} \sqrt{\mathbf{E}\left[\|\boldsymbol{X}_j\|^2\right]} \sqrt{\mathbf{E}\left[\|\boldsymbol{X}_j\|^4\right]}$$
$$\le \sqrt{\sum_{j=1}^{n} \mathbf{E}\left[\|\boldsymbol{X}_j\|^2\right]} \sqrt{\sum_{j=1}^{n} \mathbf{E}\left[\|\boldsymbol{X}_j\|^4\right]} = \sqrt{d} \sqrt{\sum_{j=1}^{n} \|\boldsymbol{X}_j\|_4^4},$$

where we also used $\sum \sigma_j^2 = d$. □

A.3 Derandomization and hypercontractivity

We now show that this result can be "derandomized" in a certain sense. This idea is essentially due to Meka and Zuckerman [MZ09, Sec. 4.1].

Definition A.4. *We say that the sequences of $\mathbb{R}^d$-valued random vectors $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ and $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ satisfy the* r-matching-moments condition*, $r \in \mathbb{N}$, if the following holds: $\mathbf{E}[\boldsymbol{\mathcal{X}}^K] = \mathbf{E}[\boldsymbol{\mathcal{Y}}^K]$ for all multi-indices $|K| \le r$, where $\boldsymbol{\mathcal{X}}$ is the $\mathbb{R}^{dn}$-valued random vector gotten by concatenating $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$, and $\boldsymbol{\mathcal{Y}}$ is defined similarly.*

In this section, we suppose that $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ satisfy the 4-matching-moments condition with respect to $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$. We will *not* suppose that they are independent, but rather that they have some limited independence. Let $\boldsymbol{T} = \boldsymbol{Y}_1 + \cdots + \boldsymbol{Y}_n$.

Proposition A.5. *Let $H_1, \dots, H_t$ form a partition of $[n]$, and write $\boldsymbol{Z}_\ell = \sum_{j \in H_\ell} \boldsymbol{Y}_j$. Assume that $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_t$ are independent. Then*

$$|\mathbf{E}[\psi(\boldsymbol{T})] - \mathbf{E}[\psi(\boldsymbol{G})]| \le 2bd^{7/2} \sqrt{\sum_{\ell=1}^{t} \| \sum_{j \in H_\ell} \boldsymbol{X}_j \|_4^4}.$$

Proof. We simply apply Corollary A.3 to the random variables $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_t$. To check that it is applicable, we note the following: The random variables are independent. They satisfy $\mathbf{E}[\boldsymbol{Z}_\ell] = 0$, because each $\mathbf{E}[\boldsymbol{Y}_j] = 0$ by 1-matching-moments. The covariance matrix $\sum_{\ell=1}^{t} \mathbf{Cov}[\boldsymbol{Z}_\ell] = C$, by 2-matching-moments.

Thus Corollary A.3 gives

$$|\mathbf{E}[\psi(\boldsymbol{T})] - \mathbf{E}[\psi(\boldsymbol{G})]| \le 2bd^{7/2} \sqrt{\sum_{\ell=1}^{t} \|\boldsymbol{Z}_\ell\|_4^4}.$$

But for each ℓ,

$$\|\boldsymbol{Z}_\ell\|_4^4 = \| \sum_{j \in H_\ell} \boldsymbol{Y}_j \|_4^4 = \mathbf{E}\Big[\langle \sum_{j \in H_\ell} \boldsymbol{Y}_j, \sum_{j \in H_\ell} \boldsymbol{Y}_j \rangle^2\Big] = \mathbf{E}\Big[\langle \sum_{j \in H_\ell} \boldsymbol{X}_j, \sum_{j \in H_\ell} \boldsymbol{X}_j \rangle^2\Big] = \| \sum_{j \in H_\ell} \boldsymbol{X}_j \|_4^4,$$

using 4-matching-moments, completing the proof. □

Remark A.6. *The full 4-matching-moments condition is not essential for our results; it would suffice to have 2-matching-moments, along with a good upper bound on the 4th moments of the $\boldsymbol{Y}_j$'s with respect to those of the $\boldsymbol{X}_j$'s.*

We can simplify the previous bounds if we assume hypercontractivity.

Corollary A.7. *If we additionally assume that the random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are η-HC, then we have*

$$\begin{aligned} |\mathbf{E}[\psi(\boldsymbol{S})] - \mathbf{E}[\psi(\boldsymbol{G})]| &\leq (2bd^{7/2}/\eta^2)\sqrt{\sum_{j=1}^{n} \sigma_j^4}, \\ |\mathbf{E}[\psi(\boldsymbol{T})] - \mathbf{E}[\psi(\boldsymbol{G})]| &\leq (2bd^{7/2}/\eta^2)\sqrt{\sum_{\ell=1}^{t} \Big(\sum_{j \in H_\ell} \sigma_j^2 \Big)^2}. \end{aligned}$$

Proof. We prove only the second statement, the first being simpler. It suffices to show

$$\| \sum_{j \in H_\ell} \boldsymbol{X}_j \|_4^4 \leq (1/\eta)^4 \Big(\sum_{j \in H_\ell} \sigma_j^2 \Big)^2.$$

Since the random variables $\{\boldsymbol{X}_j : j \in H_\ell\}$ are independent and η-HC, it follows that the (vector-valued) random variable $\sum_{j \in H_\ell} \boldsymbol{X}_j$ is η-HC. Hence

$$\| \sum_{j \in H_\ell} \boldsymbol{X}_j \|_4^4 \leq (1/\eta)^4 \Big(\| \sum_{j \in H_\ell} \boldsymbol{X}_j \|_2^2 \Big)^2.$$

But

$$\| \sum_{j \in H_\ell} \boldsymbol{X}_j \|_2^2 = \sum_{j \in H_\ell} \sigma_j^2$$

by the Pythagorean Theorem. □

We now consider the case when the partition $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_t$ chosen randomly using a C-collison preserving family of hash functions (see Definition 4.2).

Proposition A.8. *In the setting of Corollary A.7, if the partition $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_t$ is chosen using a C-collision preserving family of hash functions, then*

$$|\mathbf{E}[\psi(\boldsymbol{T})] - \mathbf{E}[\psi(\boldsymbol{G})]| \leq (2bC^{1/2}d^{7/2}/\eta^2)\sqrt{\frac{d^2}{t} + \sum_{j=1}^{n} \sigma_j^4}.$$

where the expectation $\mathbf{E}[\psi(\boldsymbol{T})]$ is with respect to both the choice of $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_t$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$.

Proof. By the triangle inequality for real numbers, it suffices to show

$$\mathop{\mathbf{E}}_{\boldsymbol{H}_1, \ldots, \boldsymbol{H}_t} \left[\sqrt{\sum_{\ell=1}^{t} \Big(\sum_{j \in \boldsymbol{H}_\ell} \sigma_j^2 \Big)^2} \right] \leq \sqrt{C \left(\frac{d^2}{t} + \sum_{j=1}^{n} \sigma_j^4 \right)}.$$

By Cauchy-Schwarz, this reduces to showing

$$\mathop{\mathbf{E}}_{\boldsymbol{H}_1, \ldots, \boldsymbol{H}_t} \left[\sum_{\ell=1}^{t} \Big(\sum_{j \in \boldsymbol{H}_\ell} \sigma_j^2 \Big)^2 \right] \leq C \left(\frac{d^2}{t} + \sum_{j=1}^{n} \sigma_j^4 \right).$$

But

$$\mathop{\mathbf{E}}_{\boldsymbol{H}_1,\ldots,\boldsymbol{H}_t}\left[\sum_{\ell=1}^{t}\Big(\sum_{j\in\boldsymbol{H}_\ell}\sigma_j^2\Big)^2\right]=\sum_{\ell=1}^{t}\mathbf{E}\left[\Big(\sum_{j=1}^{n}\mathbf{1}_{\{j\in\boldsymbol{H}_\ell\}}\sigma_j^2\Big)^2\right]=\sum_{\ell=1}^{t}\sum_{j_1,j_2=1}^{n}\sigma_{j_1}^2\sigma_{j_2}^2\,\mathbf{E}[\mathbf{1}_{\{j_1\in\boldsymbol{H}_\ell\}}\mathbf{1}_{\{j_2\in\boldsymbol{H}_\ell\}}]$$

$$\leq\sum_{\ell=1}^{t}\left(\frac{C}{t}\sum_{j=1}^{n}\sigma_j^4\right)+\sum_{j_1\neq j_2}\sigma_{j_1}^2\sigma_{j_2}^2\sum_{\ell=1}^{t}\mathbf{E}[\mathbf{1}_{\{j_1\in\boldsymbol{H}_\ell\}}\mathbf{1}_{\{j_2\in\boldsymbol{H}_\ell\}}]\leq C\sum_{j=1}^{n}\sigma_j^4+\frac{C}{t}\sum_{j_1\neq j_2}\sigma_{j_1}^2\sigma_{j_2}^2\leq\frac{Cd^2}{t}+C\sum_{j=1}^{n}\sigma_j^4,$$

as needed, because

$$\sum_{j_1\neq j_2}\sigma_{j_1}^2\sigma_{j_2}^2\leq\left(\sum_{j=1}^{n}\sigma_j^2\right)^2=d^2.$$

□

A.4 Smoothing

Ideally we would like to use the results from the previous sections with ψ equal to certain indicator functions $\chi:\mathbb{R}^d\to\{0,1\}$; however these are not $\mathcal{C}^3$. As usual in the Lindeberg method (see, e.g., [MOO05]), we overcome this by working with mollified versions of these functions. For most of this section, we will work with our underandomized result, the statement about $\boldsymbol{S}$ in Corollary A.3. Identical considerations apply to the statement about $\boldsymbol{T}$ in Proposition A.8, and we will draw the necessary conclusions at the end.

Let $\xi:\mathbb{R}\to\mathbb{R}$ be the "standard mollifier", a smooth density function supported on $[-1,1]$. We will use the fact that there is some universal constant b_0 such that $\int\left|\xi^{(k)}\right|dx\leq b_0$ for $k=1,2,3$ (where $\xi^{(k)}$ denotes the kth derivative of ξ). Given $\epsilon>0$ we define $\xi_\epsilon(x)=\xi(x/\epsilon)/\epsilon$, the standard mollifier with support $[-\epsilon,\epsilon]$. Finally, define the density function Ξ_ϵ on $\mathbb{R}^d$ by $\Xi_\epsilon(x_1,\ldots,x_d)=\prod_{i=1}^{d}\xi_\epsilon(x_i)$. We now prove an elementary lemma:

Lemma A.9. *Let $\chi:\mathbb{R}^d\to[-1,1]$ be measurable, let $\epsilon>0$, and define $\psi=\Xi_\epsilon*\chi$, a smooth function. Then for any multi-index $|K|=3$ we have $\left|\psi^{(K)}\right|\leq(b_0/\epsilon)^3$.*

Proof. Using the fact that $|\chi|\leq1$ everywhere, we have

$$\left|\psi^{(K)}(a)\right|=\left|\Xi_\epsilon^{(K)}*\chi(a)\right|\leq\int\left|\Xi_\epsilon^{(K)}\right|=\int_{[-\epsilon,\epsilon]^d}\left|\prod_{i=1}^{d}\frac{\partial^{k_i}}{\partial x_i^{k_i}}\xi_\epsilon(x_i)\right|dx_1\cdots dx_d=\prod_{i=1}^{d}\int_{-\epsilon}^{\epsilon}\left|\frac{\partial^{k_i}}{\partial x^{k_i}}\xi_\epsilon(x)\right|dx.$$

Note that $\frac{\partial^k}{\partial x^k}\xi_\epsilon(x)=\xi^{(k)}(x/\epsilon)/\epsilon^{k+1}$, from which it follows that

$$\int_{-\epsilon}^{\epsilon}\left|\frac{\partial^k}{\partial x^k}\xi_\epsilon(x)\right|dx\leq b_0/\epsilon^k$$

for $k=1,2,3$. For $k=0$ we of course have

$$\int_{-\epsilon}^{\epsilon}|\xi_\epsilon(x)|\,dx=\int_{-\epsilon}^{\epsilon}\xi_\epsilon(x)\,dx=1.$$

Since $|K|=3$, we therefore achieve the claimed upper bound of $(b_0/\epsilon)^3$. □

Suppose now $A \subseteq \mathbb{R}^d$ is a measurable set. We define:

$$A^{+\epsilon} = \{x \in \mathbb{R}^d : x + [-\epsilon/2, \epsilon/2]^d \cap A \neq \emptyset\}, \quad A^{-\epsilon} = \{x \in \mathbb{R}^d : x + [-\epsilon/2, \epsilon/2]^d \subseteq A\}, \quad \partial^\epsilon A = A^{+\epsilon} \backslash A^{-\epsilon}.$$

We also define $\psi_{A^{+\epsilon}} = \Xi_\epsilon * \chi_{A^{+\epsilon}}$ as in Lemma A.9, where $\chi_{A^{+\epsilon}}$ is the 0-1 indicator of $A^{+\epsilon}$, and similarly define $\psi_{A^{-\epsilon}}$. Applying now Corollary A.3, we conclude:

Lemma A.10. *For $\psi = \psi_{A^{+\epsilon}}$ or $\psi = \psi_{A^{-\epsilon}}$ it holds that*

$$|\mathbf{E}[\psi(\boldsymbol{S})] - \mathbf{E}[\psi(\boldsymbol{G})]| \leq (2b_0 d^{7/2}/\eta^2\epsilon^3)\sqrt{\sum_{j=1}^n \sigma_j^4}.$$

It is clear from the definitions that both $\psi_{A^{+\epsilon}}$ and $\psi_{A^{-\epsilon}}$ have range $[0,1]$, and that pointwise, $\psi_{A^{-\epsilon}} \leq \chi_A \leq \psi_{A^{+\epsilon}}$. Thus

$$\mathbf{E}[\psi_{A^{-\epsilon}}(\boldsymbol{S})] \leq \mathbf{Pr}[\boldsymbol{S} \in A] \leq \mathbf{E}[\psi_{A^{+\epsilon}}(\boldsymbol{S})],$$

$$\mathbf{E}[\psi_{A^{-\epsilon}}(\boldsymbol{G})] \leq \mathbf{Pr}[\boldsymbol{G} \in A] \leq \mathbf{E}[\psi_{A^{+\epsilon}}(\boldsymbol{G})].$$

From Lemma A.10 we have that the two left-hand sides above are close and that the two right-hand sides are close. Because of good anti-concentration of Gaussians, it may also be that the left-hand and right-hand sides on the second line are also close, in which $\mathbf{Pr}[\boldsymbol{S} \in A]$ and $\mathbf{Pr}[\boldsymbol{G} \in A]$ will also be close. This motivates the following observation: $\psi_{A^{+\epsilon}} = \psi_{A^{-\epsilon}} = 1$ on $A^{-\epsilon}$ and $\psi_{A^{+\epsilon}} = \psi_{A^{-\epsilon}} = 0$ on the complement of $A^{+\epsilon}$. Hence

$$\mathbf{E}[\psi_{A^{+\epsilon}}(\boldsymbol{G})] - \mathbf{E}[\psi_{A^{-\epsilon}}(\boldsymbol{G})] \leq \mathbf{Pr}[\boldsymbol{G} \in \partial^\epsilon A].$$

Putting together these observations, we conclude:

Theorem A.11. *We have*

$$|\mathbf{Pr}[\boldsymbol{S} \in A] - \mathbf{Pr}[\boldsymbol{G} \in A]| \leq (2b_0 d^{7/2}/\eta^2\epsilon^3)\sqrt{\sum_{j=1}^n \sigma_j^4} + \mathbf{Pr}[\boldsymbol{G} \in \partial^\epsilon A].$$

A.5 Translates of unions of orthants

Let us now specialize to the case where $A \subseteq \mathbb{R}^d$ is a translate of a union of orthants. This means that there exists some vector $\Theta \in \mathbb{R}^d$ such that $X \in A$ depends only on $\overrightarrow{\mathrm{sgn}}(X - \Theta)$. We make the following observation, whose proof is trivial.

Proposition A.12. *If $A \subseteq \mathbb{R}^d$ is a union of orthants then*

$$\partial^\epsilon A \subseteq \bigcup_{i=1}^d W_i^\epsilon,$$

where

$$W_i^\epsilon = \{X \in \mathbb{R}^d : |X[j] - \Theta[j]| \leq \epsilon/2\}.$$

But we also have the following:

Proposition A.13. *Assuming the d-dimensional Gaussian $\boldsymbol{G}$ with covariance matrix C satisfies $C[i,i] = 1$ for all $i \in [d]$, it holds that*

$$\mathbf{Pr}\left[\boldsymbol{G} \in \bigcup_{i=1}^{d} W_i^{\epsilon}\right] \leq d\epsilon/\sqrt{2\pi}.$$

Proof. By a union bound it suffices to prove that $\mathbf{Pr}[|\boldsymbol{G}[i] - \Theta[i]| \leq \epsilon/2] \leq \epsilon/\sqrt{2\pi}$. This is straightforward, as $\boldsymbol{G}[i]$ has distribution $N(0,1)$ and hence has pdf bounded above by $1/\sqrt{2/\pi}$. □

We may now prove Theorem 4.1

Proof. **(Theorem 4.1)** For any $\epsilon > 0$, we may combine Propositions A.12 and A.13 with Theorem A.11 and conclude

$$|\mathbf{Pr}[\boldsymbol{S} \in A] - \mathbf{Pr}[\boldsymbol{G} \in A]| \leq (2b_0 d^2/\eta^2\epsilon^3)\sqrt{\sum_{j=1}^{n} \sigma_j^4} + d\epsilon/\sqrt{2\pi}.$$

The proof is completed by taking $\epsilon = \eta^{-1/2} d^{5/8} (\sum_{j=1}^{n} \sigma_j^4)^{1/8}$ (which is strictly positive since $\sum \sigma_j^4 = 0$ is impossible). □

Identical reasoning gives the proof of Theorem 4.3. Combining Theorems 4.1 and 4.3 gives Theorem 4.4.

B Proof of the Multi-dimensional Critical Index

The fact that the sequence $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n$ was ordered by decreasing 2-norm in Theorem 5.3 was mainly used for notational convenience. We can extract from the proof the following corollary for unordered sequences (whose proof we omit):

Corollary B.1. *Let $\delta, \epsilon, s, b, r, L$ be as in Theorem 5.3. For the unordered collection $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n$, assume we have a sequence of indices $0 \leq j_0 < j_1 < \cdots < j_{L-1} < n$ such that:*

- *for each $0 \leq t < L$, $\sigma_{j_t}^2 \geq \sigma_{j'}^2$ for all $j' > j_t$;*
- *for each $0 \leq t < L$, $\{\boldsymbol{x}_{j_t}, \boldsymbol{x}_{j_t+1}, \ldots, \boldsymbol{x}_n\}$ is* not *δ-regular.*

Assume also that $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{j_L}$ are independent. Then for all $\theta \in \mathbb{R}$,

$$\mathbf{Pr}\left[\left|\boldsymbol{x}_0 + \cdots + \boldsymbol{x}_{j_{L-1}} - \theta\right| \leq s \cdot \tau_{j_{L-1}+1}\right] \leq \epsilon + \frac{O(\ln(1/\epsilon))}{\eta^8 s^4}.$$

The case when $j_t = t$ for $0 \leq t < L$ corresponds to Theorem 5.3.

We now prove Theorem 5.4.

Proof. We construct H_0 according to an iterative process. Initially, $H_0 = \emptyset$, and we define $c_i = 0$ for all $i \in [d]$. In each step of the process, we do the following: First, we select any i such that $c_i < L$ and such that the collection $\{\boldsymbol{X}_j[i] : j \notin H_0\}$ is *not* δ-regular. If there is no such i then we stop the whole process. Otherwise, we continue the step by choosing $j \in [n] \setminus H_0$ so as to maximize $\|\boldsymbol{X}_j[i]\|_2^2$. We then end the step by adding j into H_0 and incrementing c_i.

Note that the process must terminate with $|H_0| \leq dL$; this is because each step increments one of $c_1, \ldots, c_d$, but no c_i can exceed L. When the process terminates, for each i we have either that $\{\boldsymbol{X}_j[i] : j \notin H_0\}$ is δ-regular or that $c_i = L$.

It suffices then to show that when $c_i = L$, the anti-concentration statement holds for i. To see this, first reorder the sequence of random variables $(\boldsymbol{X}_j[i])_j$ so that the first $|H_0|$ are in the order that the indices were added to H_0, and the remaining $n - |H_0|$ are in an arbitrary order. Write $1 \leq j_0 < j_1 < \cdots < j_{L-1} \leq |H_0|$ for the indices that were added to H_0 on those steps which incremented c_i. Then by the definition of the iterative process, for each $0 \leq t < L$ we have that $\|\boldsymbol{X}_{j_t}[i]\|_2^2 \geq \|\boldsymbol{X}_{j'}[i]\|_2^2$ for all $j' > j_t$ and that $\{\boldsymbol{X}_{j_t}[i], \boldsymbol{X}_{j_t+1}[i], \cdots, \boldsymbol{X}_n[i]\}$ is not δ-regular. The anti-concentration statement now follows from Corollary B.1. □

C The Meka-Zuckerman Generator

For the Meka-Zuckerman generator, the first step is to reduce the problem of fooling functions of halfspaces under an arbitrary C-bounded product distribution to fooling an $O(C)$-bounded *discrete* product distribution with support $\mathrm{poly}(n, C, \epsilon^{-1})$ in each co-ordinate.

Lemma C.1. *Given a C-bounded dsitribution $\boldsymbol{X}$, there is a discrete product distribution $\boldsymbol{Y}$ such that if $f : \mathbb{R}^n \to \{-1, 1\}$ is a function of d halfspaces $\{h_i : \mathbb{R}^n \to \{-1, 1\}\}_{i \in [d]}$, then*

$$|\,\mathbf{E}[f(\boldsymbol{X})] - \mathbf{E}[f(\boldsymbol{Y})]| \leq O\left(\frac{d\epsilon^2}{nC^2}\right).$$

Each $\boldsymbol{y}_i$ is distributed uniformly over a multiset $\Omega_i = \{b_1(i) \leq \cdots \leq b_g(i)\}$ where $|b_j(i)| \leq (nC^2\epsilon^{-1})^{\frac{1}{4}}$. For every i, we have $|\Omega_i| = 2^s = O(n^2C^2\epsilon^{-2})$ and Further $\mathbf{E}[\boldsymbol{y}_i] = 0, \mathbf{E}[\boldsymbol{y}_i^2] = 1$, $\mathbf{E}[\boldsymbol{y}_i^4] \leq O(C)$.

We are interested in $d << n$, so the error in going from $\boldsymbol{X}$ to $\boldsymbol{Y}$ is $o(\epsilon^2)$. Since $|\Omega_i| = 2^s$ for all i, sampling k-wise independently from $\boldsymbol{Y}$ reduces to generating n strings of length s in a k-wise independent manner: this can be done using $k \max(\log n, s) = O(k \log(nC/\epsilon))$ random bits.

This lemma is proved by *sandwiching* $\boldsymbol{X}$ between two discrete product distributions $\boldsymbol{Y}^u$ and $\boldsymbol{Y}^\ell$ which are close to each other in statistical distance. The proof is in Appendix H. Henceforth, we will rename $\boldsymbol{Y}$ as $\boldsymbol{X}$ and focus on fooling discrete product distributions.

We now describe the main generator of Meka-Zuckerman, modified so that random variables take values in $\prod_j \Omega_j$ instead of simply ± 1. At a high level, the generator hashes variables into buckets and uses bounded independence for the variables within each bucket. We use a weaker property of hash functions than used in [MZ09].

Efficient constructions of size $|\mathcal{H}| = O(nt)$ are known for any constant $C \geq 1$. $C = 1$ is optimal, and can be achieved by a pairwise independent family. In our construction we use $C = 1$, but we will need larger C in our analysis. A hash function induces a partition of $[n]$. The generator first picks a partition of $[n] = H_1 \cup \ldots \cup H_t$ using a random element from $\mathcal{H}$. For each $i \in [t]$, it then generates a 5-wise independent distribution $(\boldsymbol{y}_j)_{j \in H_i}$ on $\prod_{j \in H_i} \Omega_j$. Such a distribution on n random variables can be generated using a seed of length $k \log \max(n, |\Omega|)$. These t distributions are chosen independently. The generator outputs $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$. The seedlength required is $\log(2n) + 5t \log \max(n, |\Omega|)$ where $\log(2n)$ are required for the hash function and $5 \log \max(n, |\Omega|)$ bits are needed for each H_i, $i \in [t]$.

D Analyzing the Meka-Zuckerman Generator

The first claim we prove is that the indices in the set H_0 are likely to be hashed into distinct buckets.

Definition D.1. *A hash function $h : [n] \to [t]$ is* S-isolating *if for all $x \neq y \in S$, $h(x) \neq h(y)$. A family of hash functions $\mathcal{H} = \{h : [n] \to [t]\}$ is (ℓ, β)-isolating if for any $S \subseteq [n]$, $|S| \leq \ell$,*

$$\mathop{\mathbf{Pr}}_{h \in_u \mathcal{H}}[h \text{ is not } S\text{-isolating}] \leq \beta.$$

A C-collision preserving hash family is likely to be isolating for small sets:

Lemma D.2. *Assume t is a power of 2. A C-collision preserving family of hash functions $\mathcal{H} = \{h : [n] \to [t]\}$ is (ℓ, β)-collision free for $\beta = C\ell^2/(2t)$.*

Proof. The expected number of collisions for a set S is at most

$$\binom{|S|}{2}\frac{C}{t} \leq \frac{C\ell^2}{2t}.$$

By increasing n to the next largest power of 2, since t is a power of 2, there is a field $\mathbb{F}$ of size n where $t|n$. Then there is a hash family of size n for $C = 1$. For any element $a \in \mathbb{F}$, define a hash function $h_a(x) = (ax) \bmod t$. Here x is viewed as a field element, the multiplication is done in the field, and the product is then viewed as a nonnegative integer less than n before taking the mod. We can increase n and t to be the nearest powers of 2. We can therefore take $\mathcal{H}$ to have size at most $2n$ for $C = 1$. □

We want the set H_0 to be isolated with error ϵ, so we want $\ell = dL$ and $\beta = \epsilon$. Hence we set t to be the smallest power of 2 larger than $\ell^2/\epsilon = (dL)^2/\epsilon$.

We will aim to achieve error $O(d\epsilon)$ (rather than $O(\epsilon)$), as this makes the notation easier. We set the parameters s and δ in Theorem 5.4 as

$$s = 1/(\eta^2\sqrt{\epsilon}), \quad \delta = \frac{\eta^4\epsilon^8}{d^7}.$$

This implies that

$$L = O\left(\frac{\log(s)\log(1/\epsilon)}{\eta^8}\right) \cdot \frac{1}{\delta} = O\left(\frac{d^7\log^2(\epsilon\eta)}{\eta^{12}\epsilon^8}\right), \quad t = O\left(\frac{(dL)^2}{\epsilon}\right) = O\left(\frac{d^{15}\log^4(\epsilon\eta)}{\eta^{24}\epsilon^{17}}\right).$$

D.1 Analysis for Functions of Regular Halfspaces

Recall that our goal is to fool functions of $\overrightarrow{\mathrm{sgn}}(\sum_j x_j W_j - \theta)$. Let $\boldsymbol{Y}_j = \boldsymbol{y}_j W_j$ and $\boldsymbol{T} = \sum_{j=1}^n \boldsymbol{Y}_j$. Similarly let $\boldsymbol{X}_j = \boldsymbol{x}_j W_j$ and $\boldsymbol{S} = \sum_{j=1}^n \boldsymbol{X}_j$. Thus we are interested in bounding

$$|\mathop{\mathbf{Pr}}_{\boldsymbol{X}}[\boldsymbol{S} \in A] - \mathop{\mathbf{Pr}}_{\boldsymbol{Y}}[\boldsymbol{T} \in A]|$$

where A is a translate of union of orthants: membership of a point $X \in \mathbb{R}^d$ in A is a function of $\overrightarrow{\mathrm{sgn}}(X - \Theta)$. By rescaling the W_j and Θ, we may assume without loss of generality that

$$C[i,i] = \sum_{j=1}^n \mathbf{E}\left[\boldsymbol{X}_j[i]^2\right] = 1 \qquad \text{for all } i \in [d].$$

The regular case is when the vectors $W_1, \ldots, W_n$ are such that for every i, the sequence of random variables $\{\boldsymbol{X}_j[i]\}_{j=1}^n$ is δ-regular. In this case, we can directly appeal to the Berry-Esseen theorem to prove th correctness of the MZ generator.

Theorem D.3. *If the sequence of random variables $\{\boldsymbol{X}_j[i]\}_{j=1}^n$ is δ-regular for all $i \in [d]$, then the MZ generator $O(d\epsilon)$-fools any function of* $\operatorname{sgn}(W \cdot X - \Theta)$ *for all $\Theta \in \mathbb{R}^d$.*

Proof. We can therefore apply the machinery developed above. For the regular case, we only need to use 4-wise independence. Thus, the random variables $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ satisfy the 4-matching-moments condition with respect to $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, as defined in Subsection A.3.

The definition of δ-regular is given in Definition 5.1. Let $\sigma_{i,j} = \|\boldsymbol{X}_j[i]\|_2$. Suppose that for all i, the set of real random variables $\{\boldsymbol{X}_j[i]\}$ is δ-regular, i.e.,

$$\sum_{j=1}^n \|\boldsymbol{X}_j[i]\|_4^4 \le \delta\Big(\sum_{j=1}^n \|\boldsymbol{X}_j[i]\|_2^2\Big)^2 = \delta(\sigma_{i,1}^2 + \cdots + \sigma_{i,n}^2)^2 = \delta,$$

where the last equality is from our normalization. We wish to apply Theorem 4.4. Since $\sigma_{i,j} = \|\boldsymbol{X}_j[i]\|_2 \le \|\boldsymbol{X}_j[i]\|_4$, we conclude that for all i,

$$\sum_{j=1}^n \sigma_{i,j}^4 \le \sum_{j=1}^n \|\boldsymbol{X}_j[i]\|_4^4 \le \delta.$$

Since $\sigma_j^2 = \sum_{i=1}^d \sigma_{i,j}^2$, by Cauchy-Schwarz we get

$$\sigma_j^4 = \left(\sum_{i=1}^d \sigma_{i,j}^2\right)^2 \le d\left(\sum_{j=1}^n \sigma_{i,j}^4\right) \le d\delta.$$

Therefore $\sum_{j=1}^n \sigma_j^4 \le d^2\delta$. Hence we can apply Theorem 4.4 to obtain

$$|\mathbf{Pr}[\boldsymbol{S} \in A] - \mathbf{Pr}[\boldsymbol{T} \in A]| \le O\left((1/\eta)^{1/2} d^{15/8}) \cdot (t^{-1} + \delta)^{1/8}\right) \le O(d\epsilon).$$

where the last inequality follows from the choice of t, δ. □

D.2 Analysis for Functions of General Halfspaces

We now combine Theorem 5.4 with the analysis of the Regular case (Theorem D.3), to prove that the MZ generator fools functions of arbitrary halfspaces.

Theorem D.4. *The MZ generator $O(d\epsilon)$-fools any function of d halfspaces with seed length*

$$O(t \log(\max(n, |\Omega|))) = O\left(\frac{d^{15} \log^4(\epsilon\eta) \log(n/\epsilon\eta)}{\eta^{24}\epsilon^{17}}\right).$$

Proof. Apply Theorem 5.4 with these parameters. Then there exists a set $H_0 \subseteq [n]$ of size at most dL such that the coordinates $[d]$ can be partitioned into two sets, REG and JUNTA, such that the following holds.

1. For $i \in$ REG, the set of real random variables $\{\boldsymbol{X}_j[i] : j \notin H_0\}$ is δ-regular.
2. For $i \in$ JUNTA, for all $\theta \in \mathbb{R}$,

$$\mathbf{Pr}\left[\Big|\sum_{j \in H_0} \boldsymbol{X}_j[i] - \theta\Big| \le s \cdot \sqrt{\sum_{j \notin H_0} \|\boldsymbol{X}_j[i]\|_2^2}\right] \le \epsilon + \frac{O(\log(1/\epsilon))}{\eta^8 s^4} \le 2\epsilon \tag{10}$$

We condition on the hash function h being S-collision free, which happens with probability at least $1-\epsilon$. Therefore, at most one variable from H_0 lands in each set in the partition. Since the distribution in each partition set is 5-wise independent, this means that the distribution on H_0 is fully independent. This allows us to construct a coupling of $\boldsymbol{X}$ and $\boldsymbol{Y}$: let $\boldsymbol{X}_j = \boldsymbol{Y}_j$ for $j \in H_0$, and then sample the rest according to the correct marginal distribution.

We say that the variables in H_0 are good if

$$\Big|\sum_{j\in H_0} \boldsymbol{Y}_j[i] - \theta[i]\Big| > s \cdot \sqrt{\sum_{j\notin H_0} \|\boldsymbol{Y}_j[i]\|_2^2} \text{ for all } i \in V$$

By Equation 10,

$$\mathbf{Pr}[\{\boldsymbol{X}_j = \boldsymbol{Y}_j\}_{j\in H_0} \text{ are not good}] \le 2d\epsilon. \tag{11}$$

We condition on these variables being good.

With this conditioning, we show that the halfspaces in JUNTA are nearly constant: with high probability they do not depend on the variables outside H_0. To see this, observe that conditioned on the variables in H_0, the remaining variables are still 4-wise independent (in both $\boldsymbol{X}$ and $\boldsymbol{Y}$), so by Chebychev

$$\mathbf{Pr}\left[\Big|\sum_{j\notin H_0} \boldsymbol{Y}_j[i]\Big| \ge s \cdot \sqrt{\sum_{j\notin H_0} \|\boldsymbol{Y}_j[i]\|_2^2}\right] \le 1/s^2 \le \epsilon. \tag{12}$$

But if this does not happen, then

$$\text{sign}(\sum_{j=1}^{n} \boldsymbol{Y}_j[i] - \Theta[i]) = \text{sign}(\sum_{j\in H_0} \boldsymbol{Y}_j[i] - \Theta[i])].$$

A similar analysis holds for $\boldsymbol{X}$. Thus for both $\boldsymbol{X}$ and $\boldsymbol{Y}$, with error probability at most $2d/s^2 \le 2d\epsilon$, we can assume that the halfspaces in JUNTA are fixed to constant functions for a good choice of variables in H_0.

Recall that we are interested in fooling functions of the form $g(h_1(\boldsymbol{X}), \ldots, h_k(\boldsymbol{X}))$. Conditioned on the variables in H_0 being good, the halfspaces h_j for $j \in$ JUNTA are close to constant functions. Thus, the function g is $2d\epsilon$ close to a function g' of halfspace $\{h_j\}_{j\in \text{REG}}$ under both distributions $\boldsymbol{X}$ and $\boldsymbol{Y}$. Thus it suffices to show that the bias of g' under $\boldsymbol{X}$ and $\boldsymbol{Y}$ is close.

Conditioning on $\boldsymbol{X}_j = \boldsymbol{Y}_j$ for $j \in H_0$ gives a halfspace on the remaining variables in each coordinate $i \in$ REG. Define

$$\Theta'[i] = (\Theta[i] - \sum_{j\notin H_0} \boldsymbol{X}_j[i])), \quad \boldsymbol{S}'[i] = \sum_{j\notin H_0} \boldsymbol{X}_j[i], \ \boldsymbol{T}'[i] = \sum_{j\notin H_0} \boldsymbol{Y}_j[i].$$

then

$$\text{sgn}(\boldsymbol{S}[i] - \Theta[i]) = \text{sgn}(\boldsymbol{S}'[i] - \Theta'[i]).$$

Thus there exists a union of orthants $A' \in \mathbb{R}^{|\text{REG}|}$ such that $g'(X) = 1$ if $X \in A'$. Our goal is to bound

$$\left|\mathbf{Pr}[\boldsymbol{S}' \in A'] - \mathbf{Pr}[\boldsymbol{T}' \in A']\right|.$$

The set of random variables $\{\boldsymbol{X}_j[i] : j \notin H_0\}$ is δ-regular. Hence we can apply our result for the regular case. We've already conditioned on the hash function h being H_0-collision free. Since this happens with probability at least $1-\epsilon$, the resulting function is C-collision preserving for $C = 1/(1-\epsilon) \le 2$, since conditioning on an event which happens with probability p can increase

the probability of any other event by a factor of at most $1/p$. So now applying the analysis from the regular case,

$$\left|\mathop{\mathbf{Pr}}_{\boldsymbol{S}'}[\boldsymbol{S}' \in A'] - \mathop{\mathbf{Pr}}_{\boldsymbol{T}'}[\boldsymbol{T}' \in A']\right| \leq O\left(\eta^{-1/2} d^{15/8} \cdot (\frac{1}{t} + \delta)^{1/8}\right) \leq O(d\epsilon). \tag{13}$$

Hence, conditioned on $\mathbf{h}$ and the variables in H_0 being good, we have

$$\left|\mathop{\mathbf{Pr}}_{\boldsymbol{S}}[\boldsymbol{S} \in A] - \mathop{\mathbf{Pr}}_{\boldsymbol{T}}[\boldsymbol{T} \in A]\right| \leq O(d\epsilon) + 2d\epsilon. \tag{14}$$

Removing the conditioning gives

$$\left|\mathop{\mathbf{Pr}}_{\boldsymbol{S}}[\boldsymbol{S} \in A] - \mathop{\mathbf{Pr}}_{\boldsymbol{T}}[\boldsymbol{T} \in A]\right| \leq O(d\epsilon) + 2d\epsilon + \epsilon + 2d\epsilon = O(d\epsilon)$$

□

E Generalized Monotone Trick

We generalize the "monotone trick" introduced in Meka and Zuckerman [MZ09] and show that a generator that fools small-width "monotone" branching programs also fools any monotone function of several arbitrary-width monotone branching programs.

First we define read-once branching programs. Branching programs corresponding to space S have width 2^S. We use the following notation from [MZ09].

Definition E.1 (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 $0 \leq i \leq T$, a vertex v in layer i has at most 2^D outgoing edges each labeled with an element of $\{0,1\}^D$ and pointing to a vertex in layer $i+1$.*

Let M be an (S, D, T)-branching program and v a vertex in layer i of M. We now define the set of accepting suffixes.

Definition E.2. *We say z is an* accepting suffix *from vertex v if the path in M starting at v and following edges labeled according to z leads to an accepting state. We let $\mathrm{Acc}_M(v)$ denote the set of accepting suffixes from v. If M is understood we may abbreviate this $\mathrm{Acc}(v)$.*

Nisan [Nis92] and Impagliazzo et al. [INW94] gave PRGs that fool (S, D, T)-branching programs with error $\exp(2^{-\Omega(S+D)})$ and seed length $r = O((S + D + \log T)\log T)$. For $T = \mathrm{poly}(S, D)$, the PRG of Nisan and Zuckerman [NZ96] fools (S, D, T)-branching programs with seed length $r = O(S + D)$. Meka and Zuckerman showed that the above PRGs in fact fool arbitrary width branching programs of a certain form called monotone, defined next.

Definition E.3 (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 $v \prec w$ implies $\mathrm{Acc}_M(v) \subseteq \mathrm{Acc}_M(w)$.*

Note that the natural ROBP accepting a halfspace, where states correspond to partial sums, is monotone. However, the natural ROBP accepting the intersection of just two halfspaces may not be monotone.

The following theorem is the only known way to obtain PRGs for halfspaces using seed length which depends logarithmically on $1/\epsilon$ (and polylogarithmically on n).

Theorem E.4. *[MZ09] Let $0 < \epsilon < 1$ and $G : \{0,1\}^R \to (\{0,1\}^D)^T$ be a PRG that δ-fools monotone $(\log(4T/\epsilon), D, T)$-branching programs. Then G $(\epsilon+\delta)$-fools monotone (S, D, T)-branching programs for arbitrary S with error at most $\epsilon + \delta$.*

We now generalize Theorem E.4 to the intersection of monotone branching programs, or even to any monotone function of monotone branching programs. (Of course, the intersection corresponds to the monotone function AND.)

Theorem E.5. *Let $0 < \epsilon < 1$ and $G : \{0,1\}^R \to (\{0,1\}^D)^T$ be a PRG that δ-fools monotone $(d\log(4Td/\epsilon), D, T)$-branching programs. Then G $(\epsilon + \delta)$-fools any monotone function of d monotone (S, D, T)-branching programs for arbitrary S.*

We now generalize monotone functions to decision trees. First note that the complement of a monotone branching program is a monotone branching program. Now consider any decision tree, where each node of the decision tree is a monotone branching program. Any leaf of this tree represents the intersection of monotone branching programs. Thus, the error of the function above for such decision trees is at most s times the error for each leaf. This gives the following corollary.

Corollary E.6. *Let $0 < \epsilon < 1$ and $G : \{0,1\}^R \to (\{0,1\}^D)^T$ be a PRG that δ-fools monotone $(d\log(4Td/\epsilon), D, T)$-branching programs. Then G $(s(\epsilon + \delta))$-fools any decision tree with s leaves, where each decision tree node is a monotone (S, D, T)-branching programs for arbitrary S.*

In the above, we can even take s to be the minimum of the number of 0 and 1 leaves. We now prove Theorem E.5, using the ideas of [MZ09] based on "sandwiching" monotone branching programs between small-width branching programs.

Definition E.7. *A pair of functions $(f_{\text{down}}, f_{\text{up}})$, each with the same domain and range as a function $f : B \to \{0,1\}$, is said to* ϵ-sandwich *f if the following hold.*

1. *For all $z \in B$, $f_{\text{down}}(z) \le f(z) \le f_{\text{up}}(z)$.*
2. $\mathbf{Pr}_{z \in_u B}[f_{\text{up}}(z) = 1] - \mathbf{Pr}_{z \in_u B}[f_{\text{down}}(z) = 1] \le \epsilon$.

The following lemma shows that it suffices to fool functions which sandwich the given target function. Bazzi [Baz09] used sandwiching in showing that polylog-wise independence fools DNF formulas. The lemma below is a small modification of a lemma in [MZ09].

Lemma E.8. *If $(f_{\text{down}}, f_{\text{up}})$ ϵ-sandwich f, and a PRG G δ-fools f_{down} and f_{up}, then G $(\epsilon + \delta)$-fools f.*

Meka and Zuckerman then showed that any monotone branching program can be sandwiched between two small-width branching programs.

Lemma E.9. *[MZ09] For any monotone (S, D, T)-branching program M, there exist monotone $(\log(4T/\epsilon), D, T)$-branching programs $(M^{\text{down}}, M^{\text{up}})$ that ϵ-sandwich M.*

Using this, we can show that any monotone function of monotone branching programs is sandwiched by a small-width branching program.

Lemma E.10. *Any monotone function of d (S, D, T)-branching programs has a pair of $(d\log(4T/\epsilon), D, T)$-branching programs $(M^{\text{down}}, M^{\text{up}})$ that $(d\epsilon)$-sandwich it.*

Proof. For a monotone branching program M, let $(M^{\text{down}}, M^{\text{up}})$ denote monotone $(\log(4T/\epsilon), D, T)$-branching programs that ϵ-sandwich M, as given by Lemma E.9. Suppose our given function is $f(z) = g(M_1(z), M_2(z), \ldots, M_d(z))$ for g monotone. Then $f(z)$ is sandwiched by $(f_{\text{down}}, f_{\text{up}})$ given by

$$\begin{aligned} f_{\text{down}}(z) &= f\left(M_1^{\text{down}}(z), M_2^{\text{down}}(z), \ldots, M_d^{\text{down}}(z)\right) \\ f_{\text{up}}(z) &= f\left(M_1^{\text{up}}(z), M_2^{\text{up}}(z), \ldots, M_d^{\text{up}}(z)\right). \end{aligned}$$

Moreover,

$$f_{\text{up}}^{-1}(1) - f_{\text{down}}^{-1}(1) \subseteq \bigcup_{i=1}^{d} \left((M_i^{\text{up}})^{-1}(1) - (M_i^{\text{down}})^{-1}(1)\right).$$

Since $\mathbf{Pr}_z[M_i^{\text{up}}(z) = 1] - \mathbf{Pr}_z[M_i^{\text{down}}(z) = 1] \leq \epsilon$, it follows that $\mathbf{Pr}_z[f_{\text{up}}(z) = 1] - \mathbf{Pr}_z[f_{\text{down}}(z) = 1] \leq d\epsilon$. □

Theorem E.5 now follows from Lemmas E.8 and E.10. Without using any of the hard work we've done in other sections, this theorem gives us PRGs for monotone functions of halfspaces (such as intersections) using a random seed of length $O(d(\log n)\log(n/\epsilon))$. We improve this seed length now.

E.1 Combining the Monotone Trick and the Main Construction

Fix a hash function h, which fixes the partition into t sets. Then any monotone function of $\overrightarrow{\text{sgn}}(y_1 W_1 + \ldots + y_n W_n - \Theta)$ may be computed by a monotone function of d monotone branching programs, with t layers each. Thus, we can apply Theorem E.5 and Corollary E.6 to deduce Theorem 1.2.

We can set $T = t$ and $D = O(\log n)$ to store the seed for the 5-wise independent distribution. Also note that $\log \eta^{-1} = \Theta(\log C)$. With these parameters, using Nisan's PRG gives a seed length of $O((d\log(dT/\epsilon) + D + \log T)\log T) = O((d\log(Cd/\epsilon) + \log n)\log(Cd/\epsilon))$ to fool monotone functions of d halfspaces. For functions computable by size s decision trees of halfspaces, the seed length becomes $O((d\log(Cds/\epsilon) + \log n)\log(Cds/\epsilon))$.

When $Cd/\epsilon \geq \log^{-c} n$ for any $c > 0$, then $t = \text{polylog}(n)$ and we can use the Nisan-Zuckerman PRG. This gives a seed length of $O(d\log(dT/\epsilon) + D + \log T) = O(d\log(Cd/\epsilon) + \log n)$ for monotone functions of d halfspaces. For functions computable by size s decision trees of halfspaces, the seed length becomes $O(d\log(Cds/\epsilon) + \log n)$.

More generally, using Armoni's interpolation of Nisan and Nisan-Zuckerman will shave off an extra $\log\log n$ factor off of Nisan's PRG when $t/\epsilon \leq \exp(-(\log n)^{1-\gamma})$ for some $\gamma > 0$. We omit the details.

F Bounded Independence fools functions of Halfspaces

In this section, we prove Theorem 1.5

F.1 Reduction to upper polynomials for single halfspaces

We now flesh out the reduction described in Section 2, i.e., we show how to prove Theorem 2.5 given upper sandwiching polynomials for a single halfspace with extra properties.

Lemma F.1. *Let $\boldsymbol{X}$ be a random vector on the product set Ω, and suppose we have order-k polynomials $p_1, \dots, p_d : \Omega \to \mathbb{R}$, as well as functions $h_1, \dots, h_d : \Omega \to \{0, 1\}$. Write $\boldsymbol{p}_i = p_i(\boldsymbol{X})$ and $\boldsymbol{h}_i = h_i(\boldsymbol{X})$. Assume that for each $i \in [k]$:*

1. *$\boldsymbol{p}_i \geq \boldsymbol{h}_i$ with probability 1;*
2. *$\mathbf{E}[\boldsymbol{p}_i - \boldsymbol{h}_i] \leq \epsilon_0$;*
3. *$\mathbf{Pr}[\boldsymbol{p}_i > 1 + 1/d^2] \leq \gamma$;*
4. *$\|\boldsymbol{p}_i\|_{2d} \leq 1 + 2/d^2$.*

If we write $p = p_1 p_2 \cdots p_d$, $h = h_1 h_2 \cdots h_d$, then p is a polynomial of order at most dk, $p(\boldsymbol{X}) \geq h(\boldsymbol{X})$ with probability 1, and

$$\mathbf{E}[p(\boldsymbol{X}) - h(\boldsymbol{X})] \leq 2d\epsilon_0 + 3d^2\sqrt{\gamma}. \tag{15}$$

Proof. The first two parts of the claim are immediate, so it suffices to verify (15). We use the telescoping sum (1), and thus it suffices to bound the general term as follows:

$$\mathbf{E}[\boldsymbol{h}_1 \cdots \boldsymbol{h}_{i-1}(\boldsymbol{p}_i - \boldsymbol{h}_i)\boldsymbol{p}_{i+1} \cdots \boldsymbol{p}_d] \leq 2\epsilon_0 + 3d\sqrt{\gamma}. \tag{16}$$

We have

$$\begin{aligned}
& \mathbf{E}[\boldsymbol{h}_1 \cdots \boldsymbol{h}_{i-1}(\boldsymbol{p}_i - \boldsymbol{h}_i)\boldsymbol{p}_{i+1} \cdots \boldsymbol{p}_d] \\
\leq\ & \mathbf{E}[\boldsymbol{p}_1 \cdots \boldsymbol{p}_{i-1}(\boldsymbol{p}_i - \boldsymbol{h}_i)\boldsymbol{p}_{i+1} \cdots \boldsymbol{p}_d] \\
<\ & 2\,\mathbf{E}[\boldsymbol{p}_i - \boldsymbol{h}_i] + \mathbf{E}[\mathbf{1}[\boldsymbol{p}_1 \cdots \boldsymbol{p}_{i-1}\boldsymbol{p}_{i+1} \cdots \boldsymbol{p}_d \geq 2]\, \boldsymbol{p}_1 \cdots \boldsymbol{p}_{i-1}(\boldsymbol{p}_i - \boldsymbol{h}_i)\boldsymbol{p}_{i+1} \cdots \boldsymbol{p}_d] \\
\leq\ & 2\epsilon_0 + \mathbf{E}\left[\left(\sum_{i'=1}^{d} \mathbf{1}\left[\boldsymbol{p}_{i'} > 1 + 1/d^2\right]\right) \prod_{i=1}^{d} \boldsymbol{p}_i\right],
\end{aligned}$$

where in the last term we used the bounds $(1 + 1/d^2)^{d-1} < 2$ and $\boldsymbol{p}_i - \boldsymbol{h}_i \leq \boldsymbol{p}_i$. Thus we can establish (16) by showing the bound

$$\sum_{i'=1}^{d} \mathbf{E}\left[\mathbf{1}\left[\boldsymbol{p}_{i'} > 1 + 1/d^2\right] \prod_{i=1}^{d} \boldsymbol{p}_i\right] \leq 3d\sqrt{\gamma}.$$

This follows by bounding each summand individually:

$$\begin{aligned}
& \mathbf{E}\left[\mathbf{1}\left[\boldsymbol{p}_{i'} > 1 + 1/d^2\right] \prod_{i=1}^{d} \boldsymbol{p}_i\right] \\
\leq\ & \|\mathbf{1}\left[\boldsymbol{p}_{i'} > 1 + 1/d^2\right]\|_2 \cdot \prod_{i=1}^{d} \|\boldsymbol{p}_i\|_{2d} \qquad \text{(Hölder's inequality)} \\
\leq\ & \sqrt{\gamma} \cdot (1 + 2/d^2)^d \quad \leq \quad 3\sqrt{\gamma},
\end{aligned}$$

as needed. □

F.2 Tools for upper polynomials

We construct the upper sandwiching polynomial needed in Lemma F.1 using two key tools: "DGJSV Polynomials", the family of univariate real polynomial constructed in [DGJ$^+$09] for approximating the sgn function; and, our Regularity Lemma for halfspaces over general random variables 5.3.

Regarding the DGJSV Polynomials, the following is a key theorem from [DGJ$^+$09] (slightly adjusted for our purposes):

Theorem F.2. *([DGJ$^+$09]) Let $0 < a, b < 1$. Then there exists an even integer $K = K_{a,b}$ with*

$$K \le C_0 \frac{\log(2/b)}{a} \qquad (C_0 \text{ is a universal constant})$$

as well as an ordinary univariate real polynomial $P = P_{a,b} : \mathbb{R} \to \mathbb{R}$ of degree K with the following behavior:

- $P(x) \ge 0$ *for* $x \in (-\infty, -1]$,
- $0 \le P(x) \le b$ *for* $x \in [-1, -a]$;
- $0 \le P(x) \le 1$ *for* $x \in [-a, 0]$;
- $1 \le P(x) \le 1 + b$ *for* $x \in [0, 1]$;
- $P(x) \ge 1$ *for* $x \in [1, \infty)$;
- $P(x) \le (4x)^K$ *for all* $|x| \ge 1$.

Note that the first five conditions imply $P(x) \ge \mathbf{1}[x \ge 0]$ for all $x \in \mathbb{R}$.

Regarding our Regularity Lemma for general halfspaces, we will use the following rephrasing of Theorem 5.3 with simplified parameters:

Theorem F.3. *Let $t > 1$, $0 < \delta < 1$ and $0 < \eta$ be parameters. Then there exists an integer L satisfying*

$$L \le \mathrm{poly}(\log t, 1/\eta) \cdot \frac{1}{\delta}$$

such that the following holds. Suppose $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ is a sequence of independent η-HC random variables, $\theta \in \mathbb{R}$, and $n \ge L$. Then there exists a set of coordinates $H \subseteq [n]$ of cardinality L such that, denoting

$$\boldsymbol{\theta}' = \theta - \sum_{j \in H} \boldsymbol{x}_j, \qquad \boldsymbol{z} = \sum_{j \notin H} \boldsymbol{x}_j$$

(these random variables are independent), we have three mutually exclusive and collectively exhaustive events depending only on $\boldsymbol{\theta}'$*:*

- *Event* **BAD***: $|\boldsymbol{\theta}'| \le t\|\boldsymbol{z}\|_2$ and the collection $\{\boldsymbol{x}_j : j \notin H\}$ is not δ-regular;*
- *Event* **NEAR***: $|\boldsymbol{\theta}'| \le t\|\boldsymbol{z}\|_2$ and the collection $\{\boldsymbol{x}_j : j \notin H\}$ is δ-regular;*
- *Event* **FAR***: $|\boldsymbol{\theta}'| > t\|\boldsymbol{z}\|_2$.*

Furthermore, **BAD** *has probability at most $O(1/t^4)$.*

The reader will note that events **BAD**, **NEAR**, and **FAR** are defined somewhat peculiarly: Neither $\|\boldsymbol{z}\|_2$ nor the (ir)regularity of $\{\boldsymbol{x}_j : j \notin H\}$ is actually random. Furthermore, by our original Theorem 5.3, we either have that $\{\boldsymbol{x}_j : j \notin H\}$ is δ-regular, in which case **NEAR** and **FAR** are the only possible events, or the collection is not δ-regular, in which case **BAD** and **FAR** are the only possible events. Nevertheless, this tripartition of events makes our future analysis simpler.

F.3 Statement of the main technical theorem, and how it completes the proof

The main technical result we will prove is the following:

Theorem F.4. *Let $k \geq 1$, $0 < \delta < 1$, and $t > 4$ be parameters. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be a vector of independent η-HC random variables. Furthermore, let T be an even integer such that the $\boldsymbol{x}_i$'s are $(T, 2, 4/t)$-hypercontractive. Assume $T \geq C_1 d \log(dt)$, where C_1 is a universal constant. Let $\theta \in \mathbb{R}$ and let*

$$h(x_1, \ldots, x_n) = \mathbf{1}[x_1 + \cdots + x_n - \theta \geq 0]\,.$$

Then there exists a polynomial $p(x_1, \ldots, x_n)$ of order k, with

$$k \leq \mathrm{poly}(\log t, 1/\eta) \cdot \frac{1}{\delta} + O(T/d),$$

satisfying the 4 properties appearing in Lemma F.1, with

$$\epsilon_0 = O(\sqrt{\delta}) + O(\epsilon_1), \qquad \epsilon_1 = \frac{dt \log(dt)}{T}, \qquad \gamma = 2^{-T/d}.$$

As we now show, using Theorem F.4 and Lemma F.1, we can deduce Theorem 2.5 and hence Theorem 1.5 simply by choosing parameters appropriately. Note that it is sufficient to prove Theorem 2.5 with $\epsilon \cdot \mathrm{polylog}(d/\epsilon)$ in place of ϵ.

We will apply Theorem F.4 with $\delta = \Theta(\epsilon^2/d^2)$ and

$$t = C_2 \frac{d^2}{\epsilon\alpha},$$

where C_2 is a large constant of our choosing. Regarding the hypercontractivity parameters, using Fact I.1, we may take

$$\eta = \Theta(\alpha^{-1/4}), \qquad T = \Theta(t^2 \cdot \alpha \ln(2/\alpha)).$$

The necessary assumption that

$$T \geq C_1 d \log(td) \quad \Leftrightarrow \quad C_2^2 \cdot \Theta\left(\frac{d^4 \ln(2/\alpha)}{\epsilon^2 \alpha}\right) \geq C_1 d \log\left(C_2 \frac{d^3}{\epsilon\alpha}\right)$$

is valid provided that C_2 is a sufficiently large constant.

We obtain from the theorem an upper ϵ_2-sandwiching polynomial for h with order

$$k = \widetilde{O}(d^2/\epsilon^2) \cdot \mathrm{poly}(1/\alpha) + O(d^3/\epsilon^2) \cdot \mathrm{poly}(1/\alpha) \leq O(d^3/\epsilon^2) \cdot \mathrm{poly}(1/\alpha),$$

where

$$\epsilon_0 = O(\epsilon/d) + \widetilde{O}(\epsilon/d) = \widetilde{O}(\epsilon/d)$$

and γ is exponentially small in $d/(\epsilon\alpha)$. By using such polynomials in Lemma F.1, we get upper sandwiching polynomials for intersections of d halfspaces with the claimed degree $kd = \widetilde{O}(d^4/\epsilon^2) \cdot \mathrm{poly}(1/\alpha)$ and the claimed error $d\epsilon_0 = \epsilon \cdot \mathrm{polylog}(d/\epsilon)$.

F.4 Proof of Theorem F.4

In this section, we prove Theorem F.4. Let H be the set of cardinality $L = \mathrm{poly}(\log t, 1/\eta) \cdot (1/\delta)$ coming from Theorem F.3, and assume without loss of generality that $H = \{1, \dots, L\}$. We use the notation $\theta' = \theta - (x_1 + \cdots + x_L)$, $z = x_{L+1} + \cdots + x_n$, $\mathrm{BAD} = \mathrm{BAD}(x_1, \dots, x_L)$ etc., with boldface indicating randomness as usual. Given the outcomes for $\boldsymbol{x}_1, \dots, \boldsymbol{x}_L$, we will handle the three events **BAD**, **NEAR**, and **FAR** with separate ordinary real polynomials. More precisely, our final (generalized) polynomial will be

$$p(x_1, \dots, x_n) = \mathbf{1}[\mathbf{BAD}] \cdot 1 + \mathbf{1}[\mathbf{NEAR}] \cdot p^{\mathrm{near}}_{\theta'}(z) + \mathbf{1}[\mathbf{FAR}] \cdot p^{\mathrm{far}}_{\theta'}(z),$$

where

$$p^{\mathrm{near}}_{\theta'}(z) = P\left(\frac{z - \theta'}{2t\|\boldsymbol{z}\|_2}\right),$$

and

$$p^{\mathrm{far}}_{\theta'}(z) = \mathbf{1}\big[\theta' > 0\big] \cdot 1 + \mathbf{1}\big[\theta' \le 0\big] \cdot \left(\frac{z}{\theta'}\right)^q,$$

where q is a positive integer and P is an ordinary real univariate polynomial to be specified later. For typographic simplicity, we will write simply $p_{\theta'}$ in place of $p^{\mathrm{near}}_{\theta'}$ and $p^{\mathrm{far}}_{\theta'}$, with context dictating which we are referring to.

Let us walk through the properties of p we need to prove. Regarding its order, we will prove that both

$$q \le O(T/d), \qquad \deg P \le O(T/d);$$

i.e. when θ' is fixed, $p_{\theta'}(x_{L+1}, \dots, x_n)$ has degree at most $O(T/d)$ as an ordinary multivariate real polynomial. Since θ', BAD, NEAR, and FAR are determined by $x_1, \dots, x_L$ alone, it follows that our final polynomial p is a generalized polynomial of order at most $L + O(T/d)$, as needed for the theorem.

Next, we discuss Condition 1, that $p(\boldsymbol{X}) \ge h(\boldsymbol{X})$ always. For the BAD outcomes for $\boldsymbol{x}_1, \dots, \boldsymbol{x}_L$ we have $p(\boldsymbol{X}) = 1 \ge h(\boldsymbol{X})$. For the remaining outcomes, we will have $p(\boldsymbol{X}) \ge h(\boldsymbol{X})$ as required provided that in all cases

$$p_{\theta'}(\boldsymbol{z}) \ge h_{\theta'}(\boldsymbol{z}) \qquad \text{for all } \theta' \text{ and } z \tag{17}$$

where

$$h_{\theta'}(\boldsymbol{z}) = \mathbf{1}\big[\boldsymbol{z} - \theta' \ge 0\big].$$

Next, we discuss Condition 2, the bound $\mathbf{E}[p(\boldsymbol{X}) - h(\boldsymbol{X})] \le \epsilon_1$. It suffices to prove an upper bound of $O(\epsilon_1)$. Recall that

$$\epsilon_1 = \frac{dt \log(dt)}{T}.$$

Note also that we will always $T \le t^2$, since no random variable has stronger hypercontractivity than do Gaussians, for which $T \le 1 + t^2/16$. It follows that we will always have $\epsilon_1 \ge 1/t$. Thus the probability of **BAD**, which is at most $O(1/t^4)$, is much smaller than $O(\epsilon_1)$ and can therefore be neglected. Hence it suffices to show that

$$\mathbf{E}[p_{\theta'}(\boldsymbol{z}) - h_{\theta'}(\boldsymbol{z})] \le O(\epsilon_1) \tag{18}$$

holds in both of the following cases:

Case Near: $|\theta'| \le t\|\boldsymbol{z}\|_2$ and the collection $\{\boldsymbol{x}_{L+1}, \ldots, \boldsymbol{x}_n\}$ is δ-regular.

Case Far: $|\theta'| > t\|\boldsymbol{z}\|_2$.

Next we discuss Condition 3, the bound $\mathbf{Pr}[p(\boldsymbol{X}) > 1 + 1/d^2] \le 2^{-T/d}$. Again, since $p(\boldsymbol{X}) = 1$ for the bad outcomes $x_1, \ldots, x_L$, it suffices to show that

$$\mathbf{Pr}[p_{\theta'}(\boldsymbol{z}) > 1 + 1/d^2] \le 2^{-T/d} \tag{19}$$

holds in both Case a and Case b.

Finally, we discuss the bound $\|p(\boldsymbol{X})\|_{2d} \le 1 + 2/d^2$. We have

$$\mathbf{E}[p(\boldsymbol{X})^{2k}] \le (1+1/d^2)^{2d} + \mathbf{E}[p(\boldsymbol{X})^{2d} \cdot \mathbf{1}\big[p(\boldsymbol{X}) > 1 + 1/d^2\big]] \le 1 + 3/d + \mathbf{E}[p(\boldsymbol{X})^{2d} \cdot \mathbf{1}\big[p(\boldsymbol{X}) > 1 + 1/d^2\big]].$$

If we can show that

$$\mathbf{E}[p(\boldsymbol{X})^{2d} \cdot \mathbf{1}\big[p(\boldsymbol{X}) > 1 + 1/d^2\big]] \le 1/d,$$

then we will have shown

$$\mathbf{E}[p(\boldsymbol{X})^{2d}] \le 1 + 4/d \le (1 + 2/d^2)^{2d},$$

as required. Thus it remains to establish the previous upper bound. Again, since $p(\boldsymbol{X}) = 1$ for the BAD outcomes $x_1, \ldots, x_L$, it suffices to show that

$$\mathbf{E}[p_{\theta'}(\boldsymbol{z})^{2d} \cdot \mathbf{1}\big[p_{\theta'}(\boldsymbol{z}) > 1 + 1/d^2\big]] \le 1/d \tag{20}$$

holds in both Case Near and Case Far.

Summarizing, our goal is to construct univariate polynomials $p_{\theta'}(z)$ of degree at most $O(T/d)$ for each of Case Near and Case Far so that (17), (18), (19), and (20) all hold. We will first handle Case Near, the more difficult case.

F.4.1 Case Near

In this case we have $|\theta'| \le t\|\boldsymbol{z}\|_2$, where $\boldsymbol{z} = \boldsymbol{x}_{L+1} + \cdots + \boldsymbol{x}_n$ is the sum of a δ-regular collection of independent random variables. Our task is to construct a real polynomial $p_{\theta'}(z)$ of degree at most $O(T/d)$ such that bounds (17), (18), (19), and (20) all hold with respect to the function $h_{\theta'}(z) = \mathbf{1}[z - \theta' \ge 0]$.

Given the parameters d and t, choose

$$a = 16C_0 \frac{d \log(td)}{T}, \qquad b = \min(1/d^2, 1/t^4);$$

we have $a < 1$ assuming that the C_1 in our assumption on T is large enough. Let $K = K_{a,b}$ and $P = P_{a,b}$ be the resulting even integer and univariate polynomial from Theorem F.2. Our choice of a was arranged so that

$$K \le \frac{T}{4d}. \tag{21}$$

We will define

$$p_{\theta'}(\boldsymbol{z}) = p_{\text{near}}(\theta', \boldsymbol{z}) = P(\boldsymbol{w}), \qquad \text{where } \boldsymbol{w} = \frac{\boldsymbol{z} - \theta'}{2t\|\boldsymbol{z}\|_2}.$$

Thus $p_{\theta'}(z)$ has degree $K = O(T/d)$ as necessary, and it also satisfies (17), using the property that $P \geq 0$ on $(-\infty, 0]$ and $P \geq 1$ on $[0, \infty)$.

Next we check (20). i.e.,

$$\mathbf{E}[p_{\theta'}(\boldsymbol{z})^{2d} \cdot \mathbf{1}\left[p_{\theta'}(\boldsymbol{z}) > 1 + 1/d^2\right]] \leq 1/d.$$

Since $b \leq 1/d^2$, we have that $p_{\theta'}(\boldsymbol{z}) > 1 + 1/d^2$ only if $|\boldsymbol{w}| \geq 1$.

Also notice that $p_{\theta'}(z) \leq (4w)^K$, it suffice to bound $\mathbf{E}[\mathbf{1}[|w| \geq 1] \cdot (4w)^{2dK}]$ and we will prove a stronger result:

$$\mathbf{E}[(4w)^{2dK} \cdot \mathbf{1}[|w| \geq 1]] \leq 2^{-T}. \tag{22}$$

To see this, since we are in Case Near we have $|\theta'| < t\|\boldsymbol{z}\|_2$. Thus if $|w| \geq 1$, we must have $|\boldsymbol{z}| > t\|\boldsymbol{z}\|_2$. This also implies $|\boldsymbol{z} - \theta'| < 2\,|\boldsymbol{z}|$; hence we have

$$|4\boldsymbol{w}| = 2\frac{|\boldsymbol{z} - \theta'|}{t\|\boldsymbol{z}\|_2} < \frac{4}{t} \cdot \left|\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_2}\right|.$$

Thus we have

$$\begin{aligned}
& \mathbf{E}\left[\mathbf{1}[|w| \geq 1] \cdot (4w)^{2dK}\right] \\
\leq\ & \mathbf{E}\left[\mathbf{1}[|\boldsymbol{z}| > t\|\boldsymbol{z}\|_2] \cdot \left(\frac{4}{t}\right)^{2dK} \left(\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_2}\right)^{2dK}\right] \\
=\ & \left(\frac{4}{t}\right)^{2dK} \cdot \mathbf{E}\left[\mathbf{1}\left[\left|\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_2}\right| > t\right] \cdot \left(\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_2}\right)^{2dK}\right].
\end{aligned} \tag{23}$$

It is easy to check that

$$\mathbf{1}\left[\left|\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_2}\right| > t\right] \cdot \left(\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_2}\right)^{2dK} \leq \left(\frac{\boldsymbol{z}}{t\|\boldsymbol{z}\|_2}\right)^{T} \cdot t^{2dK},$$

using the fact that $2dK \leq T$. Thus we may upper-bound (23) by

$$4^{2dK} t^{-T} \frac{\|\boldsymbol{z}\|_T^T}{\|\boldsymbol{z}\|_2^T} \leq 4^{2dK} t^{-T} (t/4)^T = 4^{2dK - T},$$

where we used the $(T, 2, 4/t)$-hypercontractivity of $\boldsymbol{z}$. Since we have

$$2dK \leq T/2, \tag{24}$$

by virtue of (21), we conclude

$$\mathbf{E}[p_{\theta'}(\boldsymbol{z})^{2d} \cdot \mathbf{1}\left[p_{\theta'}(\boldsymbol{z}) > 1 + 1/d^2\right]] \leq 4^{-T/2} = 2^{-T} \leq 1/d. \tag{25}$$

Let us move on to showing (19) in this Case Near; i.e., upper-bounding $\mathbf{Pr}[p_{\theta'}(\boldsymbol{z}) > 1 + 1/d^2]$. Since $b \leq 1/d^2$, again we have that $p_{\theta'}(\boldsymbol{z}) > 1/d^2$ only if $|\boldsymbol{w}| \geq 1$. But by (22)

$$\mathbf{E}[\mathbf{1}[|\boldsymbol{w}| \geq 1] \cdot (4\boldsymbol{w})^{dK}] \leq 2^{-T},$$

and the left-hand side is clearly an upper bound on $\mathbf{Pr}[|\boldsymbol{w}| \geq 1]$. Thus we have established (19) in Case Near.

Last, we will work to upper bound $\mathbf{E}[p_{\theta'}(\boldsymbol{z}) - h_{\theta'}(\boldsymbol{z})]$ so as to show (18) in Case Near. We analyze three subcases, depending on the magnitude of $\boldsymbol{w}$.

Case i: $-a \le \boldsymbol{w} \le 0$**.** In this case, we upper-bound $p_{\theta'}(\boldsymbol{z}) - h_{\theta'}(\boldsymbol{z})$ simply by 1, and argue that Case i occurs with low probability. Specifically,

$$\mathbf{Pr}[-a \le \boldsymbol{w} \le 0] \le \mathbf{Pr}[|\boldsymbol{w}| \le a] = \mathbf{Pr}[\left|\boldsymbol{z} - \theta'\right| \le 2ta \cdot \|\boldsymbol{z}\|_2].$$

We can upper-bound this probability using the Berry-Esseen Theorem [MZ09, Corollary 4.5]. Since we have δ-regularity of $\boldsymbol{x}_{L+1}, \ldots, \boldsymbol{x}_n$ in Case Near, we get

$$\mathbf{Pr}[\left|\boldsymbol{z} - \theta'\right| \le 2ta \cdot \|\boldsymbol{z}\|_2] \le O(\sqrt{\delta} + ta)$$

By definition of a we have $O(ta) = O(\epsilon_1)$. Thus we conclude for Case i,

$$\mathbf{E}[\mathbf{1}[\text{Case i}] \cdot (p_{\theta'}(\boldsymbol{z}) - h_{\theta'}(\boldsymbol{z}))] \le O(\sqrt{\delta} + \epsilon_1). \tag{26}$$

Case ii: $|\boldsymbol{w}| \le 1$ **but not Case i.** In this case, we have $p_{\theta'}(\boldsymbol{z}) - h_{\theta'}(\boldsymbol{z}) \le b \le 1/t^4$, by construction. Thus

$$\mathbf{E}[\mathbf{1}[\text{Case ii}] \cdot (p_{\theta'}(\boldsymbol{z}) - h_{\theta'}(\boldsymbol{z}))] \le 1/t^4 \le O(\epsilon_1). \tag{27}$$

Case iii: $|\boldsymbol{w}| > 1$**.** I.e., $|\boldsymbol{z} - \theta'| > 2t\|\boldsymbol{z}\|_2$. Notice that $p_{\theta'}(\boldsymbol{z}) - h_{\theta'}(\boldsymbol{z}) \le p_{\theta'}(\boldsymbol{z})$ and therefore

$$\mathbf{E}[\mathbf{1}[\text{Case iii}] \cdot (p_{\theta'}(\boldsymbol{z}) - h_{\theta'}(\boldsymbol{z}))] \le \mathbf{E}[p_{\theta'}(z) \cdot \mathbf{1}[|\boldsymbol{w}| \ge 1]] \le \mathbf{E}\left[\mathbf{1}[|w| \ge 1]\,(4w)^{dK}\right] \le 2^{-T} \le O(\epsilon_1)$$

(the second last inequality is due to (22)).

F.4.2 Case Far

If $\theta < 0$ then $h_{\theta'}$ is almost always 1. As stated, in this case we simply have $p_{\theta'}(z) \equiv 1$. Bounds (17), (19), and (20) become trivial; for (18) it suffices to show

$$\mathbf{Pr}[\boldsymbol{z} \le \theta'] \le \epsilon_1. \tag{28}$$

We will show a stronger statement in the course of handling the case that $\theta' > 0$.

So it remains to handle the $\theta' > 0$ case. As stated, in this case we define

$$p_{\theta'}(z) = p_{\text{far}}(\theta', z) = \left(\frac{z}{\theta'}\right)^q,$$

where

$$q = \left\lfloor \frac{T}{2d} \right\rfloor_{\text{even}},$$

meaning $T/2d$ rounded down to the nearest even integer. Note that $p_{\theta'}(z)$ has the claimed degree bound $O(T/d)$ (treating θ' as a constant). Also note that $p_{\theta'}(\boldsymbol{z}) \ge 1$ if and only if $|\boldsymbol{z}| \ge \theta'$. This establishes (17).

Let's move to (19); we need

$$\mathbf{Pr}[p_{\theta'}(\boldsymbol{z}) \ge 1 + 1/d^2] \le 2^{-T/d}.$$

Certainly

$$p_{\theta'}(\boldsymbol{z}) \ge 1 + 1/d^2 \quad \Rightarrow \quad p_{\theta'}(\boldsymbol{z}) \ge 1 \quad \Rightarrow \quad |\boldsymbol{z}| \ge \left|\theta'\right|.$$

It thus suffices to show

$$\mathbf{Pr}[|\boldsymbol{z}| \geq |\theta'|] \leq 2^{-T/d},$$

which, once shown, also establishes (28), since $2^{-T/d} \ll \epsilon_1$. We will in fact show the stronger statement

$$\mathbf{E}\left[\left(\frac{\boldsymbol{z}}{\theta'}\right)^q\right] \leq 2^{-T/d}. \tag{29}$$

And this stronger statement establishes (18), again because $2^{-T/d} \leq \epsilon_1$.

To prove (29) we appeal to the condition of Case Far, $|\theta'| > t\|\boldsymbol{z}\|_2$. Thus

$$\begin{aligned}
\mathbf{E}\left[\left(\frac{\boldsymbol{z}}{\theta'}\right)^q\right] &\leq \mathbf{E}\left[\left(\frac{\boldsymbol{z}}{t\|\boldsymbol{z}\|_2}\right)^q\right] \\
&\leq \mathbf{E}\left[\left(\frac{\boldsymbol{z}}{t\|\boldsymbol{z}\|_2}\right)^T\right]^{q/T} \qquad \text{(Jensen, since } T/q \geq 1\text{)} \\
&= t^{-q}\left(\frac{\|\boldsymbol{z}\|_T^T}{\|\boldsymbol{z}\|_2^T}\right)^{q/T} \\
&\leq t^{-q}\left(\frac{t}{4}\right)^q \qquad \text{(by } (T,2,4/t)\text{-hypercontractivity of } \boldsymbol{z}\text{)} \\
&= 4^{-q} \quad = \quad 2^{-T/d},
\end{aligned}$$

using the definition of q.

Finally, to prove (20) it certainly suffices to show

$$1/d \geq \mathbf{E}[p_{\theta'}(\boldsymbol{z})^{2d}] = \mathbf{E}\left[\left(\frac{\boldsymbol{z}}{\theta'}\right)^{2d}\right].$$

By repeating the previous inequality with $2d$ in place of q (we still have $T/2d \geq 1$), we can upper-bound the expectation by 4^{-2d}, which is indeed at most $1/d$. This concludes the verification of Case Far, and thus all of Theorem F.4.

G Fooling the Uniform Distribution on the Sphere

In this section, we will show that our PRG can also be used to fool any function of d halfspaces over the uniform distribution on the n dimensional unit sphere; building such a PRG also has an application in derandomizing the hardness of learning reduction in [KS08].

The main idea is to show that the n dimensional Gaussian distribution can be use to fool the uniform distribution on the sphere. Therefore, it suffice to fool the n dimensional Gaussian which is studied in the previous sections (either using the modified MZ generator or k-wise independence).

Specifically, we first show the following connection between the n dimensional Gaussian distribution $\mathcal{N}(0, 1/\sqrt{n})^n$ and the uniform distribution on the n dimensional unit sphere S_{n-1}.

Lemma G.1. *For any $\theta_1, \theta_2, ..\theta_d \in \mathbb{R}$ and $W_1, W_2, ..W_d \in \mathbb{R}^n$ and $h_i(X) = \mathrm{sgn}(W_i \cdot X - \theta_i)$ and $f : \{0,1\}^d \to \{0,1\}$, there is some universal constant C such that*

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{X} \in_u S^{n-1}}[f(h_1(\boldsymbol{X}), .., h_d(\boldsymbol{X}))] - \mathop{\mathbf{E}}_{\boldsymbol{X} \in_u \mathcal{N}(0,1/\sqrt{n})^n}[f(h_1(\boldsymbol{X}), h_2(\boldsymbol{X}) .. h_d(\boldsymbol{X})] \right| \leq \frac{Cd \log n}{n^{1/4}} \tag{30}$$

Proof. Notice that if we choose $x \in_u \mathcal{N}(0, 1/\sqrt{n})^n$, then $\frac{x}{\|x\|_2}$ follows the uniform distribution on the sphere. Therefore, we only need to bound:

$$\Big| \mathop{\mathbf{E}}_{\boldsymbol{X}\in_u\mathcal{N}(0,1/\sqrt{n})^n}(f(h_1(\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_2}),..,h_d(\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_2})) - \mathop{\mathbf{E}}_{x\in_u\mathcal{N}(0,1/\sqrt{n})^n} f(h_1(\boldsymbol{X}),h_2(\boldsymbol{X})..h_d(\boldsymbol{X}))|$$
$$\leq \mathop{\mathbf{Pr}}_{x\in_u\mathcal{N}(0,1/\sqrt{n})^n}\left(f(h_1(\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_2}),..,h_d(\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_2})) \neq f(h_1(\boldsymbol{X}),h_2(\boldsymbol{X})..h_d(\boldsymbol{X}))\right)$$
$$\leq \sum_{i=1}^{d} \mathop{\mathbf{Pr}}_{x\in_u\mathcal{N}(0,1/\sqrt{n})^n}(h_i(\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_2}) \neq h_i(\boldsymbol{X})) \quad (31)$$

By Lemma 6.2 in [MZ09], we know that:

$$\mathop{\mathbf{Pr}}_{\boldsymbol{X}\in_u\mathcal{N}(0,1/\sqrt{n})^n}(h_i(\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_2}) \neq h_i(x)) \leq \frac{C\log n}{n^{1/4}}.$$

Combining above inequality with (31), we prove (30). □

Therefore to fool any function of d halfspaces over the uniform distribution on the n dimensional sphere with accuracy $\Omega(\frac{C\log n}{n^{1/4}})$, it suffice to build a PRG for n dimensional Gaussian distribution with the same accuracy.

G.1 Derandomized hardness of learning intersections of halfspaces

One of the application of above PRG is that we can use it to derandomize the hardness of learning result in [KS08]. In [KS08], Khot and Saket showed that assuming NP≠RP, for any $\epsilon > 0$ and positive integer d, given a set of examples such that there is a intersection of two halfspaces that is consistent with all the examples, it is NP-hard to find a function of any d halfspaces that is consistent with a $1/2 + O(\epsilon)$ fraction of the examples. Our PRGs can be used to derandomize the hardness reduction and obtain the same hardness result assuming NP≠ P.

To see why our PRG works, we need to look into the details of [KS08]. Let us explain in high level why our PRG helps, wihtout entering into the details of the reduction. The hardness of learning result in [KS08] is based on a reduction from a Label Cover instance $\mathcal{L}$ to a distribution $\mathcal{D}_0$ on negative examples and a distribution $\mathcal{D}_1$ on positive examples. Such a reduction would preserve the following two properties:

- (Completeness) if the optimum value of $\mathcal{L}$ is 1, then there is a intersection of two halfspaces $f(x)$ that agrees with all the examples; i.e., $\mathbf{E}_{\mathcal{D}_1}[f(\boldsymbol{X})] = \mathbf{E}_{\mathcal{D}_0}[f(\boldsymbol{X})] + 1$.
- (Soundness) if the optimum value of $\mathcal{L}$ is small, then for any $h(x)$ which is a function of d halfspaces, we have that $\left|\mathbf{E}_{D_0}[h(\boldsymbol{X})] - \mathbf{E}_{D_1}[h(\boldsymbol{X})]\right| = O(\epsilon)$ which implies that $h(x)$ agrees with at most $1/2 + O(\epsilon)$ fraction of the examples.

The D_i for $(i = 0, 1)$ constructed in [KS08] is a mixture of uniform distribution on the sphere located at different center and the number of the different spheres is poly(n), where n is the size of the Label Cover instance. Then by the PRG in this paper, we can derandomize each sphere with some distribution that only has support of size poly(n) to ϵ-fool functions of d halfspaces; and overall we can get distribution $\mathcal{P}_0$ and $\mathcal{P}_1$ with poly(n) support and it has the property that for any function $h(x)$ of l halfspaces, $\mathbf{E}_{\mathcal{D}_i}[f(\boldsymbol{X})] - \mathbf{E}_{\mathcal{P}_i}[f(\boldsymbol{X})] \leq O(\epsilon)$ for $i = 0, 1$. If we replace $\mathcal{D}_i$ with $\mathcal{P}_i$ in the hardness reduction, we still get the soundness guarantee that $|\mathbf{E}_{P_1}[f(\boldsymbol{X})] - \mathbf{E}_{P_0}[f(\boldsymbol{X})]| = O(\epsilon)$.

We also need to verify that the completeness property will hold if we replace $\mathcal{D}_i$ with $\mathcal{P}_i$. If we look into the reduction of [KS08], as long as the distribution $\mathcal{P}_i$ has all its support points on the sphere, the reduction will preserve the completeness property. Therefore, to make the reduction work, we need to build a PRG for functions of d-halfspaces over the uniform distribution on the sphere with the additional property that all the points generated by the PRG are all on the unit sphere as well.

This is also achievable and we summarize the high level idea here. As is shown in Lemma G.1, it suffice to fool functions of d halfspaces over n dimensional Gaussian instead of the uniform distribution on the sphere. In addition, by the proof of Theorem 4.4, if we only want to fool any functions of d ϵ-regular halfspaces, it suffice just to fool uniform distribution on $\{-1/\sqrt{n}, 1/\sqrt{n}\}^n$ instead. For the uniform distribution over $\{-1/\sqrt{n}, 1/\sqrt{n}\}^n$. we know that it can be fooled by PRG with all the support points in $\{-1/\sqrt{n}, 1/\sqrt{n}\}^n$ which is a subset of the unit sphere. To handle the case that d halfspaces are not all ϵ-regular, we can follow the idea of [MZ09] Lemma 6.3 by showing that there exists a set of poly(n) unitary rotations and with high probability that all of the d halfspaces become regular under a rotation randomly chosen from the set.

H Discretizing the Distribution

The first step is to truncate each $\boldsymbol{x}_i$ to lie in the range $(-B, B)$.

Lemma H.1. *Set $B = (nC^2\epsilon^{-1})^{\frac{1}{4}}$. For each $i \in [n]$, let $\boldsymbol{y}_i = \boldsymbol{x}_i \cdot \mathbb{I}(|\boldsymbol{x}_i| < B)$. Define the product random variable $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)$ where the $\boldsymbol{y}_i$s are independent. Then we have*

- $\mathrm{SD}(\boldsymbol{X}, \boldsymbol{Y}) \leq \epsilon$.
- $\mathbf{E}[\boldsymbol{y}_i^2] \geq \frac{1}{2}, \mathbf{E}[\boldsymbol{y}_i^4] \leq C$.

Proof. Note that $\boldsymbol{x}_i = \boldsymbol{y}_i$ when $|\boldsymbol{x}_i| \leq B$ and $\boldsymbol{y}_i = 0$ otherwise. But we have

$$\mathbf{Pr}[|\boldsymbol{x}_i| \geq B] = \mathbf{Pr}[|\boldsymbol{x}_i|^4 \geq B^4] \leq \frac{C}{B^4} = \frac{\epsilon}{nC}.$$

Thus it follows that

$$\mathrm{SD}(\boldsymbol{x}_i, \boldsymbol{y}_i) \leq \frac{\epsilon}{nC} \;\Rightarrow\; \mathrm{SD}(\boldsymbol{X}, \boldsymbol{Y}) \leq \frac{\epsilon}{C} \leq \epsilon.$$

It is clear that $\mathbf{E}[\boldsymbol{y}_i^4] \leq \mathbf{E}[\boldsymbol{x}_i^4] \leq C$. Thus we only need to prove the claim about the two-norm. We have

$$\boldsymbol{x}_i = \boldsymbol{x}_i \cdot \mathbb{I}(|\boldsymbol{x}_i| < B) + \boldsymbol{x}_i \cdot \mathbb{I}(|\boldsymbol{x}_i| \geq B) = \boldsymbol{y}_i + \boldsymbol{x}_i \cdot \mathbb{I}(|\boldsymbol{x}_i| \geq B)$$

from which it follows that

$$\mathbf{E}[\boldsymbol{x}_i^2] = \mathbf{E}[\boldsymbol{y}_i^2] + \mathbf{E}[\boldsymbol{x}_i^2 \cdot \mathbb{I}(|\boldsymbol{x}_i| \geq B)].$$

By the Cauchy-Schwartz inequality, we have

$$\mathbf{E}[\boldsymbol{x}_i^2 \cdot \mathbb{I}(|\boldsymbol{x}_i| \geq B)] \leq \mathbf{E}[\boldsymbol{x}_i^4]^{\frac{1}{2}} (\mathbf{Pr}[|\boldsymbol{x}_i| \geq B)])^{\frac{1}{2}} \leq \sqrt{C}\sqrt{\frac{\epsilon}{nC}} = \sqrt{\frac{\epsilon}{n}} < \frac{1}{2}$$

Hence we have $\mathbf{E}[\boldsymbol{y}_i^2] \geq \frac{1}{2}$.

By a similar argument, one can show that $|\mathbf{E}[\boldsymbol{y}_i]| \leq \frac{\epsilon}{nC}$. □

By suitable shifting and rescaling, we can assume that the distribution satisfies $\mathbf{E}[\boldsymbol{x}_i] = 0, \mathbf{E}[\boldsymbol{x}_i^2] = 1$, $\mathbf{E}[\boldsymbol{x}_i^4] \leq C$ and $|\boldsymbol{x}_i| < B$.

The next step is to suitably discretize the distribution. Assume that the random variable $\boldsymbol{x}_i$ has a cumulative distibution function F_i where $F_i(x) = \mathbf{Pr}[\boldsymbol{x}_i \leq x]$. Since $|\boldsymbol{x}_i| < B$ we have $F(-B) = 0$ and $F(B) = 1$. We will define two *sandwiching* discrete distributions $\boldsymbol{x}_i^\ell$ and $\boldsymbol{x}_i^u$ whose cdfs F_i^ℓ and F_u^i satisfy:

$$F_i^\ell(x) \leq F_i(x) \leq F_i^\ell(x) + \gamma$$
$$F_i^u(x) - \gamma \leq F_i(x) \leq F_i^u(x)$$

where γ is a granularity paramater (which will be chosen as inverse polynomial in n).

Let $g = \frac{1}{\gamma}$. Our goal is to define bucket boundaries $b_0, \ldots, b_g$ by picking b_k that stisfy $F_i(b_k) = k\gamma$.

Definition H.2. *For $k \in \{0, \ldots, g\}$, let b_k be the smallest $x \in [-B, B]$ so that $F_i(x) \geq k\gamma$.*

We can sample $\boldsymbol{x}_i$ by first picking a bucket $k \in \{0, \ldots, g-1\}$ and then sampling from this bucket according to the suitable conditional distribution, resulting in $\boldsymbol{x}_i \in [b_k, b_{k+1}]$.

We now define the sandwiching distributions:

Definition H.3. *The random variable $\boldsymbol{x}_i^\ell$ is uniformly distributed on $\{b_0, \ldots, b_{g-1}\}$ while $\boldsymbol{x}_i^u$ the uniform distributed on $\{b_1, \ldots, b_g\}$. We define the family $\mathcal{F}$ of 2^n product distributions on $\mathbb{R}^n$ where each co-ordinate is distributed independently according to $\boldsymbol{x}_i^\ell$ or $\boldsymbol{x}_i^u$.*

It follows that $\mathrm{SD}(\boldsymbol{x}_i^\ell, \boldsymbol{x}_i^u) \leq \gamma$. Hence if we take any pair of variables $\boldsymbol{Y}, \boldsymbol{Z}$ from $\mathcal{F}$, by the union bound we have $\mathrm{SD}(\boldsymbol{Y}, \boldsymbol{Z}) \leq \gamma n$. The following lemma allows us to reduce the problem of fooling halfspaces under the distribution $\boldsymbol{X}$ to the problem of fooling a single distribution from the family $\mathcal{F}$.

Lemma H.4. *Let $h : \mathbb{R}^n \to \{-1, 1\}$ for $i \in [k]$ be a halfspace and let $\boldsymbol{Y} \in \mathcal{F}$. Then*

$$|\,\mathbf{E}[h(\boldsymbol{X})] - \mathbf{E}[h(\boldsymbol{Y})]| \leq 4\gamma n.$$

Proof. We will pick sandwiching distributions $\boldsymbol{Y}^\ell = (\boldsymbol{y}_1^\ell, \ldots, \boldsymbol{y}_n^\ell)$ and $\boldsymbol{Y}^u = (\boldsymbol{y}_1^u, \ldots, \boldsymbol{y}_n^u)$ from $\mathcal{F}$ (depending on the halfspace h) and construct a coupling of the three distributions $\boldsymbol{Y}^\ell, \boldsymbol{X}$ and $\boldsymbol{Y}^u$ so that

$$h(\boldsymbol{Y}^\ell) \leq h(\boldsymbol{X}) \leq h(\boldsymbol{Y}^u). \tag{32}$$

Let $h(x) = \mathrm{sgn}(\sum_i w_i \boldsymbol{x}_i - \theta)$. If $w_i \geq 0$ for all i, then we set

$$\boldsymbol{y}_i^\ell = \boldsymbol{x}_i^\ell, \quad \boldsymbol{y}_i^u = \boldsymbol{x}_i^u.$$

Whereas if $w_i < 0$, then we set

$$\boldsymbol{y}_i^\ell = \boldsymbol{x}_i^u, \quad \boldsymbol{y}_i^u = \boldsymbol{x}_i^\ell.$$

Next we describe the coupling, co-ordinate by co-ordinate. Fix co-ordinate i. Pick $k \in \{0, \ldots, g-1\}$ at random. Set $\boldsymbol{x}_i^\ell = b_k$ and $\boldsymbol{x}_i^u = b_{k+1}$. We now set the random variables $\boldsymbol{y}_i, \boldsymbol{y}_i\ell$ and $\boldsymbol{y}_i^u$ to be eiher $\boldsymbol{x}_i^\ell$ or $\boldsymbol{x}_i^u$, based on their defintion. We pick $\boldsymbol{x}_i$ conditioned on the k^{th} bucket, so that $b_k \leq \boldsymbol{x}_i \leq b_{k+1}$. It follows that

$$w_i \boldsymbol{y}_i^\ell \leq w_i \boldsymbol{x}_i \leq w_i \boldsymbol{y}_i^u$$

and hence

$$\sum_i w_i \boldsymbol{y}_i^\ell \leq \sum_i w_i \boldsymbol{x}_i \leq \sum_i w_i \boldsymbol{y}_i^u$$

which implies Equation 32.

Since a halfspace is a statistical test, we have

$$\mathbf{Pr}[h(\boldsymbol{X}) \neq h(\boldsymbol{Y}^u)] \leq \mathbf{Pr}[h(\boldsymbol{Y}^\ell) \neq h(\boldsymbol{Y}^u)] \leq \mathrm{SD}(\boldsymbol{Y}^u, \boldsymbol{Y}^\ell) \leq \gamma n. \tag{33}$$

If we replace $\boldsymbol{Y}^u$ with $\boldsymbol{Y} \in \mathcal{F}$, we have

$$\mathbf{Pr}[h(\boldsymbol{X}) \neq h(\boldsymbol{Y})] \leq \mathbf{Pr}[h(\boldsymbol{X}) \neq h(\boldsymbol{Y}^u)]| + \mathbf{Pr}[h(\boldsymbol{Y}) \neq h(\boldsymbol{Y}^u)]| \leq 2\gamma n$$

where we use Equations 33 and the fact that $\mathrm{SD}(\boldsymbol{Y}, \boldsymbol{Y}^u) \leq \gamma n$. The claim follows since $h(\boldsymbol{X})$ and $h(\boldsymbol{Y})$ take values over $\{-1, 1\}$. □

This lemma extends to fooling functions of halfspaces.

Lemma H.5. *Let $f : \mathbb{R}^n \to \{-1, 1\}$ be a function of d halfsapces $h_i : \mathbb{R}^n \to \{-1, 1\}$ given by $f = g(h_1, \ldots, h_d)$ where $g : \{-1, 1\}^k \to \{-1, 1\}$. Then for any $Y \in \mathcal{F}$,*

$$|\,\mathbf{E}[f(\boldsymbol{X})] - \mathbf{E}[f(\boldsymbol{Y})]| \leq 4\gamma dn.$$

Proof. We consider the same coupling used in Lemma H.4. We have

$$\mathbf{Pr}[g(\boldsymbol{X}) \neq g(\boldsymbol{Y})] \leq \mathbf{Pr}[(h_i(\boldsymbol{X}), \ldots, h_d(\boldsymbol{X})) \neq (h_1(\boldsymbol{Y}), \ldots, h_d(\boldsymbol{Y}))] \leq \sum_i \mathbf{Pr}[h_i(\boldsymbol{X}) \neq h_i(\boldsymbol{Y})] \leq 2\gamma dn.$$

The claim now follows since g is Boolean valued. □

Finally, we need to show that for a suitable choice of γ, the expectation and the second and fourth moments of $\boldsymbol{x}_i^\ell$ and $\boldsymbol{x}_i^u$ are nearly the same as those of $\boldsymbol{x}_i$. We prove the claim for the fourth moment, the other arguments are similar.

Lemma H.6. *We have*

$$|\,\mathbf{E}[(\boldsymbol{x}_i)^4] - \mathbf{E}[(\boldsymbol{x}_i^u)^4]| \leq 2B^4\gamma, \quad |\,\mathbf{E}[(\boldsymbol{x}_i)^4] - \mathbf{E}[(\boldsymbol{x}_i^\ell)^4]| \leq 2B^4\gamma$$

Proof. It is clear that

$$\mathbf{E}[(\boldsymbol{x}_i^\ell)^4] = \gamma(\sum_{k=0}^{g-1} b_k^4), \quad \mathbf{E}[(\boldsymbol{x}_i^u)^4] = \gamma(\sum_{k=1}^{g} b_k^4).$$

Our goal is to compare these with the 4^{th} moment of $\boldsymbol{x}_i$. The contribution of the k^{th} bucket to $\mathbf{E}[\boldsymbol{x}_i^4]$ can be upper bounded by $\gamma \max(b_k^4, b_{k+1}^4)$ and lower bounded by $\gamma \min(b_k^4, b_{k+1}^4)$. Hence

$$\gamma \sum_{k=0}^{g-1} \min(b_k^4, b_{k+1}^4) \leq \mathbf{E}[\boldsymbol{x}_i^4] \leq \gamma \sum_{k=0}^{g-1} \max(b_k^4, b_{k+1}^4).$$

By case analysis, the sequence $\max(b_k^4, b_{k+1}^4)$ takes on g distinct values from $\{b_0, \ldots, b_g\}$. Similarly, $\min(b_k^4, b_{k+1}^4)$ can take some value twice but every other value at most once. Hence both the upper and lower bounds are within $2B^4\gamma$ of both $\mathbf{E}[(\boldsymbol{x}_i^\ell)^4]$ and $\mathbf{E}[(\boldsymbol{x}_i^u)^4]$. □

A similar argument shows that the second moment changes by at most $2B^2\gamma$ and the expectation by $2B\gamma$. We pick $\gamma < \frac{\epsilon}{2nB^4} = O(\frac{\epsilon^2}{n^2C^2})$, which is of the form 2^{-s} for some integer s. We have $2^s < O(\frac{n^2C^2}{\epsilon^2})$ hence $s = \log(n^2C^2/\epsilon^2) + O(1)$. To sample from $\boldsymbol{x}_i^\ell$ (X_i^u), we pick a random bit-string of length s, treat it as a number $j \in \{0, g-1\}$, and output b_j (b_{j+1}).

Finally we rescale and shift, so that we again have $\mathbf{E}[\boldsymbol{y}_i] = 0, \mathbf{E}[\boldsymbol{y}_i^2 = 1]$ and $\mathbf{E}[\boldsymbol{y}_i^4] \leq C$.

I Hypercontractive Random Variables

We list some basic properties of η-HC random variables, all of which have elementary proofs:

Fact I.1. *[KS88, MOO05, Wol06a, Wol06b]*

1. *If $\boldsymbol{x}$ is η-HC then it is also η'-HC for all $\eta' < \eta$.*
2. *If $\boldsymbol{x}$ is η-HC then $\boldsymbol{x}$ is centered, $\mathbf{E}[\boldsymbol{x}] = 0$.*
3. *If $\boldsymbol{x}$ is η-HC then $\mathbf{E}[\boldsymbol{x}^4] \leq (1/\eta)^4 \mathbf{E}[\boldsymbol{x}^2]^2$.*
4. *Conversely, if $\mathbf{E}[\boldsymbol{x}] = 0$ and $\mathbf{E}[\boldsymbol{x}^4] \leq (1/\eta)^4 \mathbf{E}[\boldsymbol{x}^2]^2$, then $\boldsymbol{x}$ is $(\eta/2\sqrt{3})$-HC. If $\boldsymbol{x}$ is also* symmetric *(i.e., $-\boldsymbol{x}$ has the same distribution as $\boldsymbol{x}$) then $\boldsymbol{X}$ is $\min(\eta, 1/\sqrt{3})$-HC.*
5. *If $\boldsymbol{x}$ is ± 1 with probability $1/2$ each, then $\boldsymbol{x}$ is $(1/\sqrt{3})$-HC. The same is true if $\boldsymbol{x}$ has the standard Gaussian distribution or the uniform distribution on $[-1, 1]$.*
6. *If $\boldsymbol{x}$ is η-HC then in fact $\eta \leq 1/\sqrt{3}$.*
7. *If $\boldsymbol{x}$ is a centered discrete random variable and $\alpha \leq 1/2$ is the least nonzero value of $\boldsymbol{x}$'s probability mass function, then $\boldsymbol{x}$ is η-HC for $\eta = \alpha^{1/4}/2\sqrt{3}$.*
8. *If $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent η-HC random variables, then so is $c_1\boldsymbol{x}_1 + \cdots c_n\boldsymbol{x}_n$ for any real constants $c_1, \ldots, c_n$, not all 0. (Indeed, 4-wise independence suffices.)*
9. *If $\boldsymbol{x}$ is η-HC, and $\boldsymbol{y}$ is a random variable with the same rth moments as $\boldsymbol{x}$ for all $r = 0, 1, 2, 3, 4$, then $\boldsymbol{y}$ is also η-HC.*

We now prove the Concentration bound (Proposition 3.5) for η-HC variables.

Proof. Apply Markov to the event "$\boldsymbol{x}^4 \geq t^4 \mathbf{E}[\boldsymbol{x}^2]^2$". □

Next we prove the Anti-Concentration bound (Proposition 3.5) for η-HC variables, which is more involved.

Proof. By scaling $\boldsymbol{x}$ it suffices to consider the case $\|\boldsymbol{x}\|_2 = 1$. Consider the random variable $\boldsymbol{y} = (\boldsymbol{x} - \theta)^2$. We have

$$\mathbf{E}[\boldsymbol{y}] = \mathbf{E}[\boldsymbol{x}^2] - 2\theta \mathbf{E}[\boldsymbol{x}] + \theta^2 = 1 + \theta^2,$$

$$\mathbf{E}[\boldsymbol{y}^2] = \eta^{-4} \mathbf{E}[(-\eta\theta + \eta\boldsymbol{x})^4] \leq \eta^{-4} \mathbf{E}[(-\eta\theta + \boldsymbol{x})^2]^2 = \eta^{-4}(1 + \eta^2\theta^2)^2 = (\eta^{-2} + \theta^2)^2,$$

where we used the fact that $\boldsymbol{x}$ is η-HC in the second calculation (and then used the first calculation again). We now apply the Paley-Zygmund inequality (with parameter $0 < t^2/(1 + \theta^2) < 1$):

$$\begin{aligned}\mathbf{Pr}[|\boldsymbol{x} - \theta| > t] = \mathbf{Pr}[\boldsymbol{y} > t^2] = \mathbf{Pr}\left[\boldsymbol{y} > \frac{t^2}{1+\theta^2}\mathbf{E}[\boldsymbol{y}]\right] &\geq \left(1 - \frac{t^2}{1+\theta^2}\right)^2 \frac{\mathbf{E}[\boldsymbol{y}]^2}{\mathbf{E}[\boldsymbol{y}^2]} \\ &\geq \left(1 - \frac{t^2}{1+\theta^2}\right)^2 \frac{(1+\theta^2)^2}{(\eta^{-2}+\theta^2)^2} = \left(\frac{\eta^2(1-t^2) + \eta^2\theta^2}{1+\eta^2\theta^2}\right)^2. \quad (34)\end{aligned}$$

Treat η and t as fixed and θ as varying. Writing $u = \eta^2(1 - t^2)$, we have $0 < u < 1$; hence the fraction $(u + \eta^2\theta^2)/(1 + \eta^2\theta^2)$ appearing in (34) is positive and increasing as $\eta^2\theta^2$ increases. Thus it is minimized when $\theta = 0$; substituting this into (34) gives the claimed lower bound. □