2012 IEEE 53rd Annual Symposium on Foundations of Computer Science

A Structure Theorem for Poorly Anticoncentrated Gaussian Chaoses and Applications to the Study of Polynomial Threshold Functions*

Daniel M. Kane
Department of Mathematics
Stanford University
Stanford, CA
Email: dankane@math.stanford.edu

***Abstract*—We prove a structural result for degree-d polynomials. In particular, we show that any degree-d polynomial, p can be approximated by another polynomial, p_0, which can be decomposed as some function of polynomials $q_1, \ldots, q_m$ with q_i normalized and $m = O_d(1)$, so that if X is a Gaussian random variable, the probability distribution on $(q_1(X), \ldots, q_m(X))$ does not have too much mass in any small box.**

Using this result, we prove improved versions of a number of results about polynomial threshold functions, including producing better pseudorandom generators, obtaining a better invariance principle, and proving improved bounds on noise sensitivity.

***Keywords*-Polynomials, Gaussian distributions, Threshold logic functions**

I. Introduction

A. Polynomial Threshold Functions

A polynomial threshold function (PTF) is a function of the form $f(X) = \mathrm{sgn}(p(X))$ for some polynomial $p(X)$. We say that f is a degree-d polynomial threshold function of p is of degree at most d. Polynomial threshold functions are a fundamental class of functions with applications to many fields such as circuit complexity [1], communication complexity [23] and learning theory [17].

We present a new structural result for degree-d polynomials that allows us to obtain improved versions of a number of results relating to polynomial threshold functions. Our result allows us to define a new notion of regularity for polynomials for which we can prove an improved version of the Invariance Principle of [20]. We also obtain a regularity lemma (along the lines of the main theorem of [8]) for this new notion of regularity. Although neither of these theorems will be directly comparable to their classical versions (due to the different notions of regularity), the combination of our regularity lemma and Invariance Principle produces a marked improvement over previous work. These results in turn allow us to prove better bounds on the noise sensitivity of polynomial threshold functions (improving on the bounds of [6] for fixed $d \geq 3$) and provide us with an improved analysis of the pseudorandom generators of [19] and [13].

*Full version of the paper is available at http://arxiv.org/abs/1204.0543

B. Anticoncentration and Diffuse Decompositions

Many of the analytic techniques for dealing with polynomial threshold functions (most notably the replacement method (see [9], [18])) work well for dealing with smooth functions of polynomials. In order to get these techniques to yield useful results for threshold functions, it is often necessary to approximate the threshold function by a smooth one. In order to obtain a good approximation, one needs to know that with high probability that the value of $p(X)$ does not lie too close to zero. Results of this form have become known as a *anticoncentration* results. Such a result was proved by Carbery and Wright in [3]. They prove that for p a degree-d polynomial and X a random Gaussian that

$$\Pr(|p(X)| \leq \epsilon |p|_2) = O(d\epsilon^{1/d}). \quad (1)$$

This bound has proved to be an essential component of many theorems about polynomial threshold functions. Unfortunately, the presence of the $\epsilon^{1/d}$ term above often leads to results that have poor ϵ-dependence for moderately large values of d, and the lack of a stronger form of Equation (1) has proved to be a bottleneck for a number of results on polynomial threshold functions. One might hope to overcome this difficulty by proving an improved version of Equation (1). In particular, a generic polynomial p can be thought of as a sum of largely uncorrelated monomials, and thus, one might expect that $p(X)$ be Gaussian distributed. Thus, while Equation (1) tells us little more than the fact that the distribution of $p(X)$ has no point masses, one might expect the stronger condition that $p(X)$ has bounded probability density function to hold. Unfortunately, this is not the case in general. For example, if p is the d^{th} power of a linear polynomial, the probability that $|p(X)| < \epsilon$ will in fact be proportional to $\epsilon^{1/d}$. On the other hand, this counterexample is not as great an obstacle as it first appears to be. While, in this case, the probability distribution of $p(X)$ does have poor analytic properties, this is because p can be written as a composition of a well-behaved (in this case linear) polynomial, and a simple, yet poorly-behaved polynomial (the d^{th} power). The fact that the value of p is governed by the value of this linear polynomial

0272-5428/12 $26.00 © 2012 IEEE
DOI 10.1109/FOCS.2012.52

will allow one to overcome the difficulties posed by poor anticoncentration in most applications.

In fact, this principle applies more generally. In particular, as we shall show, any polynomial p may be approximately represented as the composition of a simple polynomial (i.e. a polynomial dependent on few input variables) and an analytically well-behaved polynomial (i.e. one with good anticoncentration properties). In order to make this claim rigorous, we provide the following definitions:

Definition. *Given a degree-d polynomial $p : \mathbb{R}^n \to \mathbb{R}$, we say that a set of polynomials $(h, q_1, \ldots, q_m)$ is a* decomposition of p of size m *if $q_i : \mathbb{R}^n \to \mathbb{R}$, and $h : \mathbb{R}^m \to \mathbb{R}$ are polynomials so that*

- $p(x) = h(q_1(x), \ldots, q_m(x))$.
- *For every monomial $c\prod x_i^{a_i}$ appearing in h, we have that $\sum a_1 \deg(q_i) \leq d$.*

In other words, a decomposition of p is a way of writing p as a composition of a simple polynomial, h, with another polynomial $Q = (q_1, \ldots, q_m)$. The second condition above tells us that if we expanded out the polynomial $h(q_1(x), \ldots, q_m(x))$, we would never have to write any terms of degree more than d.

Definition. *We say that a tuple of polynomials $(q_1, \ldots, q_m) : \mathbb{R}^n \to \mathbb{R}^m$, is an (ϵ, N)-*diffuse set *if for every $(a_1, \ldots, a_m) \in \mathbb{R}^m$ and Gaussian random variable X we have that*

$$Pr_X(|q_i(X) - a_i| \leq \epsilon \text{ for all } i) \leq \epsilon^m N,$$

and $\mathbb{E}[|q_i(X)|^2] \leq 1$ for all i.

We note that while an anticoncentration result need only tell us that the probability distribution of $p(X)$ contains no point masses, an (ϵ, N)-diffuse set of polynomials will have the probability density function of the vector $(q_1(X), \ldots, q_m(X))$ average no more than N on any small box. This is a much stronger notion of "analytically well-behaved". Combining the two definitions above, we define the notion of a *diffuse decomposition*.

Definition. *Given a polynomial p we say that $(h, q_1, \ldots, q_m)$ is an (ϵ, N)-*diffuse decomposition of p of size m *if $(h, q_1, \ldots, q_m)$ is a decomposition of p of size m and if $(q_1, \ldots, q_m)$ is an (ϵ, N)-diffuse set.*

It is not obvious that diffuse decompositions should exist in any useful cases. The main result of this paper will be to show that not only can any polynomial be approximated by a polynomial with a diffuse decomposition, but that the parameters of this decomposition are sufficient for use in a wide variety of applications.

Theorem 1 (The Diffuse Decomposition Theorem)**.** *Let ϵ, c and N be positive real numbers and d a positive integer. Let $p(X)$ be a degree-d polynomial. Then there exists a degree-d polynomial p_0 with $|p - p_0|_2 < O_{c,d,N}(\epsilon^N)|p|_2$ so that p_0 has an $(\epsilon, \epsilon^{-c})$-diffuse decomposition of size at most $O_{c,d,N}(1)$.*

It should be noted that if p is a polynomial with a diffuse decomposition, $(h, q_1, \ldots, q_m)$, then the distribution of $p(X)$ will be determined in large part by the polynomial h, as the distribution for $(q_1(X), \ldots, q_m(X))$ is controlled by the diffuse property. Thus, Theorem 1 may be thought of as a structural result for Gaussian chaoses. Theorem 1 may also be thought of as a continuous analogue of theorems of Green-Tao ([11]) and Kaufman-Lovett ([16]) which say that a polynomial over a finite field can be decomposed in terms of lower degree polynomials whose output distributions on random inputs are close to uniform.

Remark. *The bound on the size of the decomposition in Theorem 1 is effective, but may be quite large. Working through the details of the proof would lead to a bound of $A(d + O(1), N/c)$, where $A(m, n)$ is the Ackermann function. The author believes that a polynomial in (dN/c) should be sufficient, but does not know of a proof for this improved bound.*

C. Applications of the Main Theorem

Theorem 1 has several applications that we will discuss. The existence of diffuse decompositions allows us to make better use of the replacement method and achieve a tighter analysis of the pseudorandom generators for polynomial threshold functions presented in [13] and [19]. We can also use this theory to improve on the Invariance Principle of [20]. In particular, we come up with a new notion of regularity for a polynomial, so that for highly regular polynomials their evaluation at random Gaussian variables and at random Bernoulli variables are close in cdf distance. We then show that an arbitrary polynomial can be written as a decision tree of small depth almost all of whose leaves are either regular or have constant sign with high probability. These theorems of ours will produce a qualitative improvement over the analogous theorems of [20] and [8]. Finally, we make use of this technology to prove new bounds on the noise sensitivity of polynomial threshold functions. Each of these applications will be discussed in more detail in the relevant section of this paper.

II. Basic Results and Notation

In this section, we introduce notation that will be used throughout the rest of the paper and state some results that shall be used throughout. We use $|p|_t$ to denote the L^t norm of a function p with respect to the Gaussian measure. Namely,

$$|p|_t := \mathbb{E}_{X \sim \mathcal{N}}[|p(X)|^t]^{1/t}.$$

We use the asymptotic notation $O(X)$ to denote a quantity bounded above by an absolute constant times X, and $O_{A,B,C}(X)$ to denote a quantity bounded above by X times some function depending only on A, B and C. We

use $\tilde{O}(X)$ to denote a quantity that is bounded above by X times some polynomial in $\log(X)$, and use $\tilde{O}_{A,B,C}(X)$ when the coefficients or degree of this polynomial depend on A, B, C. We will also assume throughout that all random variables are independent n-dimensional random Gaussians unless otherwise stated.

We will also need to make use of the notion of strong anticoncentration first mentioned by the author in [13]. This is the idea that a polynomial at a random input is probably not much small than the value of its derivative at the same point. As a key technical lemma we will need a generalization of this idea to vector-valued functions. In particular, we will make use of the following Proposition:

Proposition 2 (Strong Anticoncentration)**.** *Let $p_1, \ldots, p_k$ be polynomials of bounded degree, and let X be a random Gaussian. Then*

$$Pr\left(\prod_{i=1}^{k} |p_i(X)| \leq \epsilon \left| \bigwedge_{i=1}^{k} \nabla p_i(X) \right|\right) = \tilde{O}_k(\epsilon)$$

where $\bigwedge_{i=1}^{k} \nabla p_i(X)$ above is the wedge product of the gradients of the p_i at X.

This Proposition says that after excepting events of small probability, the $p_i(X)$ can only be simultaneously small if their gradients are approximately linearly dependent.

III. The Decomposition Theorem

In this section, we provide a sketch of the proof of Theorem 1. On a high level, we begin with the trivial decomposition of p as $\mathrm{Id} \circ p$ and iteratively refine this decomposition until we have a diffuse one. If we have some decomposition $(h, q_1, \ldots, q_m)$, where $(q_1, \ldots, q_m)$ does not form an $(\epsilon, \epsilon^{-c})$-diffuse set, there must be constants a_i so that the $|q_i(X) - a_i|$ are simultaneously small with abnormally large probability. Using Strong Anticoncentration, we see that this implies that some linear combination of the q_i has a small gradient with reasonable probability.

That the entire gradient vector of a polynomial is small turns out to be a very strong condition, in particular we claim:

Proposition 3 (Small Derivative Proposition)**.** *Let p be a degree-d polynomial with $|p|_2 \leq 1$. Assume that for some $\epsilon, N, c > 0$ and for X a standard Gaussian that*

$$Pr(|\nabla p(X)|_2 \leq \epsilon) \geq \epsilon^N.$$

Then there exist $O_{c,d,N}(1)$ polynomials a_i,b_i,c of degree less than d so that

$$\left| p(X) - \sum_i a_i(X) b_i(X) - c(X) \right|_2 = O_{c,d,N}(\epsilon^{1-c}).$$

Furthermore, such polynomials can be found so that for each i, $\deg(a_i) + \deg(b_i) = d$.

In other words, a polynomial with a non-trivial probability of having a small gradient can always be approximately decomposed in terms of lower degree polynomials. Hence, unless our decomposition is already diffuse, we can further decompose at least one of the q_i in terms of lower-degree polynomials. This allows us to produce a finer decomposition of p. To show that this process terminates, we make use of an ordinal monovariant depending on the number of q_i of each degree. Using these ideas, we present the sketch of a proof of Theorem 1 given Proposition 3.

Proof of Theorem 1 given Proposition 3 (sketch): In this proof we will overlook some minor technical complications that are dealt with in the full paper. We start by making some simplifying assumptions. We assume that N and c^{-1} are both integers. We assume that ϵ is sufficiently small. Additionally, we take the normalization of p so that $|p|_2 = 1$.

To formalize the ideas described above, we define a *partial decomposition* of p to be a sequence of polynomials $(h, q_1, \ldots, q_m)$ and non-negative integers $e_1, \ldots, e_m$, with

$$e_i < 4 \cdot 3^i (N+1) c^{-1}$$

so that $|h|_2 = O(\epsilon^{-1})$, $|q_i|_2 = 1$ for each i, and so that p is within $O(\epsilon^N)$ of

$$h\left(\epsilon^{e_1 c/(2 \cdot 3^1)} q_1(x), \ldots, \epsilon^{e_m c/(2 \cdot 3^m)} q_m(x)\right).$$

To each partial decomposition, we assign a weight given by the vector of non-negative integers $v = (n_d, \ldots, n_1)$ where

$$n_i = \sum_{j : \deg(q_j) = i} 4 \cdot 3^j (N+1) c^{-1} - e_j.$$

We claim that every partial decomposition either gives us the desired diffuse decomposition, or can be used to construct a more refined partial decomposition. In particular we claim:

Claim. *For any partial decomposition (h, q_i, e_i), either $(q_1, \ldots, q_m)$ is an $(\epsilon, \epsilon^{-c})$-diffuse set, or there exists another partial decomposition whose weight is a vector that is strictly smaller in lexicographic order.*

To prove this claim, assume that $(q_1, \ldots, q_m)$ is not $(\epsilon, \epsilon^{-c})$-diffuse. Unless it is the case that $\deg(q_1) \geq \deg(q_2) \geq \ldots \geq \deg(q_m)$, an appropriate reordering of the q_i will lead to a partial decomposition of strictly smaller weight. We may thus assume that the degrees of the q_i are ordered as above. Similarly, we may assume that $e_i \leq 2 \cdot 3^i (N+1) c^{-1}$ for each i, since otherwise removing q_i entirely from our decomposition will introduce an error on the order of ϵ^N and decrease the weight of the decomposition.

By assumption, there exist real numbers a_i so that

$$\Pr(|q_i(X) - a_i| \leq \epsilon \text{ for all } i) > \epsilon^{m-c}.$$

By Strong Anticoncentration, this implies that with proba-

bility at least ϵ^m we have that

$$\left|\bigwedge_{i=1}^{m} \nabla q_i(X)\right|_2 \leq \epsilon^{c(1-3^{-m})}. \quad (2)$$

Let $V_i(X)$ be the difference between $\nabla q_i(X)$ and it's projection onto the space spanned by the $\nabla q_j(X)$ for $j > i$. Since

$$\left|\bigwedge_{i=1}^{m} \nabla q_i(X)\right|_2 = \prod_{i=1}^{m} |V_i(X)|_2,$$

Equation (2) implies that $|V_i(X)|_2 \leq \epsilon^{2c/3^i}$ for some i. There therefore exists some particular i for which this holds with probability at least ϵ^m/m.

For each X we have that

$$V_i(X) = \sum_{j=i}^{m} \beta_j \nabla q_j(X)$$

for some constants β_j with $\beta_i = 1$. Furthermore, if $|V_j| > \epsilon^{2c/3^j}$ for all $j > i$, it is not hard to show that $|\beta_j| \leq \epsilon^{-c/3^i}$. To avoid some technical complications, we will consider only the case where $|\beta_j| \leq 1$ for all j. Since a randomly chosen sequence of real numbers γ_j will have reasonable probability of approximating the β_j, we have that with probability $\epsilon^{O(m)}$ that

$$\left|\nabla q_i(X) + \sum_{j>i}^{m} \gamma_j \nabla q_j(X)\right|_2 \leq 2\epsilon^{2c/3^i}.$$

It must therefore be the case that for some specific choice of γ_j that the above holds with probability $\epsilon^{O(m)}$. Thus, if we let

$$Q(x) = q_i(x) + \sum_{j>i} \gamma_j q_j(x),$$

then

$$\Pr\left(|\nabla Q(X)|_2 \leq 2\epsilon^{2c/3^i}\right) > \epsilon^{O(m)}.$$

Applying the Small Derivative Proposition to Q, we find that Q may be rewritten as

$$Q(x) = \sum A_i(x)B_i(x) + C(x) + Q'(x)$$

with $\deg(A_i), \deg(B_i), \deg(C) < \deg(Q)$, $\deg(Q') = \deg(Q)$, and $|Q'|_2 = O_{c,d,m}(\epsilon^{2c/3^i})$. This means that we can write q_i as

$$q_i(x) = \sum A_i(x)B_i(x) + C(x) + Q'(x) - \sum_{j>i} \gamma_i q_j(x).$$

This allows us to construct a new partial decomposition of p by letting the normalized version of Q' be the new q_i, introducing A_i, B_i and C as new $q_j's$, and using the above identity to rewrite h. Since $|Q'|_2$ is small, we can construct this new partial decomposition with an e_i strictly larger than before. Since the other e_j remain the same and since the new q_j all have degree strictly smaller than q_i, the weight of our new partial decomposition is strictly smaller than before. This completes the proof of our claim.

To prove our theorem, we begin with the trivial partial decomposition, D_0, given by

$$h(x) = x, q_1 = p, e_1 = 0.$$

Iteratively applying our claim, we produce a sequence of partial decompositions, $D_0, D_1, \ldots$ each with weight strictly smaller than the last until we end up with a partial decomposition $D_n = (h, q_i, e_i)$ with $(q_1, \ldots, q_m)$ an $(\epsilon, \epsilon^{-c})$-diffuse set. Since the set of vectors of d non-negative integers is well ordered under the lexicographic ordering, we are guaranteed that this process will eventually terminate with such a D_n. On the other hand, letting

$$h'(x_1, \ldots, x_m) = h\left(\epsilon^{e_1 c/(2\cdot 3^1)} x_1, \ldots, \epsilon^{e_m c/(2\cdot 3^m)} x_m\right),$$

we have that $(h', q_1, \ldots, q_m)$ is an $(\epsilon, \epsilon^{-c})$-diffuse decomposition of an appropriate approximation to p. Since it is possible to bound the complexity of D_{i+1} in the above sequence in terms of c, d, N and the complexity of D_i, it is not hard to show that we can find an absolute bound on the size of the resulting diffuse decomposition in terms of c, d and N. ∎

The proof of the Small Derivative Proposition is quite complicated. As a first step, one notes that having a small derivative with reasonable probability implies that higher order derivatives are also small. In particular, we show:

Proposition 4 (Higher Derivatives Proposition)**.** *Let $c, N > 0$ be real numbers and d a positive integer and $\epsilon > 0$ sufficiently small. Suppose that p is a degree-d polynomial so that for a random Gaussian X*

$$Pr_X(|\nabla p(X)|_2 < \epsilon) > \epsilon^N.$$

Then we have for random Gaussians X and Y that

$$Pr_{X,Y}(|\nabla D_Y p(X)|_2 < \epsilon^{1-c}) > \epsilon^{O_{N,c,d}(1)},$$

where $D_Y p(X)$ is the directional derivative of p at X in the direction of Y.

This Proposition follows fairly easily from Strong Anticoncentration. In particular, if there is a reasonable probability that all of the partial derivatives of p are simultaneously small, then by Strong Anticoncentration, their derivatives must be approximately linearly dependent. In particular, one can show that with reasonable probability over X that the Hessian of p at X is approximately of low rank. When this is the case, $|\nabla D_Y p(X)| = (H(p)(X)) \cdot Y$ is small with reasonable probability.

The conclusion of Proposition 4 tells us that with probability at least $\epsilon^{O_{N,c,d}(1)}$ over Y, the polynomial $q = D_Y p$ satisfies

$$\Pr(|\nabla q(X)| < \epsilon^{1-c}) > \epsilon^{O_{N,c,d}(1)}.$$

This allows us to apply Proposition 4 again to q. Applying this argument iteratively, we find that if $|\nabla p(X)|$ is small with reasonable probability, that

$$\Pr_{X,Y^i}\left(|\nabla D_{Y^1}\cdots D_{Y^{d-1}}p(X)| < \epsilon^{1-c}\right) > \epsilon^{O_{N,c,d}(1)}.$$

Since the d^{th} order derivatives of p are constant, the above statement does not depend on X. In fact, it depends on p only through P, the tensor of d^{th} order partials of p. In particular, it says that P has the property:

$$\Pr_{Y^i}(|P(Y^1,\ldots,Y^{d-1},-)|_2 < \epsilon^{1-c}) > \epsilon^{O_{N,c,d}(1)}. \quad (3)$$

One way in which Equation (3) might hold is if P is of the form

$$P = \sum_\ell A^\ell \otimes B^\ell + O(\epsilon^{1-c})$$

for some lower-rank tensors A^ℓ and B^ℓ. The necessary inequality will then hold as long as $A^\ell(Y^i)$ is small for each ℓ. As it turns out, the converse of this statement is also true. This decomposition of P corresponds more or less directly to the desired decomposition of p.

Proposition 5 (Tensor Decomposition Proposition)**.** *Let d be an integer, and let $c, N, \epsilon > 0$ be real numbers. Then for all rank-d tensors A with $|A|_2 \le 1$ and*

$$Pr_{X^i}(|A(X^1,\ldots,X^{d-1},-)|_2 < \epsilon) > \epsilon^N$$

there exist tensors U^ℓ, V^ℓ, $1 \le \ell \le k = O_{c,d,N}(1)$ defined on complimentary proper subsets of the coordinates of A such that $|U^\ell|_2|V^\ell|_2 \le O_{c,d,N}(|A|_2\epsilon^{-c})$ for all ℓ and

$$\left|A - \sum_{\ell=1}^k U^\ell(A)\otimes V^\ell(A)\right|_2 = O_{c,d,N}(\epsilon^{1-c}).$$

The proof of the Tensor Decomposition Proposition is itself quite difficult. We show the stronger statement that there is a probability distribution over sequences of tensor-valued polynomials $U^\ell(A), V^\ell(A)$, so that if A satisfies the necessary hypotheses, $U^\ell(A), V^\ell(A)$ satisfy the conclusion with probability at least $\epsilon^{O_{c,d,N}(1)}$. This is in turn proved by induction on d.

We note that if A satisfies the hypothesis of the Tensor Decomposition Proposition, that $A(X^1,-)$ also does with probability $\epsilon^{O_{c,d,N}(1)}$ over X^1. By the inductive hypothesis, there must exist specific tensor-valued polynomials U^ℓ, V^ℓ so that with probability $\epsilon^{O_{c,d,N}(1)}$ over $B = A(X^1,-)$ we have that

$$\left|B - \sum_{\ell=1}^k U^\ell(B)\otimes V^\ell(B)\right|_2 = O_{c,d,N}(\epsilon^{1-c}).$$

By Strong Anticoncentration, this implies that the derivative of the above tensor with respect to X^1 is approximately low rank. This says that $A = \nabla_{X^1}B$ is approximated by

$$\sum_{\ell=1}^k \left(\nabla_{X^1}U^\ell(B)\right)\otimes V^\ell(B) + U^\ell(B)\otimes\left(\nabla_{X^1}V^\ell(B)\right)$$

plus the sum of a bounded number of tensor products of a tensor on the first coordinate with a tensor on the remaining coordinates. This approximates A as a sum of products of lower-rank tensors. We have left to show that such an approximation can be found with reasonable probability by an appropriate probability distribution over tensor-valued polynomials in A. We consider the tensor T defined by

$$T(X^1) = D_{X^1}\left[B - \sum_{\ell=1}^k U^\ell(B)\otimes V^\ell(B)\right].$$

We know that T can be approximated by the sum of a bounded number of products of a tensor over X^1 with a tensor over its other coordinates. In particular, thinking of T as a linear transformation taking X^1 to the appropriate $(d-1)$-tensor, this says that T is approximately a matrix of low-rank. Thus T is approximated by a sum $\sum_{i=1}^k C_i v_i \otimes w_i$ with $\{v_i\}$ and $\{w_i\}$ orthonormal sets. For random $Z^1,\ldots,Z^k$, we expect that $\{T(Z^i)\}$ will approximately span $\{w_i\}$. By guessing the corresponding change of basis, we can with reasonable probability obtain approximations $\tilde{w}_i \approx w_i$. If this succeeds, we may then approximate T as

$$T(X) \approx \sum_{i=1}^k \tilde{w}_i \left\langle \tilde{w}_i, T(X)\right\rangle.$$

This completes the proof of Proposition 5, providing the final piece in the proofs of Proposition 3 and Theorem 1.

IV. Diffuse Decompositions and the Replacement Method

One of the main uses of the theory of diffuse decompositions is to improve upon applications of the replacement method. Given a polynomial threshold function $f = \text{sgn}(p(x))$, standard applications of the replacement method approximates f by the smooth function $g(x) = \rho(p(x))$. This has the problem that if p is poorly anticoncentrated near zero, then $g(X)$ and $f(X)$ will necessarily disagree with relatively high probability unless ρ is picked to have very large derivatives near zero. On the other hand, if p has a diffuse decomposition $(h, q_1, \ldots, q_m)$, then f may be approximated instead by $g(x) = \rho(q_1(x),\ldots,q_m(x))$ for ρ some smooth approximation to $\text{sgn}\circ h$. Since $(q_1,\ldots,q_m)$ is a diffuse set, ρ can now be allowed to transition between 1 and -1 over a much larger distance without producing a large discrepancy between $g(X)$ and $f(X)$. In particular, we show:

Proposition 6. *Let $(h, q_1, \ldots, q_m)$ by an (ϵ, N)-diffuse decomposition of a degree-d polynomial p for $1/2 > \epsilon > 0$. There exists a function $g : \mathbb{R}^m \to [-1, 1]$ so that:*

1) $g(q_1(x), q_2(x), \ldots, q_m(x)) \geq sgn(p(x))$ *pointwise.*
2) $\mathbb{E}[g(q_1(X), q_2(X), \ldots, q_m(X))] - \mathbb{E}[sgn(p(X))] = O_{m,d}(\epsilon N \log(\epsilon^{-1})^{dm/2+1})$.
3) *For any* $k \geq 0$, $|g^{(k)}|_\infty = O_{m,k}(\epsilon^{-k})$, *where* $|g^{(k)}|_\infty$ *denotes the largest* k^{th} *order mixed partial derivative of* g *at any point.*

Using the above g in a more or less standard application of the replacement method, we obtain the following Proposition:

Proposition 7. *Let* $p_0 : \mathbb{R}^n \to \mathbb{R}$ *be a degree-*d *polynomial with an* (ϵ, N)*-diffuse decomposition* $(h, q_1, \ldots, q_m)$ *for some* $1/2 > \epsilon > 0$. *Let* n_i *be positive integers so that* $n = \sum_{i=1}^{\ell} n_i$. *We can then consider* p_0 *and each of the* q_i *as functions on* $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_\ell}$.

Let $X^1, \ldots, X^\ell$ *and* $Y^1, \ldots, Y^\ell$ *be any independent random variables, where* X^j *and* Y^j *take values in* $\mathbb{R}^{n_j}$ *and* Y^j *is a random Gaussian. Furthermore, assume that for some integer* $k > 1$ *that for any polynomial* g *in* m *variables of degree less than* k, *any* $1 \leq j \leq \ell$ *and any* z^i *that*

$$\begin{aligned}&\mathbb{E}\left[g(q_i(z^1, \ldots, z^{j-1}, X^j, z^{j+1}, \ldots, z^\ell))\right] \\ &\quad = \mathbb{E}\left[g(q_i(z^1, \ldots, z^{j-1}, Y^j, z^{j+1}, \ldots, z^\ell))\right].\end{aligned}$$

For each $1 \leq i \leq m$ *and each* $1 \leq j \leq \ell$ *define*

$$\begin{aligned}Q_{i,j}(x^1, \ldots, &x^{j-1}, x^{j+1}, \ldots, x^\ell) := \\ &\mathbb{E}_{Y^j}[q_i(x^1, \ldots, x^{j-1}, Y^j, x^{j+1}, \ldots, x^\ell)],\end{aligned}$$

and

$$\begin{aligned}V_{i,j}(x^1, \ldots, x^\ell) &:= q_i(x^1, \ldots, x^\ell) \\ &\quad - Q_{i,j}(x^1, \ldots, x^{j-1}, x^{j+1}, \ldots, x^\ell).\end{aligned}$$

Define $T_{i,j}$ *to be*

$$\begin{aligned}&\mathbb{E}\left[\left|V_{i,j}(Y^1, \ldots, Y^j, X^{j+1}, \ldots, X^\ell)\right|^k\right] \\ +&\mathbb{E}\left[\left|V_{i,j}(Y^1, \ldots, Y^{j-1}, X^j, \ldots, X^\ell)\right|^k\right].\end{aligned}$$

And let

$$T := \sum_{i=1}^{m} \sum_{j=1}^{\ell} T_{i,j}.$$

Then we have that

$$\begin{aligned}&\left|Pr\left(p_0(X^1, \ldots, X^\ell) \leq 0\right) - Pr\left(p_0(Y^1, \ldots, Y^\ell) \leq 0\right)\right| \\ &\leq O_{d,m,k}\left(\epsilon^{-k} T + \epsilon N \log(\epsilon^{-1})^{dm/2+1}\right).\end{aligned}$$

Furthermore, if p *is a degree-*d *polynomial so that for some parameters* $\delta, \eta > 0$ *the probabilities*

$$Pr\left(|p(X) - p_0(X)| < \delta |p|_2\right), Pr\left(|p(Y) - p_0(Y)| < \delta |p|_2\right)$$

are each at least $1 - \eta$, *then*

$$\begin{aligned}&\left|Pr\left(p(X^1, \ldots, X^\ell) \leq 0\right) - Pr\left(p(Y^1, \ldots, Y^\ell) \leq 0\right)\right| \\ &\leq O_{d,m,k}\left(\epsilon^{-k} T + \epsilon N \log(\epsilon^{-1})^{dm/2+1} + \delta^{1/d} + \eta\right).\end{aligned}$$

This will prove to be the main technical tool for proving the later results in this paper.

The basic idea of the proof of Proposition 7 is to sandwich $f(x) = \text{sgn}(p_0(x))$ between smooth functions $g_- \leq f \leq g_+$ of the type given in Proposition 6. We then show for these smooth functions that

$$\mathbb{E}[g_\pm(X)] \approx \mathbb{E}[g_\pm(Y)].$$

This is done by replacing the X^j in the above expectation by the corresponding Y^j one at a time and bounding the errors. When performing the j^{th} replacement, we may fix the values of $Y^1, \ldots, Y^{j-1}, X^{i+1}, \ldots, X^\ell$ and approximate h by its degree $k-1$ Taylor polynomial about $(Q_{1,j}, \ldots, Q_{m,j})$. By assumption, the expectations of the terms of this polynomial are the same for j^{th} coordinate X^j and for j^{th} coordinate Y^j. Thus, the error in expectations is the difference in expectations of the Taylor error. This is bounded in terms of the k^{th} derivative of $g_\pm$ and the k^{th} moments of $q_i - Q_{i,j}$, and is thus, easily seen to be

$$O_{d,m,k}\left(\epsilon^{-k} \sum_{i=1}^{m} T_{i,j}\right).$$

Summing over j, shows that

$$|\mathbb{E}[g_\pm(X)] - \mathbb{E}[g_\pm(Y)]| = O_{d,m,k}(\epsilon^{-k} T).$$

Our result now follows from the observation that

$$\begin{aligned}\mathbb{E}[f(Y)] \approx \mathbb{E}[g_-(Y)] \approx \mathbb{E}[g_-(X)] \leq \mathbb{E}[f(X)] \leq \mathbb{E}[g_+(X)] \\ \approx \mathbb{E}[g_+(Y)] \approx \mathbb{E}[f(Y)].\end{aligned}$$

V. The Diffuse Invariance Principle and Regularity Lemma

While there are many powerful analytic tools for dealing with polynomials evaluated at Gaussian inputs, a number of questions are both more difficult and perhaps more interesting when Bernoulli inputs are considered. A key tool for dealing with this challenge is that of an invariance principle. An Invariance Principle is a theorem which states that for certain polynomials, p, that

$$\text{Dist}(p(\text{Gaussian})) \sim \text{Dist}(p(\text{Bernoulli})),$$

usually in terms of cdf distance. If, for example, p is a linear polynomial, the above reduces to a special case of the Berry-Esseen Theorem.

As in the case of the Berry-Esseen Theorem, this approximation will fail if any coordinate contributes too large a portion of the final value of the polynomial. In order to quantify this, we define the i^{th} *influence* of a polynomial p

to be

$$\mathrm{Inf}_i(p) := \left|\frac{\partial p}{\partial x_i}\right|_2^2.$$

(Note that this is equivalent to the standard definition.) This influence is a measure of how much changing the value of the i^{th} coordinate effects the value of p. We say that a polynomial, p, is τ-regular if $\mathrm{Inf}_i(p) \leq \tau \mathrm{Var}(p(X))$ for each i.

The celebrated Invariance Principle of [20] states that if p is a τ-regular, multilinear, degree-d polynomial, then the cdf distance between p(Gaussian) and p(Bernoulli) is $O(d\tau^{1/(8d)})$.

Unfortunately, the dependence of this bound on $\tau^{1/d}$ is necessary. In particular, consider (an appropriate multilinear version of) the polynomial

$$p(x_0, \ldots, x_N) = \tau x_0 + \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i\right)^d$$

for d an even integer. It is easy to see that $\mathrm{Inf}_0(p) = \tau^2$, while all other influences are $O(1/N)$, and $\mathrm{Var}(p) = \Theta_d(1)$. It is also clear that under Bernoulli inputs, $p(B)$ is always at least $-\tau$. On the other hand, under Gaussian inputs $p(G) < -\tau$ whenever $x_0 < -2$ and $\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i\right| < \tau^{1/d}$, an event of probability $\Omega(\tau^{1/d})$. Thus, despite p being $O(\tau^2)$-regular, these distributions differ in cdf distance by at least $\Omega(\tau^{1/d})$.

The essential problem arising in this example is that despite its relatively small influence, x_0 can have a large effect on the sign of $p(x) + \tau$ when $|p(x)|$ is small, which happens with decent probability. Thus, its influence alone does not properly measure the impact of the value of x_0 on the final distribution of $p(x)$, as it cannot see the anticoncentration-related properties of the polynomial. Since this information is summarized by a diffuse decomposition of p, we suspect that the correct notion of regularity for the purposes of an Invariance Principle will make use of the notion of a diffuse decomposition. In particular, we use the following definition:

Definition. *For p a degree-d multilinear polynomial, we say that p has a* (τ, N, m, ϵ)-regular decomposition *if there exists a polynomial p_0 of degree-d so that*

- $\mathbb{E}_{X\sim B}[|p(X) - p_0(X)|^2] \leq \epsilon^2 \mathit{Var}(p_0(X))$.
- *p_0 has a $(\tau^{1/5}, N)$-diffuse decomposition of size m, $(h, q_1, \ldots, q_m)$ so that q_i is multilinear for each i and $\mathit{Inf}_j(q_i) \leq \tau$ for each i, j.*

Using this notion of regularity we can show:

Theorem 8 (The Diffuse Invariance Principle)**.** *If p is a degree-d multilinear polynomial that has a (τ, N, m, ϵ)-regular decomposition for $1/2 > \epsilon, \tau > 0$, A and X and random Bernoulli and Gaussian variables respectively and t is a real number, then*

$$|\mathit{Pr}(p(A) \leq t) - \mathit{Pr}(p(X) \leq t)| = O_{d,m}(\tau^{1/5} N \log(\tau^{-1})^{dm/2+1} + \epsilon^{1/d}\log(\epsilon^{-1})^{1/2}).$$

The proof of Theorem 8 is more or less a direct application of Proposition 7, although there are some technical complications relating to the fact that p_0 will not be multilinear.

There are a number of applications for which one would like to be able to apply an Invariance Principle to a polynomial that is not necessarily regular. A standard technique for dealing with this issue is to consider restrictions of the polynomial which are regular. A number of *regularity lemmas* have been proved to show that this can be done in various contexts (for example see [6] or [8]). They tend to say something along the lines of the following:

Theorem 9. *If p is a degree-d polynomial on $\{-1, 1\}^n$, then p can be re-written as a decision tree of depth $O_d(\tau^{-1})$ whose internal nodes correspond to coordinates, and whose leaves correspond to restrictions of p, so that with high probability a random leaf of this decision tree corresponds to a polynomial p_ρ which is either τ-regular or has constant sign with high probability.*

A new regularity lemma is needed when dealing with regular decompositions. The following theorem will suffice for many such purposes.

Theorem 10 (Diffuse Regularity Lemma)**.** *Let p be a degree-d polynomial with Bernoulli inputs. Let $\tau, c, M > 0$ with $\tau < 1/2$. Then p can be written as a decision tree of depth at most*

$$O_{c,d,M}\left(\tau^{-1}\log(\tau^{-1})^{O(d)}\right)$$

with variables at the internal nodes and a degree-d polynomial at each leaf, with the following property: with probability at least $1 - \tau$, a random path from the root reaches a leaf ρ so that the corresponding polynomial p_ρ either satisfies $\mathit{Var}(p_\rho) < \tau^M |p_\rho|_2^2$ or p_ρ has an $(\tau, \tau^{-c}, O_{c,d,M}(1), O_{c,d,M}(\tau^M))$-regular decomposition.

The Diffuse Regularity Lemma is proved by maintaining a diffuse decomposition of an approximation of p and repeatedly throwing any coordinates which have too large an influence on any of the q_i into the decision tree. It is not hard to see that with high probability, this leads to a decomposition in which all of the q_i have small influences. This will yield us a regular decomposition of the restricted polynomial, unless it has small variance. A restriction with small variance will satisfy our secondary condition unless it additionally has a small L^2 norm. Thus, we can construct a decision tree that has the desired properties on any path for which the L^2 norm of the restricted polynomial does not repeatedly drop.

On the other hand, a martingale argument can be used to show that in any decision tree, that the L^2 norm of the restricted polynomial does not drop by a factor of 2 more than $O_d(\log(\delta^{-1}))$ times along any path except with probability at most $O(\delta)$. This completes the proof.

It should be noted that while Theorem 8 is not directly comparable to the classical Invariance Principle (as they use different notions of regularity), the combination of Theorems 8 and 10 can be compared to the combination of the classical versions of the Invariance Principle and regularity lemma. When combined, the later pair states that upon fixing $\tilde{O}_d(\tau^{-1})$ coordinates a polynomial can be made to be either constant sign with high probability or have $O(d\tau^{1/(8d)})$ cdf distance between its Gaussian and Bernoulli evaluations. The diffuse versions of these theorems can be combined to produce a similar statement with the bound on cdf distance replaced by $O_{c,d}(\tau^{1/5-c})$. This improvement will be critical for many applications.

VI. Application to Noise Sensitivity of Polynomial Threshold Functions

The noise sensitivity of a Boolean function is a measure of the probability that a small change to the input of the function will change the value of the output. There are several different notions of noise sensitivity that are useful in different contexts. The *average sensitivity* of a function f, $\mathbb{AS}(f)$, is the expectation over random Bernoulli inputs of the number of input coordinates that one could flip in order to change the value of f. The corresponding notion in the Gaussian setting is the Gaussian average sensitivity, denoted $\mathbb{GAS}(f)$. There is also a notion called noise sensitivity, which for a parameter $0 < \delta < 1$ measures the probability that $f(A) \neq f(B)$ for Bernoulli inputs A and B that differ on a δ-fraction of their inputs, denoted $\mathbb{NS}_\delta(f)$. There is also a Gaussian notion of noise sensitivity, denoted $\mathbb{GNS}_\delta(f)$, defined using a notion of two Gaussians that are noisy versions of one another.

The problem of studying the noise sensitivity of polynomial threshold functions was first considered in the 1994 paper of Gotsman and Linial ([10]). They Conjecture that the largest possible average sensitivity of degree-d polynomial threshold function in n variables is $O(d\sqrt{n})$. By some reductions proved in [6] (and another in similar spirit for the case of Gaussian average sensitivity), this Conjecture would imply bounds of $\mathbb{GAS}(f) = O(d\sqrt{n}), \mathbb{NS}_\delta(f) = O(d\sqrt{\delta})$, and $\mathbb{GNS}_\delta(f) = O(d\sqrt{\delta})$. All of these bounds would be tight if true.

Progress towards proving the Gotsman-Linial Conjecture has been limited. Gotsman and Linial already knew the degree-1 case of their Conjecture in [10]. For $d > 1$, the first non-trivial bounds were proved independently by Diakonikolas-Raghavendra-Servedio-Tan and Harsha-Klivans-Meka in [7] and [12] (these papers were later combined to form [6]). These papers prove bounds of roughly $O(n^{1-O(1/d)})$ on average sensitivity and Gaussian average sensitivity and bounds of roughly $O(\delta^{1-O(1/d)})$ on noise sensitivity and Gaussian noise sensitivity. In [15], the author proved an optimal bound of $O(d\sqrt{\delta})$ for the special case of Gaussian noise sensitivity. Unfortunately, the techniques used to prove this bound do not appear to generalize to the other notions of sensitivity.

We use the Diffuse Invariance Principle and Regularity Lemma to obtain a bound on the Bernoulli noise sensitivity by relating it via an Invariance Principle to the Gaussian noise sensitivity. While such an attack could have been mounted with previously existing technology, the limited power of the existing Invariance Principle would have made it difficult to prove any bound better than $\mathbb{NS}_\delta(f) = O(\delta^{1-O(1/d)})$. Our improved Invariance Principle though will allow us to do better. In particular, we prove:

Theorem 11. *If f is a degree-d polynomial threshold function, and if $c, \delta > 0$, then*

$$\mathbb{NS}_\delta(f) = O_{c,d}(\delta^{1/6-c}).$$

Using reductions between the various notions of sensitivity, we also prove bounds of $O_{c,d}(n^{5/6+c})$ on the average sensitivity and Gaussian average sensitivity of a degree-d polynomial threshold function in n variables.

Theorem 11 is proved in two stages. First, we use a modified form of the Diffuse Invariance Principle, to relate the noise sensitive and Gaussian noise sensitivity of regular PTFs. In particular, since good bounds are known on the Gaussian noise sensitivity, we obtain the following bound in the Bernoulli case:

Proposition 12. *Iff $= sgn \circ p$ is a polynomial threshold function for p a degree-d polynomial with a (τ, N, m, ϵ)-regular decomposition for $1/2 > \epsilon, \tau > 0$, and if $1 > \delta > 0$ is a real number, then*

$$\mathbb{NS}_\delta(f) = O(d\sqrt{\delta}) + O(d\epsilon^{1/2d}\log(\epsilon^{-1})) + O_{d,m}(N\tau^{1/5}\log(\tau^{-1})^{dm/2+1}).$$

In order to reduce the general case to this one, we use the regularity lemma to write p as a decision tree of depth $\tilde{O}(\delta^{-5/6})$, most of whose leaves correspond to polynomials that are either constant with high probability, or have a $(\delta^{5/6}, \delta^{-c/2}, O_{d,c}(1), \delta^{2d})$-regular decomposition. Two inputs to f that differ on only a δ-fraction of inputs, agree on all of the inputs defining a path along this decision tree with probability $1 - \tilde{O}(\delta^{1/6})$. If this is the case, the probability that f differs on these inputs is the average of the noise sensitivities of f over all leaves of our decision tree. On the other hand, in either the regular case or the near-constant-sign case, these noise sensitivities are $O_{c,d}(\delta^{1/6-c})$, yielding our result.

VII. Application to Pseudorandom Generators for Polynomial Threshold Functions

A fundamental problem in the study of polynomial threshold functions is that of finding pseudorandom generators to fool them. In particular, we wish to find some random variable Y given by some easily computable function of a small, purely random seed so that for any degree-d polynomial threshold function, f, in n variables,

$$|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]| \leq \epsilon, \tag{4}$$

where X above is either an n-dimensional Gaussian or n-dimensional Bernoulli random variable. In [19], Meka and Zuckerman show using the probabilistic method, that such Y must exist with seed length as small as $O(d\log(n) + \log(\epsilon^{-1}))$, but the problem of finding explicit generators has proved to be quite difficult.

A number of papers have attempted to produce generators using simple k-wise independence. If it can be shown that any k-wise independent family Y of Bernoullis or Gaussians satisfies Equation (4), then a generator can be produced with seed length $O(k\log(n))$ (for Gaussians, the story is slightly more complicated since the Gaussian distribution must be discretized first). Diakonikolas-Gopalan-Jaiswal-Servedio-Viola show in [4] that $k = O(\epsilon^{-2}\log^2(\epsilon^{-1}))$-independence suffices to fools degree-1 polynomial threshold functions of Bernoullis. In [5], Diakonikolas, Kane and Nelson show that $k = O(\epsilon^{-2})$ is sufficient for $d = 1$ and Gaussians, and $k = O(\epsilon^{-8})$ and $k = \tilde{O}(\epsilon^{-9})$ for $d = 2$ in the Gaussian and Bernoulli case respectively (the later result can be improved using Theorems 8 and 10 of this paper in place of the use of the classical Invariance Principle and Regularity Lemma in [5] to show that $k = O(\epsilon^{-8})$ suffices for both cases). Finally, in [14], it was shown that $k = O_d\left(\epsilon^{-2^{O(d)}}\right)$-independence suffices for arbitrary degree polynomial threshold functions. Unfortunately, a relatively simple construction can be used to show that $\Omega(d^2\epsilon^{-2})$-independence is required in both cases, eliminating the possibility of finding particularly small seed lengths in this way.

There have also been some attempts to come up with pseudorandom generators not based entirely on limited independence. In [19], Meka and Zuckerman come up with a generator of seed length $O(\log(n) + \log^2(\epsilon^{-1}))$ for $d = 1$ using pseudorandom generators against space bounded computation. They also develop a generator of seed length $2^{O(d)}\log(n)\epsilon^{-8d-3}$ for arbitrary degree functions in the Bernoulli case. In [13], in what was in many ways a predecessor to this work, the author developed a generator of seed length $2^{O_c(d)}\log(n)\epsilon^{-4-c}$ for any $c > 0$ in the Gaussian case.

These last two generators were both analyzed using some form of the replacement method. By replacing the standard replacement method in these cases with an appropriate diffuse version, the analysis of these generators can be improved. In particular, it is the case that with slightly modified versions of these generators, degree-d PTFs can be fooled with seed length $O_{c,d}(\log(n)\epsilon^{-2-c})$ in the Gaussian case and $O_{c,d}(\log(n)\epsilon^{-11-c})$ in the Bernoulli case.

A. The Gaussian Generator

The generator in [13] is given by

$$X = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} X^i,$$

where the X^i are chosen independently from k-independent families of Gaussians. The basic idea of the analysis is that for Y^i taken independently from fully independent families of Gaussians that

$$Y = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} Y^i$$

is simply an n-dimensional Gaussian. One then attempts to show that

$$\mathbb{E}[f(X)] \approx \mathbb{E}[f(Y)]$$

using some sort of replacement method.

The error bound that ones obtains using this technique is greatly dependent on which smooth approximation, $g \approx f$ is used in the replacement method. The standard choice of $g = \rho(p(x))$ will lead to errors on the order of $N^{-1/d}$. In [13], the author improves on this by using a g that approximates f well so long as

$$|p(Y)| \geq \epsilon|p'(Y)| \geq \ldots \geq \epsilon^d|p^{(d)}(Y)|.$$

This holds with high probability by Strong Anticoncentration, and the added control over the derivatives of g allow one to obtain error bounds on the order of $N^{-1/4+c}$ for large enough k.

This can be improved further using the theory of Diffuse Decompositions. A direct application of Proposition 7 yields an error on the order of $N^{-1/2+c}$ for large enough k.

B. The Bernoulli Generator

The generator of Meka and Zuckerman is slightly more complicated. Essentially, the coordinates are first hashed into N bins by some k-wise independent hash function $h : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, N\}$. The coordinates in the i^{th} bin are then set by some other k-wise independent hash function $Z^i : \{1, 2, \ldots, n\} \to \{\pm 1\}$. Together, this gives us our generator

$$X(i) := Z^{h(i)}(i).$$

In order to analyze this generator, we note that upon fixing h we can consider our polynomial, p, as some polynomial in the Z^i. Assuming that p is regular and assuming that the sum of the influences of the coordinates that h maps to a given bin is never too large (an event which happens with

high probability), then one can use the replacement method to show that

$$\mathbb{E}[\text{sgn}(p(Z^1, \ldots, Z^N))] \approx \mathbb{E}[\text{sgn}(p(Y^1, \ldots, Y^N))],$$

where the Y^i are standard Gaussians. Since the same approximation would hold for fully independent Z^i (thus, yielding a fully independent X), we know that

$$\mathbb{E}[\text{sgn}(p(X))] \approx \mathbb{E}[\text{sgn}(p(B))]$$

when p is sufficiently regular.

In order to analyze the case of a general polynomial p, we reduce to the case of a regular one using an appropriate regularity lemma. In order for this reduction to work, we need to know that after conditioning on the values of the coordinates in our decision tree, that the resulting conditional distribution on X is still of the above form. This is done by increasing the independence with which the Z^i are chosen by the depth of the decision tree. If we are using the classic notion of regularity, then in order to obtain an error of ϵ, the tree must have depth ϵ^{-d}, and thus, the generator must have seed length more than ϵ^{-d}. On the other hand, using the diffuse notion of regularity, the decision tree can be as small as ϵ^{-5-c}, and a detailed analysis shows that a seed length of $O_{c,d}(\epsilon^{-11-c})$ is sufficient.

ACKNOWLEDGEMENTS

This research was done with the support of an NSF postdoctoral fellowship.

REFERENCES

[1] Richard Beigel *The polynomial method in circuit complexity*, Proc. of 8th Annual Structure in Complexity Theory Conference (1993), pp. 82-95.

[2] Aline Bonami *Étude des coefficients Fourier des fonctions de $L^p(G)$*, Annales de l'Institute Fourier Vol. 20(2), p. 335-402, 1970.

[3] A. Carbery, J. Wright *Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$* Mathematical Research Letters, Vol. 8(3), pp. 233248, 2001.

[4] I. Diakonikolas, P. Gopalan, R. Jaiswal, R. Servedio, E. Viola, *Bounded Independence Fools Halfspaces* SIAM Journal on Computing, Vol. 39(8), p. 3441-3462, 2010.

[5] Ilias Diakonikolas, Daniel M. Kane, Jelani Nelson, *Bounded Independence Fools Degree-2 Threshold Functions*, Foundations of Computer Science (FOCS), 2010.

[6] Ilias Diakonikolas, Prahladh Harsha, Adam Klivans, Raghu Meka, Prasad Raghavendra, Rocco A. Servedio, Li-Yang Tan *Bounding the average sensitivity and noise sensitivity of polynomial threshold functions* Proceedings of the 42nd ACM symposium on Theory of computing (STOC), 2010.

[7] Ilias Diakonikolas, Prasad Raghavendra, Rocco A. Servedio, Li-Yang Tan *Average sensitivity and noise sensitivity of polynomial threshold functions* http://arxiv.org/abs/0909.5011.

[8] Ilias Diakonikolas, Rocco Servedio, Li-Yang Tan, Andrew Wan *A Regularity Lemma, and Low-Weight Approximators, for Low-Degree Polynomial Threshold Functions*, 25th Conference on Computational Complexity (CCC), 2010

[9] W. Feller *An introduction to probability theory and its applications* Vol. II. Second edition. John Wiley & Sons Inc., New York, 1971.

[10] Craig Gotsman, Nathan Linial *Spectral properties of threshold functions* Combinatorica, Vol. 14(1), p. 3550, 1994.

[11] Ben Green, Terence Tao, *The distribution of polynomials over Fnite Felds, with applications to the Gowers norms*, Contrib. Discrete Math Vol. 4(2), p. 1-36, 2009.

[12] Prahladh Harsha, Adam Klivans, Raghu Meka *Bounding the Sensitivity of Polynomial Threshold Functions* http://arxiv.org/abs/0909.5175.

[13] Daniel M. Kane *A Small PRG for Polynomial Threshold Functions of Gaussians* Symposium on the Foundations Of Computer Science (FOCS), 2011.

[14] Daniel M. Kane *k-Independent Gaussians Fool Polynomial Threshold Functions*, Conference on Computational Complexity (CCC), 2011.

[15] Daniel M. Kane *The Gaussian Surface Area and Noise Sensitivity of Degree-d Polynomial Threshold Functions*, in Proceedings of the 25th annual IEEE Conference on Computational Complexity (CCC 2010), pp. 205-210.

[16] Tali Kaufman, Shachar Lovett *Worst Case to Average Case Reductions for Polynomials*, The 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2008).

[17] Adam R. Klivans, Rocco A. Servedio *Learning DNF in time $2^{O(n^{1/3})}$*, J. Computer and System Sciences Vol. 68, p. 303-318, 2004.

[18] J. W. Lindeberg *Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeit srechnung*, Math. Zeit., Vol. 15, p.211-235, 1922.

[19] Raghu Meka, David Zuckerman *Pseudorandom generators for polynomial threshold functions*, Proceedings of the 42nd ACM Symposium on Theory Of Computing (STOC 2010).

[20] E. Mossel, R. ODonnell, and K. Oleszkiewicz *Noise stability of functions with low influences: invariance and optimality* Proceedings of the 46th Symposium on Foundations of Computer Science (FOCS), pages 2130, 2005.

[21] Nelson *The free Markov field*, J. Func. Anal. Vol. 12(2), p. 211-227, 1973.

[22] R.E.A.C.Paley and A.Zygmund, *A note on analytic functions in the unit circle*, Proc. Camb. Phil. Soc. Vol. 28, p. 266272, 1932.

[23] Alexander A. Sherstov *Separating AC0 from depth-2 majority circuits*, SIAM J. Computing Vol. 38, p. 2113-2129, 2009.