*The Annals of Probability*
2017, Vol. 45, No. 3, 1612–1679
DOI: 10.1214/16-AOP1097


# A STRUCTURE THEOREM FOR POORLY ANTICONCENTRATED POLYNOMIALS OF GAUSSIANS AND APPLICATIONS TO THE STUDY OF POLYNOMIAL THRESHOLD FUNCTIONS[1]

BY DANIEL KANE

*University of California*

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.

## CONTENTS



Received December 2013; revised December 2014.
[1]Support by NSF postdoctoral fellowship.
*MSC2010 subject classifications.* Primary 60G15; secondary 68R05.
*Key words and phrases.* Polynomial decompositions, Gaussian chaos, anticoncentration, invariance principle.

## 1. Introduction.

1.1. *Anticoncentration of Gaussian polynomials.* In this paper, we study low degree polynomials of Gaussian random variables, especially with regards to their *anticoncentration* properties, that is the extent to which they cluster probability mass into small intervals (or more particularly, the extent to which they fail to do so). The most that can be said along these lines for general polynomials is given to us by a result of Carbery and Wright in [3]. Namely, they show that for $p$ a degree-$d$ polynomial in $n$ variables and $X$ an $n$-dimensional Gaussian random variable, then

$$\Pr\big(|p(X)| \le \varepsilon |p|_2\big) = O\big(d\varepsilon^{1/d}\big). \tag{1}$$

While this bound does tell us that the probability of $p(X)$ lying in a small interval goes to zero as the length of the interval does, it leaves much to be desired from a quantitative standpoint. In particular, the $\varepsilon^{1/d}$-dependence in equation (1) tends to produce poor bounds if $d$ is moderately large. In particular, one might expect that the probability of $p(X)$ lying in an interval of length $\varepsilon$ should be proportional to $\varepsilon$ rather than $\varepsilon^{1/d}$. Unfortunately, while this is true for most polynomials, there are cases in which it fails. For example, if $p(X)$ is given by the $d$th power of a linear polynomial $L(X)$, then $|p(X)| \le \varepsilon$ if and only if $|L(X)| \le \varepsilon^{1/d}$, which happens with probability proportional to $\varepsilon^{1/d}$. While equation (1) implies that this is approximately the worst case in terms of poor anticoncentration, it is far from the only case where the naive bound fails. For example, if $p(X)$ is given by the sum of $d$th powers of a small number of linear polynomials $q_1(X), \dots, q_m(X)$, then if $|q_i(X)| < (\varepsilon/m)^{1/d}$ for all $i$, $|p(X)| \le \varepsilon$. This happens with probability roughly $\varepsilon^{m/d}$. There are also a number of more complicated counter-examples. For example, if

$$p(X) = q_1(X)^7 + q_2(X)^7 + q_1(X)^2 q_2(X)^2 + \varepsilon^{2/3} q_3(X)$$

for any polynomials $q_1, q_2, q_3$ then as long as $|q_1(X)| \ll \varepsilon^{1/4}$, $|q_2(X)| \ll \varepsilon^{1/4}$, $|q_3(X)| \ll \varepsilon^{1/3}$, an event that we expect to take place with probability roughly $\varepsilon^{5/6}$, then $|p(X)| \leq \varepsilon$.

A common theme in the above examples seems to be that if $p(X)$ is poorly anticoncentrated it is because $p$ can be decomposed in such a way that makes this poor anticoncentration apparent. In particular, in all of the above examples we were able to write $p(X)$ as $h(q_1(X), \ldots, q_m(X))$ for some polynomials $h, q_i$ in such a way that even if the $q_i$ were assumed to be jointly Gaussian distributed, the anticoncentration properties of $p$ are already accounted for in the structure of $h$. In fact, as we will show, this is true in general. Any polynomial $p$ may be approximately decomposed in terms of other polynomials $q_i$ such that the $q_i$ are nearly as well anticoncentrated as one might hope. This provides us with a useful structural result for polynomials of Gaussian inputs. In order to produce a rigorous statement of this result, we must first introduce some terminology.

DEFINITION 1. Given a degree-$d$ polynomial $p : \mathbb{R}^n \to \mathbb{R}$, we say that a sequence 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 2. We say that a tuple of polynomials $(q_1, \ldots, q_m) : \mathbb{R}^n \to \mathbb{R}^m$, is an $(\varepsilon, N)$-*diffuse set* if for every $(a_1, \ldots, a_m) \in \mathbb{R}^m$ and standard Gaussian random variable $X$ we have that

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

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 $(\varepsilon, N)$-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 provides a much stronger notion of anticoncentration. Combining the two definitions above, we define the notion of a *diffuse decomposition*.

DEFINITION 3. Given a polynomial $p$ we say that $(h, q_1, \ldots, q_m)$ is an $(\varepsilon, N)$-*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 $(\varepsilon, N)$-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 $\varepsilon$, $c$ 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 $|p - p_0|_2 < O_{c,d,N}(\varepsilon^N)|p|_2$ so that $p_0$ has an $(\varepsilon, \varepsilon^{-c})$-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 polynomials of Gaussians. Theorem 1 may also be thought of as a continuous analogue of theorems of Green-Tao ([10]) and Kaufman–Lovett ([15]) 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 1. 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.

1.2. *Applications to the study of polynomial threshold functions.* A number of results such as the invariance principle (see [18]) or various pseudorandom generators for polynomial threshold functions (see [13, 17]) compare the output distributions of a polynomial evaluated at different input distributions. An important technique for dealing with such issues is the replacement method of Lindeberg (see [7, 16]). While the replacement method is well adapted to comparing the expectations of smooth functions at different inputs, it is less well adapted to comparing the outputs of threshold functions [i.e., functions of the form $f(x) = \operatorname{sgn}(p(x))$], which are often required for this analysis. The standard method for resolving this issue is approximating the threshold function, $f$, in question by a smooth function $g$. Unfortunately, this will itself introduce a substantial error if there is a large discrepancy between $f$ and $g$. Since a continuous function must fail to approximate a discontinuous one near the locus of discontinuity, bounding this source of error will generally depend on proving an appropriate anticoncentration result, showing that $X$ has a small probability of lying near this locus of discontinuity.

For example, in previous applications, $g$ was often taken to be of the form $g(x) = \rho(p(x))$ for some suitable smooth function $\rho$. In these cases, $g(x)$ would equal $f(x)$ except when the absolute value of $p(x)$ was small. Unfortunately,

in this context, the relatively weak anticoncentration bounds provided by equation (1) have proven to be a major bottleneck in terms of the bounds that have been obtainable. By making use of Theorem 1, we will be able to make substantial improvements to several of these results by making a better choice of the function $g$. In particular, if $f(x) = \operatorname{sgn}(p(x))$ where $p$ is approximated by a $p_0$ with an appropriate diffuse decomposition $(h, q_1, \ldots, q_m)$, we can approximate $f(x)$ by $g(x) = \rho(q_1(x), \ldots, q_m(x))$ for a suitable smooth function $\rho$. The error introduced by this approximation is now bounded by the anticoncentration properties of $(q_1, \ldots, q_m)$, which are controlled by the diffuse property. This technique produces an improvement over several previous results.

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 [17]. We can also use this theory to improve on the invariance principle of [18]. 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 c.d.f. 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 [18] and [6]. Finally, we make use of this technology to prove bounds on the noise sensitivity of polynomial threshold functions, although this result has since been superseded by [14]. Each of these applications will be discussed in more detail in the relevant section of this paper.

1.3. *Overview of the paper.* In Section 2, we introduce a number of basic concepts that will be used throughout the paper. Section 3 will contain the proof of Theorem 1 along with some associated lemmas. In Section 4, we discuss some basic facts about diffuse decompositions that will prove useful to us later on. In Section 5, we discuss our application to pseudorandom generators for polynomial threshold functions of Gaussians. In Section 6, we state and prove our versions of the invariance principle and regularity lemma. In Section 7, we discuss our results relating to noise sensitivity problems. In Section 8, we discuss our results for pseudorandom generators for polynomial threshold functions with Bernoulli inputs. Finally, in Section 9, we provide some closing remarks.

## 2. Basic results and notation.

2.1. *Basic notation.* We will use the notation $O_a(N)$ to denote a quantity whose absolute value is bounded above by $N$ times some constant depending only on $a$.

Throughout this paper, the variables $G, X, Y, Z, X^i, Y^i, Z^i$, etc. will be used to denote standard Gaussian random vectors unless stated otherwise. The coordinates

of these variables will be denoted using subscripts. Thus, $X^i_j$ will denote the $j$th coordinate of the variable $X^i$.

We also recall here the definition of a polynomial threshold function.

DEFINITION 4. A function $f : \mathbb{R}^n \to \{\pm 1\}$ is a (degree-$d$) *polynomial threshold function* (or PTF) if it is of the form $f(x) = \operatorname{sgn}(p(x))$ for some (degree-$d$) polynomial $p$.

2.2. *Basic facts about polynomials of Gaussians.* We recall some basic facts about polynomials of Gaussians. We begin by recalling the $L^t$-norm of a function.

DEFINITION 5. For a function $p : \mathbb{R}^n \to \mathbb{R}$, we let

$$|p|_t = \left(\mathbb{E}_X\left[|p(X)|^t\right]\right)^{1/t}.$$

We now recall some basic distributional results about polynomials evaluated at random Gaussians.

LEMMA 2 (Carbery and Wright). *If $p$ is a degree-$d$ polynomial then*

$$\Pr\left(|p(X)| \le \varepsilon |p|_2\right) = O\left(d\varepsilon^{1/d}\right),$$

*where the probability is over $X$, a standard Gaussian random vector.*

We will make use of the hypercontractive inequality.

LEMMA 3. *If $p$ is a degree-$d$ polynomial and $t > 2$, then*

$$|p|_t \le \sqrt{t-1}^d |p|_2.$$

The proof follows from Theorem 2 of [19], or more directly from Theorem 1.6.2 of [1] by setting the values of $p, q, t, f$ that appearing in that theorem (which we call $p', q', t', f'$ to avoid confusion) to $p' = 2, q' = t, t' = \log(q-1)/2$ and $f' = T_{t'}^{-1}(p)$.

In particular, this implies the following corollary.

COROLLARY 4 (Weak anticoncentration). *Let $p$ be a degree-$d$ polynomial in $n$ variables. Let $X$ be a standard Gaussian random vector, then*

$$\Pr\left(|p(X)| \ge |p|_2/2\right) \ge 9^{-d}/2.$$

PROOF. This follows immediately from the Paley—Zygmund inequality ([20]) applied to $p^2$. □

We also have the following concentration bound.

COROLLARY 5. *If $p$ is a degree-$d$ polynomial and $N > 0$, then*

$$\Pr_X\big(|p(X)| > N|p|_2\big) = O\big(2^{-(N/2)^{2/d}}\big).$$

PROOF. Apply the Markov inequality and Lemma 3 with $t = (N/2)^{2/d}$. □

Note that we will often need to apply a version of Lemma 2 or Corollary 5 when $p$ is a vector valued polynomial. This can be achieved by applied the appropriate result to the degree-$2d$ polynomial $q(X) := |p(X)|^2$.

2.3. *Multilinear algebra.* The conventions and results discussed in the remainder of this section will be used primarily in Section 3, and sparingly in the rest of the paper.

We will later need to make some fairly complicated constructions making use of multilinear algebra. We take this time to review some of the basic definitions and go over some of the notation that we will be using. We recall that a $k$-tensor is an element of a $k$-fold tensor product of vector spaces $A \in V_1 \otimes \cdots \otimes V_k$. Equivalently, it may be thought of as the $k$-linear form $V_1 \times \cdots \times V_k \to \mathbb{R}$ given by $(v_1, \ldots, v_k) \to \langle A, v_1 \otimes \cdots \otimes v_k\rangle$ (assuming that each of the $V_i$ come equipped with an inner product). If the $V_i$ come with isomorphisms to $\mathbb{R}^{n_i}$, then we can associate $A$ with the sequence of coordinates $A_{i_1 \cdots i_k} = A(e_{i_1}, \ldots, e_{i_k})$.

We recall Einstein summation notation which says that if we are given a product of tensors with stated indices that it is implied that we sum over any shared indices. In particular, if $A$ is a $k_1$-tensor and $B$ a $k_2$-tensor then the expression

$$A_{i_1,i_2,\ldots,i_m,j_1,j_2,\ldots,j_{k_1-m}} B_{i_1,i_2,\ldots,i_m,j_{k_1-m+1},j_{k_1-m+2},\ldots,j_{k_1+k_2-2m}}$$

denotes the $(k_1 + k_2 - 2m)$-tensor $C$ with coordinates

$$C_{j_1,j_2,\ldots,j_{k_1+k_2-2m}} = \sum_{i_1,i_2,\ldots,i_m} A_{i_1,i_2,\ldots,i_m,j_1,j_2,\ldots,j_{k_1-m}} \cdot B_{i_1,i_2,\ldots,i_m,j_{k_1-m+1},j_{k_1-m+2},\ldots,j_{k_1+k_2-2m}}.$$

Note that if there are no overlapping indices that this product simply denotes the tensor product of $A$ and $B$. If on the other hand, all indices overlap, this denotes the dot product of $A$ and $B$. We will also sometimes group several coordinates into a single coordinate of larger dimension. We will try to use upper case letters for indices to indicate that this is happening.

We define the $L^2$ norm of a tensor $A$ to be the square root of the sum of the squares of its coordinates. If $A$ is a $k$-tensor, we have the equivalent definitions:

$$\begin{aligned} |A|_2^2 &= \langle A, A\rangle \\ &= \sum_{i_1,\ldots,i_k} |A_{i_1,\ldots,i_k}|^2 \\ &= \mathbb{E}_{X^1,\ldots,X^k}\big[\big|A_{i_1,\ldots,i_k} X^1_{i_1} X^2_{i_2} \cdots X^k_{i_k}\big|^2\big]. \end{aligned}$$

For tensor-valued functions $A(X)$, we define the $L^2$-norm by

$$|A|_2^2 := \mathbb{E}_X\big[|A(X)|_2^2\big].$$

We will also need the notion of a wedge product of tensors over some subset of their coordinates. In particular, if $A$ is a rank-$(k+m)$ tensor with its first $k$ indices corresponding to spaces of the same dimension, we define

$$\bigwedge_{i_1,\ldots,i_k} A_{i_1,\ldots,i_k,j_1,\ldots,j_m} := \sum_{\sigma\in S_k} (-1)^\sigma A_{i_{\sigma(1)},\ldots,i_{\sigma(k)},j_1,\ldots,j_m}.$$

Note the important special case here where $A$ is a tensor product of $k$ different 1-tensors $A_{i_1,\ldots,i_k} = A^1_{i_1}\cdots A^k_{i_k}$. It is then the case that

$$\Big(\bigwedge_{i_1,\ldots,i_k} A^1_{i_1}\cdots A^k_{i_k}\Big) B^1_{i_1}\cdots B^k_{i_k} = \det\big(\langle A^i, B^j\rangle\big).$$

We will think of the derivative operator as taking functions on $\mathbb{R}^n$ whose values are $k$-tensors to functions on $\mathbb{R}^n$ whose values are $(k+1)$-tensors. In particular, given a tensor valued function $A_S(x)$, we define the tensor $\nabla_i A_S(x)$ to have $(i,S)$-coordinate $\frac{\partial A_S(x)}{\partial x_i}$. For example, for a function $f$, we have that $\nabla_i f$ is the gradient of $f$, $\nabla_i\nabla_j f$ is the Hessian matrix, and $\nabla_i\nabla_i f$ is the Laplacian. Furthermore, if $X$ is a vector, then $X_i\nabla_i f$ is the standard directional derivative $D_X f$.

Lastly, note that if $p$ is a homogeneous, degree-$d$ polynomial that it has an associated $d$-tensor $A$ given by $A_{i_1,\ldots,i_d} := \nabla_{i_1}\cdots\nabla_{i_d} p$ (note that this $d$th order derivative is independent of the point at which it is being evaluated). Note that $A$ is determined by the property that it is a symmetric tensor (it is invariant under any permutation of coordinates) so that for any vector $X$, $A(X,X,\ldots,X) = d!p(X)$.

2.4. *Strong anticoncentration*. Strong anticoncentration was an idea first espoused by the author in [13]. It is a heuristic which states that a polynomial is generally not much smaller than its derivative. We will need to make use of a generalization of this to sets of several tensor-valued polynomials. In particular, we will prove the following proposition.

PROPOSITION 6 (Strong anticoncentration). *For $1\le i\le k$, let $A^i_{S_i}(x)$ (for multiindices $S_i$) be a degree-$d_i$, tensor-valued polynomial on $\mathbb{R}^n$ (i.e., a tensor whose coefficients are degree-$d_i$ polynomials on $\mathbb{R}^n$). Let $1/2>\varepsilon>0$. We have that*

$$\Pr\Bigg(\prod_{j=1}^k |A^j_{S_j}(X)|_2 < \varepsilon \,\Bigg|\, \bigwedge_{i_1,\ldots,i_k}\prod_{j=1}^k \nabla_{i_j} A^j_{S_j}(X)\Bigg|_2\Bigg) \le \varepsilon 2^{O(d_1+d_2+\cdots+d_k)} O(\sqrt{k})^{k+1}\log\big(\varepsilon^{-1}\big)^k.$$

In order to prove Proposition 6, we will need to following lemma.

LEMMA 7. *For $1 \le i \le k$ let $p^i$ be a degree $d_i$ polynomial on $\mathbb{R}^n$ and let $\delta, \varepsilon_i > 0$. Then*

$$\Pr_{X,Y^1,\ldots,Y^k}\bigl(|p^i(X)| < \varepsilon_i \text{ for all } i, \text{ and } |\det(D_{Y^j} p^i(X))| > \delta\bigr) \le \frac{2^{k+1}\prod_{i=1}^k d_i \prod_{i=1}^k \varepsilon_i}{\delta V_k},$$

*where $V_k = \frac{2\pi^{k+1/2}}{\Gamma((k+1)/2)}$ is the volume of the unit $k$-sphere.*

PROOF. Define the function $f : S^k \to \mathbb{R}^k$ by letting

$$f(a_0, a_1, \ldots, a_k)_i := p^i\bigl(a_0 X + a_1 Y^1 + a_2 Y^2 + \cdots + a_k Y^k\bigr).$$

Notice that the matrix with coefficients $D_{Y^j} p^i(X)$ is simply the Jacobian of $f$ at the point $(1, 0, 0, \ldots, 0)$. Notice that if we replace the random variables $X, Y^1, \ldots, Y^k$ by linear combinations of each other by making an orthonormal change of coordinates, that they are still independent Gaussians, and thus, the probability in question is unchanged. We claim that for any fixed values of $X, Y^i$ that the probability over a random such change of variables that

$$|p^i(X)| < \varepsilon_i \qquad \text{for all } i, \quad \text{and} \quad |\det(D_{Y_j} p^i(X))| > \delta$$

is at most $\frac{2^k \prod_{i=1}^k d_i \prod_{i=1}^k \varepsilon_i}{\delta V_k}$. Such a statement would clearly imply our lemma.

Note that making such a random change of variables is equivalent to precomposing $f$ with a random element of the orthogonal group $O(k+1)$. Thus, it suffices to bound

$$\Pr_{x \in S^k}\bigl(f(x) \in R, \text{ and } |\det(\mathrm{Jac}(f(x)))| > \delta\bigr),$$

where $R \subset \mathbb{R}^k$ is given by $\prod_i [-\varepsilon_i, \varepsilon_i]$. Let $T$ be the set of $x \in S^k$ so that $f(x) \in R$, and $|\det(\mathrm{Jac}(f(x)))| > \delta$. We know by the change of variables formula for integration that

$$\int_T |\det(\mathrm{Jac}(f(x)))|\, dx = \int_R |f^{-1}(y)|\, dy. \tag{2}$$

We note that the right-hand side of equation (2) is $\int_R |f^{-1}(y)|\, dy$. We note that $f^{-1}(y)$ is at most the number of isolated points in the intersection of the roots polynomials of degree $d_1, \ldots, d_k$ and $\sum x_i^2 - 1$ in $\mathbb{R}^{k+1}$. Applying Bezout's theorem (see [8], Example 12.3.1) to the homogenized versions of these polynomials, we find that the integrand above is at most $2\prod_{i=1}^k d_i$. Thus, $\int_R |f^{-1}(y)|\, dy \le 2^{k+1}\prod_{i=1}^k d_i \prod_{i=1}^k \varepsilon_i$. On the other hand, the left-hand side of equation (2) is at least $\delta \operatorname{Vol}(T) = \delta V_k \Pr_{x \in S^k}(x \in T)$. Thus,

$$\Pr_{x \in S^k}(x \in T) \le \frac{2^{k+1}\prod_{i=1}^k d_i \prod_{i=1}^k \varepsilon_i}{\delta V_k}. \qquad \square$$

COROLLARY 8. *For polynomials $p^i : \mathbb{R}^n \to \mathbb{R}$ of degree $d_i$ for $1 \le i \le k$ and for $1/2 > \varepsilon > 0$,*

$$\Pr_{X, Y^1, \dots, Y^k}\left(\prod_{i=1}^{k} |p^i(X)| < \varepsilon \left|\det\left(D_{Y^i} p^j\right)\right|\right) \le \varepsilon 2^{O(d_1+d_2+\cdots+d_k)} O(\sqrt{k})^{k+1} \log\left(\varepsilon^{-1}\right)^k.$$

PROOF. We note that the problem in question is invariant under scalings of the $p^i$ and, therefore, we may assume that $|p^i|_2 = 1$ for all $i$. We note by Lemma 2 and Corollary 5 that we may ignore the case where some $|p^i(X)| < \varepsilon^{d_i}$ or where some $|p^i(X)| > \varepsilon^{-1}$ [as the probability that such an event happens for any $i$ is at most $O(\sum_i d_i \varepsilon)$]. For each $i$, we may partition the interval $[\varepsilon^{d_i}, \varepsilon^{-1}]$ into $O(d_i \log(\varepsilon^{-1}))$ many intervals each of whose endpoints differ by at most a factor of 2. Up to a factor of $O(\log(\varepsilon^{-1}))^k \prod_i d_i$, it suffices to bound the probability that each of the $|p^i(X)|$ lies in a specified such interval and that $\prod_{i=1}^k |p^i(X)| < \varepsilon |\det(D_{Y^i} p^j)|$. If the upper endpoints of these intervals are $\varepsilon_i$, then this probability, is at most the probability that

$$|p^i(X)| < \varepsilon_i \qquad \text{for all } i, \quad \text{and} \quad \left|\det\left(D_{Y^j} p^i(X)\right)\right| > \left(2^k \varepsilon\right)^{-1} \prod \varepsilon_i.$$

By Lemma 7, the above probability is at most $\varepsilon 2^{O(d_1+d_2+\cdots+d_k)} O(\sqrt{k})^{k+1}$. Multiplying by $O(\log(\varepsilon^{-1}))^k \prod_i d_i$, yields our bound. □

PROOF OF PROPOSITION 6. For $Z$, a tensor of the same dimension as $A^j$, let $f_Z^j : \mathbb{R}^n \to \mathbb{R}$ be the function $f_Z^j(x) = \langle A^j(x), Z \rangle$. Note that

$$\left| \bigwedge_{i_1,\dots,i_k} \prod_{j=1}^{k} \nabla_{i_j} A^j_{S_j}(X) \right|_2^2 = \mathbb{E}_{Y^1,\dots,Y^k, Z^1,\dots,Z^k}\left[\left|\det\left(D_{Y^i} f^j_{Z^j}(X)\right)\right|^2\right].$$

Furthermore,

$$\left(\prod_{j=1}^{k} |A^j_{S_j}(X)|_2\right)^2 = \mathbb{E}_{Z^1,\dots,Z^k}\left[\left|\prod_{j=1}^{k} f^j_{Z^j}(X)\right|^2\right].$$

Now suppose that for some choice of $X$ that

$$\prod_{j=1}^{k} |A^j_{S_j}(X)|_2^2 < \varepsilon^2 \left| \bigwedge_{i_1,\dots,i_k} \prod_{j=1}^{k} \nabla_{i_j} A^j_{S_j}(X) \right|_2^2. \tag{3}$$

We have by Corollary 4 that with probability at least $2^{O(k)}$ over the random Gaussians $Y^1, \dots, Y^k$ and $Z^1, \dots, Z^k$ that the left-hand side of equation (3) is at least

$$\left|\prod_{j=1}^{k} f^j_{Z^j}(X)\right|^2 \Big/ 2.$$

By the Markov bound, we have that except for a probability of at most $2^{-\Omega(k)}$ the right-hand side of equation (3) is at most

$$\varepsilon^2 2^{O(k)} \left|\det\left(D_{Y^i} f^j_{Z^j}(X)\right)\right|^2.$$

Thus, whenever equation (3) holds, with probability at least $2^{O(k)}$ over $Y^i$ and $Z^i$ we have that

$$\left|\prod_{j=1}^{k} f^j_{Z^j}(X)\right| \leq \varepsilon 2^{O(k)} \left|\det\left(D_{Y^i} f^j_{Z^j}(X)\right)\right|.$$

But by Corollary 8, the probability of this happening (even for fixed $Z^i$) is at most

$$\varepsilon 2^{O(d_1+d_2+\cdots+d_k)} O(\sqrt{k})^{k+1} \log\left(\varepsilon^{-1}\right)^k.$$

Thus, the probability of equation (3) holding is at most $2^{O(k)}$ times as much, which is still

$$\varepsilon 2^{O(d_1+d_2+\cdots+d_k)} O(\sqrt{k})^{k+1} \log\left(\varepsilon^{-1}\right)^k. \qquad \square$$

2.5. *Orthogonal polynomials.* Here, we review some basic facts about orthogonal polynomials. Recall that the Hermite polynomials are an orthonormal basis for polynomials in one variable with respect to the Gaussian inner product. In particular, they are defined by the properties that:

- $H_n : \mathbb{R} \to \mathbb{R}$ is a degree-$n$ polynomial.
- $\mathbb{E}[H_n(X)H_m(X)] = \delta_{n,m}$ where $X$ is a one-dimensional Gaussian random variable.

Furthermore, we have the relation that $H'_n(x) = \sqrt{n} H_{n-1}(x)$. We can extend this theory to polynomials in $n$ variables as follows. For $a = (a_1, \ldots, a_n)$, a vector of nonnegative integers, we define the corresponding polynomial $H_a(x) = \prod_{i=1}^n H_{a_i}(x_i)$ on $\mathbb{R}^n$. It is easy to check that the total degree of $H_a$ is $|a|_1 := \sum_{i=1}^n a_i$ and that $\mathbb{E}[H_a(X)H_b(X)] = \delta_{a,b}$.

Given a polynomial $p$ in $n$ variables, we can always write $p$ as a linear combination of Hermite polynomials. In fact, it is easy to check that

$$p(X) = \sum_{|a|_1 \leq \deg(p)} c_a(p) H_a(X),$$

where

$$c_a(p) = \mathbb{E}\left[p(X)H_a(X)\right].$$

We define the $k$th harmonic component of $p$ to be

$$p^{[k]} := \sum_{|a|_1 = k} c_a(p) H_a(X).$$

We say that $p$ is harmonic of degree $k$ if it equals its $k$th harmonic part.

Note that we can compute the derivative of $H_a$ as

$$\frac{\partial H_a(X)}{\partial X_i} = \sqrt{a_i} H_{a-e_i}(X).$$

Thus, $\nabla_i H_a(X)$ is a vector of polynomials that are harmonic of degree $|a|_1 - 1$. Furthermore, we have that

$$\begin{aligned}
\mathbb{E}\big[(\nabla_i H_a(X))(\nabla_i H_b(X))\big] &= \sum_i \mathbb{E}\big[\sqrt{a_i b_i} H_{a-e_i}(X) H_{b-e_i}(X)\big] \\
&= \sum_i \sqrt{a_i b_i}\delta_{a-e_i,b-e_i} \\
&= \delta_{a,b} \sum_i \sqrt{a_i b_i} \\
&= \delta_{a,b} \sum_i a_i \\
&= |a|_1 \delta_{a,b}.
\end{aligned}$$

Additionally, for $a \neq b$ each of the components of $\nabla_i H_a$ is a Hermite polynomial orthogonal to the corresponding component of $\nabla_i H_b$. Iterating this, we can see that

$$\begin{aligned}
&\mathbb{E}\big[(\nabla_{i_1}\nabla_{i_2}\cdots\nabla_{i_k} H_a(X))(\nabla_{i_1}\nabla_{i_2}\cdots\nabla_{i_k} H_b(X))\big] \\
&\quad = |a|_1(|a|_1 - 1)\cdots(|a|_1 - k + 1)\delta_{a,b}.
\end{aligned}$$

Hence, we have

LEMMA 9. *For $p$ a polynomial of degree $d$,*

$$|\nabla_{i_1}\cdots\nabla_{i_k} p|_2^2 \leq d(d-1)\cdots(d-k+1)|p|_2^2$$

*with equality if and only if $p$ is harmonic of degree $d$.*

## 3. Proof of the decomposition theorem.

3.1. *Overview of the proof.* The proof of Theorem 1 comes in two steps. The first is Proposition 10 (below), which states roughly that if $p$ is a degree-$d$ polynomial so that for a random Gaussian $X$, $|p'(X)|$ is small with nonnegligible probability, then $p$ can be decomposed as a polynomial with smaller $L^2$ norm, plus a sum of products of lower degree polynomials. Given this proposition, the proof of Theorem 1 is relatively straightforward. We begin by writing a trivial decomposition of $p$ as $p(x) = \mathrm{Id}(p(x))$. If this is a diffuse decomposition, we are done. Otherwise, by Proposition 6, there must be a reasonable probably that $|p'(X)|$ is small. Thus, Proposition 10 allows us to decompose $p$ in terms of lower-degree polynomials.

This gives us a new decomposition of $p$. If it is diffuse, we are done, otherwise it is not hard to show that at least one of the polynomials in this decomposition can be decomposed further. We show that this procedure will eventually terminate by demonstrating an ordinal monovariant which decreases with each step.

In Section 3.2, we state and prove Proposition 10, and in Section 3.3 complete the proof of Theorem 1.

3.2. *The decomposition lemma.* In this section, we will prove the following important proposition that will allow us to write a nondiffuse polynomial in terms of lower-degree polynomials.

PROPOSITION 10. *Let $p(X)$ be a degree $d$ polynomial with $|p|_2 \leq 1$ and let $\varepsilon, c, N > 0$ be real numbers so that*

$$\Pr_X\left(\left|\nabla_i p(X)\right|_2 < \varepsilon\right) > \varepsilon^N.$$

*Then there exist polynomials $a_i(X)$, $b_i(X)$ of degree strictly less than $d$ with $|a_i(X)|_2|b_i(X)|_2 \leq O_{N,c,d}(\varepsilon^{-c})|p^{[d]}|_2$ and so that*

$$\left|\left(p(X) - \sum_{i=1}^{k} a_i(X)b_i(X)\right)^{[d]}\right|_2 < O_{N,c,d}\left(\varepsilon^{1-c}\right),$$

*where $k = O_{N,c,d}(1)$. Furthermore, this can be done in such a way that for each $i$, $\deg(a_i) + \deg(b_i) = d$.*

REMARK 2. Unlike the constants implied in Theorem 1, the implied constants in Proposition 10 are primitive recursive functions of the parameters. Although we do not bound them explicitly, our techniques show that they are at worst an iterated exponential.

Our proof of Proposition 10 will proceed in stages. First, we will show that for such polynomials $p$, there is a reasonable probability (over $X, Y^i$) that $\nabla_i D_{Y^1} D_{Y^2} \cdots D_{Y^{d-1}} p(X)$ will be small. This is easily seen to reduce to a statement about the rank-$d$ tensor, $A_{i_1 \cdots i_d} = \nabla_{i_1} \cdots \nabla_{i_d} p$. In particular, we know that $A_{i_1 \cdots i_d} Y^1_{i_1} \cdots Y^{d-1}_{i_{d-1}}$ has a reasonable probability of being small. We then prove a structure theorem telling us that such tensors can be approximated as a sum of tensor products of lower-rank tensors. This in turn will translate into our being able to approximate the degree-$d$ part of $p$ by a sum of products of lower degree polynomials.

We begin with the following proposition.

PROPOSITION 11. *Let $c, N > 0$ be real numbers and $d$ a positive integer. Let $\varepsilon > 0$ be a real number that is sufficiently small given $c, d$ and $N$. Suppose that $p$ is a degree-$d$ polynomial so that*

$$\Pr_X\left(\left|\nabla_i p(X)\right|_2 < \varepsilon\right) > \varepsilon^N.$$

*Then we have that*

$$\Pr_{X,Y}\big(\big|\nabla_i D_Y p(X)\big|_2 < \varepsilon^{1-c}\big) > \varepsilon^{O_{N,c,d}(1)}.$$

We begin with the following lemma.

LEMMA 12. *Let $N > 0$ be a real number and let $d$ and $k$ be positive integers. Suppose that $A_i(X)$ is a degree-$d$, tensor-valued polynomial so that for some $1/2 > \varepsilon > 0$,*

$$\Pr_X\big(\big|A_i(X)\big|_2 < \varepsilon\big) \geq \varepsilon^N.$$

*Then the probability over Gaussian $X$ that $|A_i(X)|_2 < \varepsilon$ and*

$$\Big|\bigwedge_{i_1\cdots i_k}\big(\nabla_{j_1}A_{i_1}(X)\big)\cdots\big(\nabla_{j_k}A_{i_k}(X)\big)\Big|_2 < O_{d,k,N}\big(\varepsilon^{k-N}\big)\log\big(\varepsilon^{-1}\big)^k$$

*is at least $\varepsilon^N/2$.*

PROOF. Note that by decreasing $N$, we may assume that

$$\Pr_X\big(\big|A_i(X)\big|_2 < \varepsilon\big) = \varepsilon^N.$$

Note that for any tensor $B_{ij}$

$$\begin{aligned}\bigwedge_{i_1,\ldots,i_k} B_{i_1j_1}\cdots B_{i_kj_k} &= \sum_{\sigma\in S_k}(-1)^\sigma B_{i_{\sigma(1)}j_1}\cdots B_{i_{\sigma(k)}j_k}\\ &= \sum_{\sigma\in S_k}(-1)^\sigma B_{i_1j_{\sigma^{-1}(1)}}\cdots B_{i_kj_{\sigma^{-1}(k)}}\\ &= \bigwedge_{j_1,\ldots,j_k} B_{i_1j_1}\cdots B_{i_kj_k}.\end{aligned}$$

Thus, for fixed $X$, we have by Lemma 2 that with a probability of at least $9/10$ over $Y^\ell$ we have that

$$\begin{aligned}&\Big|\bigwedge_{i_1,\ldots,i_k}\big(\nabla_{j_1}A_{i_1}(X)\big)\cdots\big(\nabla_{j_k}A_{i_k}(X)\big)\Big|_2\\ &\quad=\Big|\bigwedge_{j_1,\ldots,j_k}\big(\nabla_{j_1}A_{i_1}(X)\big)\cdots\big(\nabla_{j_k}A_{i_k}(X)\big)\Big|_2\\ &\quad>\Omega(1/kd)^{kd}\Big|Y^1_{i_1}\cdots Y^k_{i_k}\bigwedge_{j_1,\ldots,j_k}\big(\nabla_{j_1}A_{i_1}(X)\big)\cdots\big(\nabla_{j_k}A_{i_k}(X)\big)\Big|_2.\end{aligned}$$

Therefore, it suffices to show that with probability at least $3\varepsilon^N/5$ that $|A_i(X)|_2 < \varepsilon$ and

$$\Big|Y^1_{i_1}\cdots Y^k_{i_k}\bigwedge_{j_1,\ldots,j_k}\big(\nabla_{j_1}A_{i_1}(X)\big)\cdots\big(\nabla_{j_k}A_{i_k}(X)\big)\Big|_2 < O_{d,k,N}\big(\varepsilon^{k-N}\big)\log\big(\varepsilon^{-1}\big)^k.$$

For fixed $X$, by Corollary 5 we have that with probability at least $9/10$ that for random $Y^1, \ldots, Y^k$ that $|Y_i^j A_i(X)| < O_k(1)|A_i(X)|_2$ for all $1 \le j \le k$. Thus, with probability at least $9\varepsilon^N/10$ over $X$ and the $Y^j$, we have that $|Y_i^j A_i(X)| < O_k(\varepsilon)$ for all $j$.

On the other hand, Proposition 6 implies that with probability at least $1-\varepsilon^N/10$ that

$$\left| \bigwedge_{j_1,\ldots,j_k} \prod_{\ell=1}^{k} \nabla_{j_\ell} Y_{i_\ell}^{\ell} A_{i_\ell}(X) \right|_2 \le O_{k,d}(1)\varepsilon^{-N}\bigl(\log(\varepsilon^{-1})\bigr)^k \prod_{\ell=1}^{k} |Y_i^{\ell} A_i(X)|. \tag{4}$$

Recall that with probability at least $9\varepsilon^N/10$ we have that $|A_i(X)|_2 < \varepsilon$ and $|Y_i^j A_i(X)| < O_k(\varepsilon)$. When this holds, the right-hand side of equation (4) is at most

$$O_{k,d}(1)\varepsilon^{k-N}\log^k(\varepsilon^{-1}).$$

Hence, with probability at least $4\varepsilon^N/5$, we have that $|A_i(X)|_2 < \varepsilon$ and

$$\left| \bigwedge_{j_1,\ldots,j_k} \prod_{\ell=1}^{k} \nabla_{j_\ell} Y_{i_\ell}^{\ell} A_{i_\ell}(X) \right|_2 < O_{k,d}(1)\varepsilon^{k-N}\log^k(\varepsilon^{-1}),$$

as desired. □

Lemma 12 tells us some very strong information about the tensor $\nabla_j A_i(X)$. In order to understand this better, we will study what it means for a 2-tensor $B_{ij}$ to have $|\bigwedge_{i_1,\ldots,i_k} B_{i_1,j_1}\cdots B_{i_k,j_k}|_2$ small. Recall that a 2-tensor can be thought of as a matrix. We will show that this condition implies that $B_{ij}$ is approximately a matrix of rank at most $k$.

LEMMA 13. *Suppose that $B_{ij}$ is a tensor and suppose that for some integer $k$ and some $\varepsilon > 0$ that*

$$\left| \bigwedge_{i_1,\ldots,i_k} B_{i_1,j_1}\cdots B_{i_k,j_k} \right|_2 < \varepsilon^k.$$

*Then there exist some vectors $V_i^{\ell}$, $W_j^{\ell}$ so that*

$$\left| B_{ij} - \sum_{\ell=1}^{k-1} V_i^{\ell} W_j^{\ell} \right|_2 < O_k(\varepsilon).$$

PROOF. We proceed by induction on $k$. If $k = 1$, we have by assumption that $|B_{ij}|_2 < \varepsilon$, so we are done.

For larger values of $k$, we may assume that

$$\left| \bigwedge_{i_1,\ldots,i_{k-1}} B_{i_1,j_1}\cdots B_{i_{k-1},j_{k-1}} \right|_2 \ge \varepsilon^{k-1},$$

or otherwise we would be done by the inductive hypothesis.

Consider random Gaussians $X^1, \dots, X^k$. We have that

$$\mathbb{E}\Big[\Big|\bigwedge_{i_1,\dots,i_{k-1}} B_{i_1,j_1}\cdots B_{i_{k-1},j_{k-1}} X^1_{j_1}\cdots X^{k-1}_{j_{k-1}}\Big|_2^2\Big]$$
$$= \Big|\bigwedge_{i_1,\dots,i_{k-1}} B_{i_1,j_1}\cdots B_{i_{k-1},j_{k-1}}\Big|_2^2 \geq \varepsilon^{2k-2}.$$

Similarly,

$$\mathbb{E}\Big[\Big|\bigwedge_{i_1,\dots,i_k} B_{i_1,j_1}\cdots B_{i_k,j_k} X^1_{j_1}\cdots X^k_{j_k}\Big|_2^2\Big]$$
$$= \Big|\bigwedge_{i_1,\dots,i_k} B_{i_1,j_1}\cdots B_{i_k,j_k}\Big|_2^2 \leq \varepsilon^{2k}.$$

By Lemma 2, we have that with probability at least $1/2$ that

$$\Big|\bigwedge_{i_1,\dots,i_{k-1}} B_{i_1,j_1}\cdots B_{i_{k-1},j_{k-1}} X^1_{j_1}\cdots X^{k-1}_{j_{k-1}}\Big|_2 \geq \Omega_k\big(\varepsilon^{k-1}\big).$$

Furthermore, by the Markov bound, we can find such $X^1, \dots, X^{k-1}$ so that

$$\mathbb{E}_{X^k}\Big[\Big|\bigwedge_{i_1,\dots,i_k} B_{i_1,j_1}\cdots B_{i_k,j_k} X^1_{j_1}\cdots X^k_{j_k}\Big|_2^2\Big] \leq 2\varepsilon^{2k}.$$

Let $V_i^\ell$ be the vector $B_{ij}X_j^\ell$. We have that

$$\Big|\bigwedge_{i_1,\dots,i_{k-1}} V^1_{i_1}\cdots V^{k-1}_{i_{k-1}}\Big|_2^2 = \Omega_k\big(\varepsilon^{2k-2}\big)$$

and

$$\mathbb{E}_{X^k}\Big[\Big|\bigwedge_{i_1,\dots,i_k} V^1_{i_1}\cdots V^k_{i_k}\Big|_2^2\Big] \leq 2\varepsilon^{2k}.$$

Notice that the wedge products above are simply standard wedges of vectors. Note that if we have vectors $u^1, \dots, u^k$ that

$$u^1 \wedge u^2 \wedge \cdots \wedge u^k = u^1 \wedge u^2 \wedge \cdots \wedge u^{k-1} \wedge u^{k,\perp},$$

where $u^{k\perp}$ is the projection of $u^k$ onto the space perpendicular to $\langle u^1, u^2, \dots, u^{k-1}\rangle$. From here, it is easy to see that we have

$$\frac{|u^1 \wedge u^2 \wedge \cdots \wedge u^k|_2}{|u^1 \wedge u^2 \wedge \cdots \wedge u^{k-1}|_2} = |u^{k,\perp}|_2.$$

Therefore, we have that

$$\mathbb{E}_{X^k}\big[\big|V_i^{k,\perp}\big|_2^2\big] = O_k\big(\varepsilon^2\big).$$

On the other hand, we have that

$$V_i^{k,\perp} = B_{ij}^{\perp} X_j^k,$$

where $B^{\perp}$ is the tensor obtained from $B$ by replacing each row $B_{ij}e_j$ with its projection onto $\langle V^1, V^2, \ldots, V^{k-1}\rangle^{\perp}$. In particular, this means that each row of $B^{\perp}$ can be written as the corresponding row of $B$ plus an element of $\langle V^1, V^2, \ldots, V^{k-1}\rangle$. This means that for some appropriate vectors $W^{\ell}$, we have that $B_{ij}^{\perp} = B_{ij} - \sum_{\ell=1}^{k-1} V_i^{\ell} W_j^{\ell}$. On the other hand, we note that

$$\begin{aligned}\big|B^{\perp}\big|_2^2 &= \mathbb{E}\big[\big|B_{ij}^{\perp} X_j\big|_2^2\big] \\ &= \mathbb{E}\big[\big|V_i^{\perp}\big|_2^2\big] \\ &= O_k\big(\varepsilon^2\big).\end{aligned}$$

Thus, $|B^{\perp}|_2 = O_k(\varepsilon)$, completing our proof. □

We are now prepared to prove Proposition 11.

PROOF. Suppose we are given $c, d, N$ and $\varepsilon > 0$ sufficiently small. Suppose that we have a degree-$d$ polynomial $p$ so that

$$\Pr_X\big(\big|\nabla_i p(X)\big|_2 < \varepsilon\big) > \varepsilon^N.$$

Applying Lemma 2 to the polynomial $x \to |\nabla_i p(x)|^2$, this implies that $\mathbb{E}_X[|\nabla_i p(X)|_2^2] \le O_d(\varepsilon^{-2dN})$, and hence that $\mathbb{E}_X[|\nabla_i \nabla_j p(X)|_2^2] \le O_d(\varepsilon^{-2dN})$.

Let $k$ be an integer so that $k > 2N/c$. By Lemma 12 applied to $\nabla_i p(X)$, we have that with probability at least $\varepsilon^N/2$ that

$$\left|\bigwedge_{i_1,\ldots,i_k} \prod_{\ell=1}^{k} \nabla_{i_\ell}\nabla_{j_\ell} p(X)\right|_2 < O_{c,d,N}\big(\varepsilon^{k(1-c/2)}\big).$$

Let $B_{ij}(X)$ be the tensor $\nabla_i \nabla_j p(X)$. By the above and Corollary 5 we have that with probability at least $\varepsilon^N/3$ over $X$ that $|B(X)|_2 < O_d(\varepsilon^{-2dN})$ and

$$\left|\bigwedge_{i_1,\ldots,i_k} \prod_{\ell=1}^{k} B_{i_\ell j_\ell}\right|_2 < O_{c,d,N}\big(\varepsilon^{k(1-c/2)}\big).$$

Applying Lemma 13 to $B$ at such values of $X$, we have that there are vectors $V^{\ell}, W^{\ell}$ so that

$$\left|B_{ij} - \sum_{\ell=1}^{k-1} V_i^{\ell} W_j^{\ell}\right|_2 = O_{c,d,N}\big(\varepsilon^{1-c/2}\big).$$

We note that we can replace the $V^\ell$ in such a decomposition with an orthonormal basis for the space that they span by adjusting the $W^\ell$ accordingly. We then have that

$$\begin{aligned}\sum_{\ell=1}^{k}|W_j^\ell|_2^2 &= \left|\sum_{\ell=1}^{k-1}V_i^\ell W_i^\ell\right|_2^2\\ &\le \left(|B|_2 + O_{c,d,N}(1)\right)^2\\ &\le O_{c,d,N}\left(\varepsilon^{-4dN}\right).\end{aligned}$$

Therefore, $|W_j^\ell|_2 \le O_{c,d,N}(\varepsilon^{-2dN})$ for each $\ell$.

Now given a standard Gaussian random vector, $Y$, there is a probability of at least $\Omega_k(\varepsilon^{2dkN+k})$ that $|Y_i V_i^\ell| \le \varepsilon^{2dN+1}$ for each $\ell$. Furthermore, by Corollary 5, for $\varepsilon$ sufficiently small the probability that

$$\left|\left(B_{ij} - \sum_{\ell=1}^{k-1}V_i^\ell W_j^\ell\right)Y_i\right|_2 > \varepsilon^{1-3c/4}$$

is much less than this. Hence, for such $X$ (which occur with probability at least $\varepsilon^N/2$), there is a probability of at least $\varepsilon^{O_{c,d,N}(1)}$ over $Y$ that

$$\left|\left(B_{ij} - \sum_{\ell=1}^{k-1}V_i^\ell W_j^\ell\right)Y_i\right|_2 < \varepsilon^{1-3c/4}$$

and

$$\left|Y_i V_i^\ell\right| \le \varepsilon^{2dN+1}$$

for each $\ell$. The latter implies that $|Y_i V_i^\ell W_j^\ell|_2 \le \varepsilon$ for each $\ell$, and thus,

$$|B_{ij}Y_i|_2 \le \left|\left(B_{ij} - \sum_{\ell=1}^{k-1}V_i^\ell W_j^\ell\right)Y_i\right|_2 + \sum_{\ell=1}^{k-1}|Y_i V_i^\ell W_j^\ell|_2 < \varepsilon^{1-c}.$$

Thus, with probability at least $\varepsilon^{O_{c,d,N}(1)}$,

$$\left|\nabla_i D_Y p(X)\right|_2 < \varepsilon^{1-c}. \qquad \square$$

Iterating Proposition 11 will tell us that a polynomial with a reasonable chance of having a small derivative will also have partial higher order derivatives that are small. Considering the $d$th order derivatives, this reduces to a statement about the rank-$d$ tensor corresponding to our polynomial. We would like to claim that such tensors can be approximately decomposed as a sum of products of lower rank tensors. In order to conveniently talk about such products we introduce some notation. If $S = \{a_1, \ldots, a_k\}$ is a set of natural numbers, we let $U_{i_S}$ denote a tensor on the indices $i_{a_1}, i_{a_2}, \ldots, i_{a_k}$.

PROPOSITION 14. *Let $d$ be an integer, and let $c, N, \varepsilon > 0$ be real numbers. Then for all rank-$d$ tensors $A$ with $|A|_2 \leq 1$ and*

$$\Pr_{X^1,\ldots,X^{d-1}}\left(\left|A_{i_1,\ldots,i_d}X^1_{i_1}\cdots X^{d-1}_{i_{d-1}}\right|_2 < \varepsilon\right) > \varepsilon^N.$$

*Then there exist tensors $U^\ell$, $V^\ell$, $1 \leq \ell \leq k = O_{c,d,N}(1)$ and sets*

$$\varnothing \subsetneq S(\ell) \subsetneq \{1,2,\ldots,d\}, \qquad \overline{S(\ell)} = \{1,2,\ldots,d\} - S(\ell)$$

*such that $|U^\ell|_2|V^\ell|_2 \leq O_{c,d,N}(\varepsilon^{-c})$ for all $\ell$ and*

$$\left|A_{i_1,\ldots,i_d} - \sum_{\ell=1}^{k} U^\ell_{i_{S(\ell)}} V^\ell_{i_{\overline{S(\ell)}}}\right|_2 = O_{c,d,N}\left(\varepsilon^{1-c}\right).$$

PROOF. We will instead prove the stronger claim that given $c, d, N, \varepsilon$ that there exists a probability distribution over sequences of tensor-valued polynomials $U^\ell$, $V^\ell$ of degree $O_{c,d,N}(1)$ in the coefficients of $A$, so that for any tensor $A$ satisfying the hypothesis of the proposition that with probability at least $\varepsilon^{O_{c,d,N}(1)}$ over our choice of $U^\ell$, $V^\ell$ in this family that

$$\left|U^\ell(A)\right|_2\left|V^\ell(A)\right|_2 \leq O_{c,d,N}\left(\varepsilon^{-c}\right)$$

for all $\ell$, and

$$\left|A_{i_1,\ldots,i_d} - \sum_{\ell=1}^{k} U^\ell_{i_{S(\ell)}}(A) V^\ell_{i_{\overline{S(\ell)}}}(A)\right|_2 = O_{c,d,N}\left(\varepsilon^{1-c}\right).$$

Given this statement, our proposition can be recovered by picking an appropriate set of $U^\ell$ and $V^\ell$ for our $A$. We assume throughout this proof that $\varepsilon$ is at most a sufficiently small function of $c, d$ and $N$, since otherwise there would be nothing to prove.

We prove this statement by induction on $d$. For $d = 1$, we already have that $|A_{i_1}|_2 < \varepsilon$, and there is nothing to prove. Hence, we assume that our statement holds for rank-$(d-1)$ tensors. The basic idea of our proof will be as follows. By assumption with reasonable probability over $X$, $AX$ will satisfy the inductive hypothesis for a rank-$(d-1)$ tensor. This means that we can write $U^\ell$ and $V^\ell$ as polynomials in $X$ so that with reasonable probability over $X$, $|AX - \sum U^\ell(X)V^\ell(X)|_2$ is small. Applying Lemmas 12 and 13, we can show that the derivative of this tensor with respect to $X$ is approximately low-rank. This means that the tensor

$$A - \sum_\ell \left(\nabla_{i_1} U^\ell(X)\right)V^\ell(X) + U^\ell(X)\left(\nabla_{i_1} V^\ell(X)\right)$$

is approximated by a small sum of products of rank-1 tensors with rank-$(d-1)$ tensors. By making some random guesses, these remaining tensors can be written as polynomials in the coefficients of $A$ with reasonable probability.

Suppose that $A$ is a rank-$d$ tensor satisfying the hypothesis of our proposition. Then with probability at least $\varepsilon^N/2$ over a choice of $X^1$, there is a probability of at least $\varepsilon^N/2$ over our choice of $X^2, \dots, X^{d-1}$ that

$$|A_{i_1,\dots,i_d} X^1_{i_1} \cdots X^{d-1}_{i_{d-1}}|_2 < \varepsilon.$$

Furthermore, by Corollary 5, with probability at least $1 - \varepsilon^N/4$ we have that $|A_{i_1,\dots,i_d} X^1_{i_1}|_2 < \varepsilon^{-c/20}$. Hence, with probability at least $\varepsilon^N/4$ over our choice of $X^1$, $\varepsilon^{c/20} A_{i_1,\dots,i_d} X^1_{i_1}$ satisfies the hypotheses of our proposition as a rank-$(d-1)$ tensor. For each such $X^1$, the induction hypothesis implies that there is a probability of $\varepsilon^{O_{c,d,N}(1)}$ over our choice of $U^\ell$, $V^\ell$ that the appropriate conclusion holds. Therefore, there must be some particular choice of $U^\ell$, $V^\ell$ so that with probability at least $\varepsilon^{O_{c,d,N}(1)}$ over our choice of $X^1$ we have that

$$|U^\ell(\varepsilon^{c/20} A X^1_{i_1})|_2 |V^\ell(\varepsilon^{c/20} A X^1_{i_1})|_2 \le O_{c,d,N}(\varepsilon^{-c/20})$$

and

$$\left| \varepsilon^{c/20} A X^1_{i_1} - \sum_{\ell=1}^{k} U^\ell_{i_{S(\ell)}}(\varepsilon^{c/20} A X^1_{i_1}) V^\ell_{i_{\overline{S(\ell)}}}(\varepsilon^{c/20} A X^1_{i_1}) \right|_2 = O_{c,d,N}(\varepsilon^{1-c/20}).$$

Letting $U'^\ell(X^1) := \varepsilon^{-c/40} U^\ell(\varepsilon^{c/20} A X^1)$ and $V'^\ell(X^1) := \varepsilon^{-c/40} V^\ell(\varepsilon^{c/20} A X^1)$, we can rephrase the last two equations as

$$|U'^\ell(X^1)|_2 |V'^\ell(X')|_2 \le O_{c,d,N}(\varepsilon^{-c/10})$$

and

$$\left| A X^1_{i_1} - \sum_{\ell=1}^{k} U'^\ell_{i_{S(\ell)}}(X^1) V'^\ell_{i_{\overline{S(\ell)}}}(X^1) \right|_2 = O_{c,d,N}(\varepsilon^{1-c/10}).$$

We will demonstrate that given a correct choice of such $U'^\ell$ and $V'^\ell$ we can construct new polynomials $U^\ell(A)$, $V^\ell(A)$ that satisfy the necessary conditions with probability at least $\varepsilon^{O_{c,d,N}(1)}$.

Let $T_i(X^1)$ be the tensor-valued polynomial whose coefficients are the concatenation of the coefficients of

$$A X^1_{i_1} - \sum_{\ell=1}^{k} U'^\ell_{i_{S(\ell)}}(X^1) V'^\ell_{i_{\overline{S(\ell)}}}(X^1)$$

and the coefficients of the $\varepsilon U'^\ell(X^1)$ and $\varepsilon V'^\ell(X^1)$. We have that for some $N_1 = O_{c,d,N}(1)$ that with probability at least $\varepsilon^{N_1}$ that $|T_i(X^1)|_2 < O_{c,d,N}(\varepsilon^{1-c/10})$. We apply Lemma 12 with $k' > 10N_1/c$ and then Lemma 13 (as in the proof of Proposition 11) to show that there exist tensors $W^\ell$, $Z^\ell$ so that

$$\left| \nabla_j T_i(X^1) - \sum_{\ell=1}^{k'-1} W^\ell_i Z^\ell_j \right|_2 \le O_{c,d,N}(\varepsilon^{1-3c/20}).$$

Note that by considering only the coordinates of $T$ that correspond to entries of $A - \sum U'V'$, we have that for appropriate values of $X^1$ that

$$\left| A - \left[ \sum_{\ell=1}^{k} V'^{\ell}_{i_{\overline{S(\ell)}}}(X^1) \nabla_{i_d} U'^{\ell}_{i_{S(\ell)}}(X^1) + U'^{\ell}_{i_{S(\ell)}}(X^1) \nabla_{i_d} V'^{\ell}_{i_{\overline{S(\ell)}}}(X^1) + \sum_{\ell=1}^{k'-1} W^{\ell}_{i_1,\ldots,i_{d-1}} Z^{\ell}_{i_d} \right] \right|_2 \tag{5}$$

is $O_{c,d,N}(\varepsilon^{1-3c/20})$. This is nearly enough to complete our proof as we have shown that $A$ can be approximated by a sum of a bounded number of products of lower rank tensors. However, for our inductive hypothesis to hold, we need to verify that the above can be obtained with reasonable probability while taking $W^{\ell}$ and $Z^{\ell}$ to be probabilistic polynomials in the coefficients of $A$. In order to analyze this, let $S_{ij}$ be $\nabla_i T_j(X^1)$ restricted to the coordinates $j$ for which $T_j$ corresponds to a coordinate of $A(X^1)$.

We know that with probability $\varepsilon^{O_{c,d,N}(1)}$ over the choice of $X^1$ that for some $Z^{\ell}$, $W^{\ell}$ that

$$\left| S_{ij} - \sum_{\ell=1}^{k'-1} Z^{\ell}_i W^{\ell}_j \right|_2 = O_{c,d,N}\left(\varepsilon^{1-3c/20}\right) \tag{6}$$

and that $|T(X^1)|_2 = O_{c,d,N}(\varepsilon^{1-3c/20})$. The latter implies that $|U^{\ell}|_2, |V^{\ell}|_2 = O_{c,d,N}(\varepsilon^{-3c/20})$ and, therefore, that $|S_{ij}|_2 = O(\varepsilon^{-c})$.

We wish to show that for such $S$, we can satisfy equation (6) with decent probability by taking $Z^{\ell}$ and $W^{\ell}$ to be specific random polynomials in the coefficients of $S$. In particular, we show the following.

LEMMA 15. *There exists an explicit probability distribution over vector valued polynomials $Z^{\ell}(S)$ and $W^{\ell}(S)$ so that for any tensor $S$ with $|S|_2 = O(\varepsilon^{-c})$ and so that equation (6) holds for some vectors $Z^{\ell}$ and $W^{\ell}$, then*

$$\left| S_{ij} - \sum_{\ell=1}^{k'-1} Z^{\ell}_i W^{\ell}_j \right|_2 = O_{c,d,N}\left(\varepsilon^{1-c/5}\right),$$

*and $|Z^{\ell}(S)|_2, |W^{\ell}(S)|_2 = O(\varepsilon^{-c})$ with probability at least $\varepsilon^{O_{c,d,N}(1)}$.*

PROOF. First, we note that we may write

$$S_{ij} = \sum_{\ell=1}^{k'-1} Z^{\ell}_i W^{\ell}_j + E_{ij},$$

where $|E_{ij}|_2 = O_{c,d,N}(\varepsilon^{1-3c/20})$. Replacing the $W^\ell$'s and $Z^\ell$'s by linear combinations and employing the theory of singular values we may instead write

$$S_{ij} = \sum_{\ell=1}^{k'-1} C^\ell Z_i^\ell W_j^\ell + E_{ij},$$

where now $\{Z^\ell\}$ and $\{W^\ell\}$ are orthonormal sets and $C^\ell$ are some nonnegative integers. We note that $|S_{ij}|_2 \geq C^\ell - |E_{ij}|$, and thus $C^\ell = O(\varepsilon^{-1})$ for all $\ell$. Furthermore, if $C^\ell < \varepsilon^{1-3c/20}$ for any $\ell$, we may remove that term from the sum and add it to $E_{ij}$. Thus, we may assume that $C^\ell > \varepsilon^{1-3c/20}$ for all $\ell$.

Our basic strategy is as follows: by taking dot products of $S$ with random vectors, there is a decent probability that we get very close approximations to $Z^\ell$ and $W^\ell$. Taking an appropriate combination gives our result. In particular, let $X_i^\ell$ and $Y_j^\ell$ be Gaussian random vectors and let $C'^\ell$ be random numbers chosen uniformly from $[0, \varepsilon^{-2}]$. With probability $\varepsilon^{O_{k'}(1)}$, all of the following hold:

- For all $a, b$, we have $|X_i^a Z_i^b - \delta_{a,b}| = O(\varepsilon^5)$.
- For all $a, b$, we have $|Y_j^a W_j^b - \delta_{a,b}| = O(\varepsilon^5)$.
- For all $\ell$, we have $|C^\ell - C'^\ell| = O(\varepsilon^5)$.
- For all $\ell$, $|E_{ij} X_i^\ell|_2, |E_{ij} Y_j^\ell|_2 = O(\varepsilon^{1-c/5})$.

This holds because all of the first three types of events listed [for each possible value of $(a, b)$ or $\ell$] are independent and occur with probability $\varepsilon^{O(1)}$, and the last holds with high probability. We let

$$Z_i^\ell(S) = S_{ij} Y_j^\ell (C'^\ell)^{-1/2},$$

and

$$W_j^\ell(S) = S_{ij} X_i^\ell (C'^\ell)^{-1/2}.$$

Given the assumptions above, this means that

$$\begin{aligned}
Z_i^\ell(S) &= S_{ij} Y_j^\ell (C'^\ell)^{-1/2} \\
&= \left( \sum_{l=1}^{k'-1} C^l Z_i^l W_j^l Y_j^\ell + E_{ij} Y_j^\ell \right) (C'^\ell)^{-1/2} \\
&= \left( \sum_{l \neq \ell}^{k'-1} C^l Z_i^l O(\varepsilon^5) + Z_i^\ell (1 + O(\varepsilon^5)) + O(\varepsilon^{1-c/5}) \right) (C^\ell)^{-1/2} (1 + O(\varepsilon^4)) \\
&= (C^\ell)^{1/2} Z_i^\ell + O(\varepsilon^{1-c/5} (C^\ell)^{-1/2}).
\end{aligned}$$

Similarly,

$$W_j^\ell(S) = (C^\ell)^{1/2} W_j^\ell + O(\varepsilon^{1-c/5} (C^\ell)^{-1/2}).$$

Thus,

$$\sum_{\ell=1}^{k'-1} Z_i^\ell(S) W_j^\ell(S) = \sum_{\ell=1}^{k'-1} C^\ell Z_i^\ell W_j^\ell + O\big(\varepsilon^{1-c/5}\big) = S_{ij} + O_{k'}\big(\varepsilon^{1-c/5}\big).$$

Furthermore, under the given assumptions $|Z^\ell(S)|_2, |W^\ell(S)|_2$ satisfy appropriate bounds. This completes the proof. □

Using $Z^\ell(S)$ and $W^\ell(S)$ in equation (5), completes the inductive step and thus completes the proof. □

We are finally ready to prove Proposition 10.

PROOF. Assume that $\varepsilon$ is sufficiently small as a function of $c, d$ and $N$ (for otherwise there is nothing to prove).

Consider such a polynomial $p$. We claim that for any $k < d$ and any $c' > 0$ that

$$\Pr_{X,Y^1,\ldots,Y^k}\big(\big|\nabla_i \nabla_{Y^1} \cdots \nabla_{Y^k} p(X)\big|_2 < \varepsilon^{1-c'}\big) > \varepsilon^{O_{k,c',d,N}(1)}.$$

This is proved by induction on $k$. The $k = 0$ case is given and the inductive step follows immediately from Proposition 11. Applying this statement for $k = d - 1$, we note that

$$\nabla_i \nabla_{Y^1} \cdots \nabla_{Y^k} p(X)$$

is independent of $X$. Let $A_{i_1,\ldots,i_d} = \nabla_{i_1} \cdots \nabla_{i_d} p(X)$ be the symmetric $d$-tensor associated to $p$. We have by Lemma 9 that $|A|_2 = \sqrt{d!}|p^{[d]}|_2 \le \sqrt{d!}$, and thus, $A/d!$ satisfies the hypothesis of Proposition 14. Hence, we can find tensors $U^\ell$ and $V^\ell$ with the properties specified by that proposition so that $|U^\ell|_2 |V^\ell|_2 \le O_{c,d,N}(\varepsilon^{-c})|p^{[d]}|_2$. Since $A$ is symmetric, we have that

$$\left| A - \sum_{\sigma \in S_d} \sum_{\ell=1}^{k} U^\ell_{i_{\sigma(S(\ell))}} V^\ell_{i_{\sigma(\overline{S(\ell)})}} \right|_2 = O_{c,d,N}\big(\varepsilon^{1-c}\big). \tag{7}$$

Note that in the above since the sum over $\sigma$ has already added the permutations of $U^\ell$ over its indices, we may replace $U^\ell$ and $V^\ell$ by their symmetrizations without affecting the above sum. Let $U^\ell$ be rank $d_\ell$ and $V^\ell$ be rank $d - d_\ell$. Let $a_\ell(X)$ be the degree-$d_\ell$ harmonic part of the polynomial $X \to U^\ell(X, X, \ldots, X)$. Define $b_\ell(X)$ similarly with respect to $V^\ell$. By Lemma 9, we have that $|a_\ell|_2 |b_\ell|_2 \le d!|U^\ell|_2 |V^\ell|_2 = O_{c,d,N}(\varepsilon^{-c})|p^{[d]}|_2$. Now consider the tensor given by

$$\nabla_{i_1} \nabla_{i_2} \cdots \nabla_{i_d} \left[ p(X) - \sum_{\ell=1}^{k} a_\ell(X) b_\ell(X) \right].$$

This is easily seen to be the tensor given in equation (7), and hence has size $O_{c,d,N}(\varepsilon^{1-c})$. On the other hand by Lemma 9, this can be seen to be $\sqrt{d!}$ times the size of the degree-$d$ harmonic part of the polynomial

$$p(X) - \sum_{\ell=1}^{k} a_\ell(X) b_\ell(X).$$

This completes our proof. □

3.3. *Proof of the main theorem.* We are now prepared to prove the diffuse decomposition theorem. The basic idea of the proof is fairly simple. We maintain decompositions of polynomials approximately equal to $p$. We show using Proposition 10 that if this decomposition is not diffuse that we can replace it by a simpler one by introducing at most a small error. This new decomposition is simpler in the sense that an associated ordinal number is smaller, and we will use transfinite induction to prove that this process will eventually terminate, yielding an appropriate decomposition.

PROOF OF THEOREM 1. We assume for convenience that $N$ and $c^{-1}$ are integers. Throughout we will assume that $N, c, d$ and $\varepsilon$ are fixed.

We define a *partial decomposition* of our polynomial $p$ to be a set of the following data:

- A positive integer $m$.
- A polynomial $h : \mathbb{R}^m \to \mathbb{R}$.
- A sequence of polynomials $(q_1, \ldots, q_m)$ each on $\mathbb{R}^n$ with $|q_i|_2 = 1$ for each $i$.
- A sequence of integers $(a_1, \ldots, a_m)$ with $a_i$ between 0 and $4 \cdot 3^i(N+1)/c - 1$.

Furthermore, we require that each $q_i$ is nonconstant, and that for any monomial $\prod x_i^{\alpha_1}$ appearing in $h$ that $\sum \alpha_i \deg(q_i) \le d$.

We say that such a partial decomposition has complexity at most $C$ if the following hold:

- $m \le C$.
- $|h|_2 \le C\varepsilon^{-1+C^{-1}}$.
- $|p(X) - h(\varepsilon^{a_1 c/(2\cdot 3^1)} q_1(X), \varepsilon^{a_2 c/(2\cdot 3^2)} q_2(X), \ldots, \varepsilon^{a_m c/(2\cdot 3^m)} q_m(X))|_2 \le C\varepsilon^N$.

Finally, we define the weight of a partial decomposition as follows. First, we define the polynomial

$$w(x) = \sum_{i=1}^{m} x^{\deg(q_i)}\big(4 \cdot 3^i(N+1)/c - a_i\big).$$

We then let the weight of the decomposition be $w(\omega)$.

Our result will follow from the following lemma.

LEMMA 16. *Let $p$ be a degree-$d$ polynomial with a partial decomposition of weight $w$ and complexity at most $C$. Then there exists a polynomial $p_0$ with an $(\varepsilon, \varepsilon^{-c})$-diffuse decomposition of size at most $O_{c,d,N,w,C}(1)$ so that $|p - p_0|_2 \leq O_{c,d,N,w,C}(\varepsilon^N)$.*

PROOF. We prove this by transfinite induction on $w$. In particular, we show that either $(h, q_1, \ldots, q_m)$ provides an appropriate diffuse decomposition of a polynomial approximately equal to $p$ or that $p$ has another partial decomposition of complexity $O_{c,d,N,C,w}(1)$ and weight strictly less than $w$ (with finitely many possibilities for the new weight). The inductive hypothesis will imply that we have an appropriate diffuse decomposition in the latter case.

First, note that if some $a_i$ at least $2(N+1)3^i/c$ that the sum of the coefficients of $q_i$ appearing in

$$h\big(\varepsilon^{a_1 c/(2\cdot 3^1)} q_1(X), \varepsilon^{a_2 c/(2\cdot 3^2)} q_2(X), \ldots, \varepsilon^{a_m c/(2\cdot 3^m)} q_m(X)\big)$$

is $O_C(\varepsilon^N)$. Thus, these terms can be thrown away without introducing an error of more than $O_{C,d}(\varepsilon^N)$. Doing so to the largest such $q_i$ and shifting all of the larger indices down, perhaps changing the $a_i$, and modifying $h$ appropriately will lead to a new partial decomposition with a new value of $C$ dependent only on $d$ and the old one, and a strictly smaller weight. Hence, we assume that $a_i < 2(N+1)3^i/c$ for all $i$.

If $\deg(q_{i+1}) > \deg(q_i)$ for some $i$, we may swap $q_i$ and $q_{i+1}$ (making a similar adjustment to $h$ and modifying $a_i$ and $a_{i+1}$ as necessary) to get a partial decomposition of complexity $C$ and strictly smaller weight. Hence, we may assume that $\deg(q_1) \geq \deg(q_2) \geq \cdots \geq \deg(q_m)$.

Were it the case that for all $x_1, \ldots, x_m$ that

$$\Pr\big(|q_i(X) - x_i| < \varepsilon \text{ for all } i\big) < \varepsilon^{m-c},$$

then we would already have an appropriate diffuse decomposition and would be done. Hence, we may assume that there is a set of $x_i$ so that the above does not hold. By Proposition 6, we have that with probability at least $1 - \varepsilon^{m-c}/2$ that

$$\prod_{i=1}^m |q_i(X) - x_i| \geq \Omega_{C,d}\big(\varepsilon^{m-c/2}\big) \left| \bigwedge_{j_1,\ldots,j_m} \prod_{i=1}^m \nabla_{j_i} q_i(X) \right|_2.$$

Thus, with probability at least $\varepsilon^{m-c}/2$ both of the above hold, which would imply that

$$\left| \bigwedge_{j_1,\ldots,j_m} \prod_{i=1}^m \nabla_{j_i} q_i(X) \right|_2 = O_{C,d}\big(\varepsilon^{c/2}\big).$$

Now the wedge product above is a wedge product of vectors, and hence its size is unchanged by making a determinant 1 change of basis to the vectors $\nabla_{j_i} q_i$. Hence,

letting $V^i$ be the projection of $\nabla q_i$ onto the orthogonal compliment of the space spanned by the $\nabla q_j$ for $j > i$ we have that the size of the wedge product equals $\prod_{i=1}^m |V^i|_2$. This means that for some $i$ that $|V^i|_2 \leq O_{C,d}(\varepsilon^{c/3^i})$. Hence, for some $i$, we have with probability at least $\Omega_{C,d}(\varepsilon^m)$ over $X$ that $|V^i(X)|_2 \leq O_{C,d}(\varepsilon^{c/3^i})$, and that this is the largest $i$ for that $X$ for which this holds. Furthermore, by Lemma 9 and Corollary 5 we know that when this happens with high probability we also have that the first derivatives of all the $q_i$ have size $O_{C,d}(\log(\varepsilon^{-1})^d)$.

When the above happens, $V^j$ is given by the derivative of $q_j$ minus an appropriate linear combination of the $V^k$ for $k > j$. Note that for each coefficient, the size of the coefficient times the size of $V^k$ is at most the size of the derivative of $q_j$. Hence, for $k > i$, the size of the coefficient is at most $O_{C,d}(\varepsilon^{-c/3^k}\log^d(\varepsilon^{-1}))$. From this, it is easy to see that $V^i$ is given by a linear combination of the derivatives of the $q_j$ with $j \geq i$ such that the $i$th coefficient is 1 and that all other coefficients have size at most

$$\prod_{k=i+1}^{m} O_{C,d}\big(\varepsilon^{-c/3^k}\log^d(\varepsilon^{-1})\big) = O_{C,d}\big(\varepsilon^{-c/(2\cdot 3^i)+c/(2\cdot 3^m)}\big).$$

Hence, for each such $X$, there are constants $C_j = O_{C,d}(\varepsilon^{-c/(2\cdot 3^i)+c/(2\cdot 3^m)})$ (for $j > i$) so that the derivative of $q_i + \sum_j C_j q_j$ at $X$ has size at most $O_{C,d}(\varepsilon^{c/3^i})$. Note that this statement still holds if the $C_j$ are rounded to the nearest multiple of $\varepsilon$. Since there are $\varepsilon^{-O_C(1)}$ such possible roundings, there is some set of $C_j$ so that for the polynomial $q(X) = q_i(X) + \sum_j C_j q_j(X)$, we have that $|\nabla_j q(X)|_2 \leq O_{C,d}(\varepsilon^{c/3^i})$ with probability at least $\varepsilon^{O_{C,d}(1)}$ over $X$.

We now can apply Proposition 10 to $\Omega_{C,d}(\varepsilon^{c/(2\cdot 3^i)-c/(4\cdot 3^m)})q(X)$. Let $D = \deg(q_i)$. Let $Q(X)$ be the degree-$D$ harmonic part of $\Omega_{C,d}(\varepsilon^{c/(2\cdot 3^i)-c/(2\cdot 3^m)})q(X)$. Proposition 10 tells us that there are polynomials $A_\ell, B_\ell$ of degree strictly less than $D$ with $|A_\ell|_2|B_\ell|_2$ at most $O_{c,C,d}(\varepsilon^{-1/(2C^2d)})|Q|_2$ for each $\ell$, and so that $Q - \sum_\ell A_\ell B_\ell$ equals a polynomial of degree less than $D$ plus an error of $L^2$ norm at most $O_{c,C,d}(\varepsilon^{c/3^i - c/(2\cdot 3^m)})$. Note that the lower degree polynomial has size at most

$$|Q|_2 + \sum |A_\ell B_\ell|_2.$$

By Corollary 5 and Hölder's inequality, we have that

$$|A_\ell B_\ell|_2 \leq |A_\ell|_4|B_\ell|_4 \leq O_d\big(|A_\ell|_2|B_\ell|_2\big) = O_{c,C,d}\big(|Q|_2\varepsilon^{-1/(2C)}\big).$$

Consider the $j$ among those for which $\deg(q_j) = D$ for which $C_j\varepsilon^{-a_j c/(2\cdot 3^j)}$ is the largest. $Q(X)$ is then some multiple of the degree-$D$ harmonic part of $q_j\varepsilon^{a_j c/(2\cdot 3^j)}$ plus smaller multiples of the degree-$D$ harmonic parts of other $q_k\varepsilon^{a_k c/(2\cdot 3^k)}$.

We are now ready to modify our partial decomposition to obtain one of smaller weight. First, we take each of the $q_i$ of degree equal to $D$ and replace $q_i$ by the sum of its harmonic degree-$D$ part and the remainder, introducing each as a new $q_j$. This increases the complexity by at most a factor of $2^D$, and increases the weight by an ordinal less than $\omega^D$.

Next, we note that $q_j\varepsilon^{a_jc/(2\cdot 3^j)}$ can be written as a linear combination of the other $q_k\varepsilon^{a_kc/(2\cdot 3^k)}$ (with coefficients less than 1) plus a sum of $A_\ell(X)B_\ell(X)$ plus a polynomial of degree less than $D$ plus a degree-$D$ polynomial of size at most $O_{c,C,d}(\varepsilon^{(a_j+1)c/(2\cdot 3^j)})$. Replacing $q_j$ by a normalized version of this error polynomial, and adding new $q$'s corresponding to the normalized versions of $A_\ell$ and $B_\ell$ and the remaining part of degree less than $D$ and modifying $h$ appropriately, we find that we have a new partial decomposition of weight smaller by $\omega^D - O_{c,C,d}(\omega^{D-1})$. Thus, our new decomposition has smaller weight since the coefficient of $\omega^D$ is strictly smaller than before and the higher degree coefficients are no bigger.

The last thing that we need to check is that this new decomposition has complexity bounded solely in terms of $C, c, d$ and $N$. It is clear from the construction that $m$ increases by at most a bounded amount and that the error between $p$ and $h(\varepsilon^{a_ic/(2\cdot 3^i)}q_i)$ remains the same. However, we need to show that the size of $h$ does not increase by too much. For this we need to analyze more carefully what we are doing to the function $h$. The idea is that we have a relation of the form

$$\varepsilon^{a_jc/(2\cdot 3^j)}q_j = \sum a_k\varepsilon^{a_kc/(2\cdot 3^j)}q_k + \varepsilon^{(a_j+1)c/(2\cdot 3^j)}q'_j + \sum_{\ell=1}^{K} A_\ell B_\ell,$$

where the first sum is over $k \neq j$ with $\deg(q_k) = D$, $q'_j$ is the error term and $|A_\ell|_2, |B_\ell|_2 \leq 1$. The new version of $h$ is now obtained from the old by replacing every occurrence of the $j$th coordinate, $x_j$, by $x_j + \sum a_k x_k + \sum_\ell s_\ell x_{m+2\ell-1}x_{m+2\ell}$, where $s_\ell = |A_\ell|_2|B_\ell|_2$. We note that this replacement increases the size of $h$ by at most $O(1 + \sum|a_k| + \sum|s_\ell|)^d$. Thus, it suffices to show that $\sum|a_k| + \sum|s_\ell| = O_{c,C,d}(\varepsilon^{-1/2C^2d})$. On the one hand, it should be noted that $|a_k| \leq 1$ by assumption. Thus, $\sum|a_k| \leq C$. We have left to deal with the $|s_\ell|$ terms. By assumption, each $|s_\ell|$ is $O_{c,C,d}(\varepsilon^{-1/2C^2d})$ and the number of them is $O_{c,C,d}(1)$. Thus, the sum is appropriately bounded, and the complexity of the new decomposition is at most $O_{c,C,d}(1)$. This completes the proof. □

Our theorem follows from applying Lemma 16 to the partial decomposition $m = 1$, $h(x_1) = |p|_2x_1$, $q_1(X) = p(X)/|p|_2$ and $a_1 = 0$ of complexity 1 and weight $[6(N+1)/c]\omega^d$. □

**4. Basic facts about diffuse decompositions.** The primary use of a diffuse decomposition will be that the existence of a diffuse decomposition will allow

us to approximate the corresponding threshold function by a smooth function. In particular, we show the following.

PROPOSITION 17. *Let $(h, q_1, \dots, q_m)$ by an $(\varepsilon, N)$-diffuse decomposition of a degree-$d$ polynomial $p$ for $1/2 > \varepsilon > 0$. There exists a function $f : \mathbb{R}^m \to [-1, 1]$ so that*:

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

In order to prove this and for some other applications, we will also need the following statement about the distribution of values of $(q_i(X))$ in a diffuse decomposition.

LEMMA 18. *Let $(h, q_1, \dots, q_m)$ be an $(\varepsilon, N)$-diffuse decomposition of a degree-$d$ polynomial for some $1/2 > \varepsilon > 0$. Letting $Q = (q_1(X), q_2(X), \dots, q_m(X))$ for $X$ a random Gaussian we have that with probability $1 - O_{m,d}(N\varepsilon \log(\varepsilon^{-1})^{dm/2+1})$ that*

$$\big|h(Q)\big| \geq \varepsilon\big|\nabla_{i_1} h(Q)\big|_2 \geq \varepsilon^2\big|\nabla_{i_2}\nabla_{i_2} h(Q)\big|_2 \geq \cdots \geq \varepsilon^d\big|\nabla_{i_1}\cdots\nabla_{i_d} h(Q)\big|_2. \tag{8}$$

PROOF. First, we note that for some $B = O_m(\log(\varepsilon^{-1})^{d/2})$ that by Corollary 5 that $|q_i(X)| \leq B$ for all $i$ with probability at least $1 - \varepsilon$. Hence, it suffices to bound the probability that equation (8) fails while $|q_i(X)| \leq B$ for all $i$. We let $R \subset \mathbb{R}^m$ be the region for which this fails. We bound the probability that $x \in R$ by covering $R$ by axis aligned boxes of side length $2\varepsilon$ and using the fact that $(q_1, \dots, q_m)$ is a diffuse set. In particular, consider the union of all axis aligned boxes of side length $2\varepsilon$ whose endpoints are integer multiples of $2\varepsilon$ and which contain some point of $R$. Call the union of all such boxes $R'$. Note that since $R'$ is a disjoint union of boxes so that for each such box the probability that $x$ lies in this box is at most $N$ times its volume, we have that $\Pr(x \in R) \leq \Pr(x \in R') \leq N\operatorname{Vol}(R')$. Let $R''$ be the set of points $y \in \mathbb{R}^m$ so that $y$ is within $2\sqrt{m}\varepsilon$ of some point in $R$. Note that $R'' \supset R'$. Thus, it suffices to prove that

$$\operatorname{Vol}\big(R''\big) = O_{m,d}\big(\varepsilon \log\big(\varepsilon^{-1}\big)^{dm/2+1}\big).$$

Let $Y$ be an $m$-dimensional Gaussian. Note that $R''$ is contained in a ball of radius $O_m(B)$. Hence, since the probability density function of $BY$ is at least

$\Omega_m(B^{-m}\,dV)$ on this region, we have that $\mathrm{Vol}(R'') = O_m(B^m \Pr(BY \in R''))$. Define the polynomial $H(x) = h(Bx)$. It now suffices to show that with probability at most $O_{d,m}(\varepsilon \log(\varepsilon^{-1}))$ that $Y$ is within $O_m(\varepsilon)$ of a point, $x$ for which

$$\big|H(x)\big| \geq \varepsilon \big|\nabla_{i_1} H(x)\big|_2 \geq \varepsilon^2 \big|\nabla_{i_2}\nabla_{i_2} H(x)\big|_2 \geq \cdots \geq \varepsilon^d \big|\nabla_{i_1}\cdots\nabla_{i_d} H(x)\big|_2$$

fails to hold.

Note that by Proposition 6 for $k = 1$ we have that for any $1/2 > \delta > 0$ that with probability $1 - O_{d,m}(\delta \log(\delta^{-1}))$,

$$\big|H(Y)\big| \geq \delta \big|\nabla_{i_1} H(Y)\big|_2 \geq \delta^2 \big|\nabla_{i_2}\nabla_{i_2} H(Y)\big|_2 \geq \cdots \geq \delta^d \big|\nabla_{i_1}\cdots\nabla_{i_d} H(Y)\big|_2.$$

If the above holds and $x = Y + z$ for $|z|_2 = O_m(\varepsilon)$, we have by Taylor's theorem that

$$\nabla_{i_1}\cdots\nabla_{i_k} H(x) = \nabla_{i_1}\cdots\nabla_{i_k} H(Y) + \sum_{t=1}^{d-k} \frac{(\nabla_{i_1}\cdots\nabla_{i_{k+t}} H(Y)) z_{i_{k+1}}\cdots z_{i_{k+t}}}{t!}.$$

Hence, we have that

$$\begin{aligned}
&\big|\nabla_{i_1}\cdots\nabla_{i_k} H(x) - \nabla_{i_1}\cdots\nabla_{i_k} H(Y)\big|_2 \\
&\quad\leq \sum_{t=1}^{d-k} \left| \frac{(\nabla_{i_{k+1}}\cdots\nabla_{i_{k+t}} H(Y)) z_{i_{k+1}}\cdots z_{i_{k+t}}}{t!} \right|_2 \\
&\quad\leq \sum_{t=1}^{d-k} \frac{|(\nabla_{i_1}\cdots\nabla_{i_{k+t}} H(Y))|_2 |z|_2^t}{t!} \\
&\quad\leq \sum_{t=1}^{d-k} \frac{|\nabla_{i_1}\cdots\nabla_{i_k} H(Y)|_2 (|z|_2/\delta)^t}{t!} \\
&\quad\leq \big|\nabla_{i_1}\cdots\nabla_{i_k} H(Y)\big|_2 \big(\exp\big(O_m\big(|z|_2/\delta\big)\big) - 1\big).
\end{aligned}$$

Thus, if $\delta = 4\sqrt{m}\varepsilon$ and the above holds [which it does with probability $1 - O_{d,m}(\varepsilon \log(\varepsilon^{-1}))$], then for any point $x$ within $2\sqrt{m}\varepsilon$ of $Y$ we have that

$$\big|\nabla_{i_1}\cdots\nabla_{i_k} H(x) - \nabla_{i_1}\cdots\nabla_{i_k} H(Y)\big|_2 \leq \big|\nabla_{i_1}\cdots\nabla_{i_k} H(Y)\big|_2 \big(e^{1/2} - 1\big),$$

and thus, equation (8) holds. Thus, $\Pr(BY \in R'') \leq O_{d,m}(\varepsilon \log(\varepsilon^{-1}))$, so $\mathrm{Vol}(R'') = O_{d,m}(\varepsilon \log(\varepsilon^{-1})^{dm/2+1})$, completing our proof. □

COROLLARY 19. *Let $(h, q_1, \ldots, q_m)$ be an $(\varepsilon, N)$-diffuse decomposition of a degree-$d$ polynomial for $1/2 > \varepsilon > 0$. Letting $Q = (q_1(X), q_2(X), \ldots, q_m(X))$ for $X$ a random Gaussian, the probability that $Q$ is within $\varepsilon$ of a point $y$ for which $h(y) = 0$ is $O_{d,m}(N\varepsilon \log(\varepsilon^{-1})^{dm/2+1})$.*

PROOF. Note that by the analysis given above, if equation (8) holds for $2\varepsilon$ then for any $y$ with $|Q - y| \le \varepsilon$

$$|h(Q) - h(y)| \le |h(Q)|\big(e^{1/2} - 1\big) < |h(x)|.$$

Thus, as long as $h(Q) \neq 0$ (which happens with probability 1) $h(y) \neq 0$. Since an $(\varepsilon, N)$-diffuse decomposition is also an $(2\varepsilon, 2^m N)$-diffuse decomposition, this happens with probability at least $1 - O_{d,m}(N\varepsilon \log(\varepsilon^{-1})^{dm/2+1})$ by Lemma 18. □

PROOF OF PROPOSITION 17. We construct $f$ in a straightforward manner. Let $\rho : \mathbb{R}^m \to \mathbb{R}$ be any smooth, nonnegative-valued, function supported on the ball of radius 1 so that

$$\int_{\mathbb{R}^m} \rho(x)\,dx = 1.$$

Let $\rho_\varepsilon(x) = \varepsilon^{-m}\rho(\varepsilon^{-1}x)$. We note that

$$\int_{\mathbb{R}^m} \rho_\varepsilon(x)\,dx = 1.$$

Let $g : \mathbb{R}^m \to \mathbb{R}$ be the function

$$g(x) = \begin{cases} 1, & \text{if there exists a } y \in \mathbb{R}^n \text{ so that } |x - y| < \varepsilon \text{ and } h(y) \ge 0, \\ -1, & \text{otherwise.} \end{cases}$$

We let $f$ be the convolution $g * \rho_\varepsilon$.

$f$ takes values in $[-1, 1]$ because

$$f(x) = \int_{\mathbb{R}^m} \rho_\varepsilon(y) g(x - y)\,dy \le \int_{\mathbb{R}^m} \rho_\varepsilon(y)\,dy = 1$$

and similarly $f(x) \ge -1$.

$f(q_1, \ldots, q_m)$ is a point-wise upper bound for $\operatorname{sgn} \circ p = \operatorname{sgn}(h(q_1, \ldots, q_m))$ since if $h(x) \ge 0$ then

$$f(x) = \int_{\mathbb{R}^m} \rho_\varepsilon(y) g(x - y)\,dy \int_{B(\varepsilon)} \rho_\varepsilon(y) g(x - y)\,dy = \int_{B(\varepsilon)} \rho_\varepsilon(y)\,dy = 1.$$

Bounds on the derivatives of $f$ come from the fact that

$$|f^{(k)}|_\infty = |g * \rho_\varepsilon^{(k)}|_\infty \le |g|_\infty |\rho_\varepsilon^{(k)}|_1 = O_{m,k}(\varepsilon^{-k}).$$

The fact that $f$ and $\operatorname{sgn} \circ p$ have similar expectations follows from the fact that $f(x) = \operatorname{sgn}(h(x))$ unless $x$ is within $2\varepsilon$ of a point $y$ for which $h(y) = 0$. Noting that an $(\varepsilon, N)$-diffuse decomposition of size $m$, is also a $(2\varepsilon, 2^m N)$-diffuse decomposition; this happens with probability $O_{d,m}(N\varepsilon \log(\varepsilon^{-1})^{dm/2+1})$. Since $|f(x) - \operatorname{sgn}(h(x))|$ is never more than 2, this provides the necessary bound. □

Another lemma that will be useful to us is the following.

LEMMA 20. *Let $(h, q_1, \ldots, q_m)$ be an $(\varepsilon, N)$-diffuse decomposition of a degree-$d$ polynomial $p$ for $1/2 > \varepsilon > 0$ and $N\varepsilon \log(\varepsilon^{-1})$ less than a sufficiently small function of $m$ and $d$. Then $|h|_2 \leq O_{m,d}(N^d |p|_2)$.*

PROOF. Consider the probability that $|p(X)| > 2|p|_2$. On the one hand, it is at most $1/4$ by the Markov bound. We will show that if $|h|_2$ is more than a sufficiently large constant times $N^d |p|_2$, then the probability must be more than this.

We note by Corollary 5 that with probability at least $7/8$ that each $q_i(X)$ is $O_d(\log(m)^{d/2})$. We consider the probability that each $q_i(X)$ is at most this size and that $|p(X)| \leq 2|p|_2$. We bound this probability above by coving the set of $x \in \mathbb{R}^m$ with each $|x_i| = O_d(\log(m)^{d/2})$ so that $|h(x)| \leq 2|p|_2$ with boxes of side length $\varepsilon$. The probability is at most $N$ times the volume of the union of these boxes. Furthermore, the union of these boxes is contained in the set of $x \in \mathbb{R}^m$ with $|x_i| \leq O_d(\log(m)^{d/2})$ for each $i$ and so that $x$ is within $\varepsilon m$ of some point $y$ with $|h(y)| \leq 2|p|_2$. Call this region $R$. We note that since $R$ is contained in a ball of radius $O_m(1)$ that the volume of $R$ is bounded by some constant times the probability that a random Gaussian $X$ lies in $R$.

By Proposition 6, we have that with probability $1 - O_{d,m}(\varepsilon \log(\varepsilon^{-1}))$ over Gaussian $X$ that

$$\bigl|h(X)\bigr| > 4\varepsilon m \bigl|\nabla_{i_1} h(X)\bigr|_2 > \cdots > (4\varepsilon m)^d \bigl|\nabla_{i_1} \cdots \nabla_{i_d} h(X)\bigr|_2.$$

This would imply that for any $y$ within $m\varepsilon$ of $X$ that $|h(X) - h(y)| \leq |h(X)|/2$ by means of the Taylor series for $h(y)$. On the other hand, $|h(X)| \geq 4|p|_2$ with probability at least $1 - O_d((|p|_2/|h|_2)^{1/d})$ by Lemma 2. Thus, the probability that $|h(X)| \leq 2|p|_2$ is at most

$$O_{d,m}\bigl(\varepsilon \log\bigl(1 + \varepsilon^{-1}\bigr) + \bigl(|p|_2/|h|_2\bigr)^{1/d}\bigr) + 1/8.$$

Hence, we have that

$$3/4 \leq \Pr\bigl(\bigl|p(X)\bigr| < 2|p|_2\bigr) \leq O_{d,m}\bigl(N\varepsilon \log\bigl(1 + \varepsilon^{-1}\bigr) + N\bigl(|p|_2/|h|_2\bigr)^{1/d}\bigr) + 1/8.$$

Thus, if $N\varepsilon \log(\varepsilon^{-1})$ is less than some sufficiently small function of $d, m$, then $|h|_2 = O_{d,m}(N^d)$. $\square$

Fundamentally, having a diffuse decomposition is useful because it allows us to improve our application of the replacement method. The following proposition presents this technique in fair generality.

PROPOSITION 21. *Let $p_0 : \mathbb{R}^n \to \mathbb{R}$ be a degree-$d$ polynomial with an $(\varepsilon, N)$-diffuse decomposition $(h, q_1, \ldots, q_m)$ for some $1/2 > \varepsilon > 0$. 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 $X^1, \dots, X^\ell$ and $Y^1, \dots, Y^\ell$ be 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 $1 \le j \le \ell$ and any $z^i$ that*

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

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

$$Q_{i,j}(x^1, \dots, x^{j-1}, x^{j+1}, \dots, x^\ell) := \mathbb{E}_{Y^j}[q_i(x^1, \dots, x^{j-1}, Y^j, x^{j+1}, \dots, x^\ell)].$$

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

$$\begin{aligned}&\mathbb{E}[|q_i(Y^1, \dots, Y^j, X^{j+1}, \dots, X^\ell) - Q_{i,j}(Y^1, \dots, Y^{j-1}, X^{j+1}, \dots, X^\ell)|^k] \\ &\quad + \mathbb{E}[|q_i(Y^1, \dots, Y^{j-1}, X^j, \dots, X^\ell) \\ &\quad - Q_{i,j}(Y^1, \dots, Y^{j-1}, X^{j+1}, \dots, X^\ell)|^k].\end{aligned}$$

*And let*

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

*Then we have that*

$$\begin{aligned}&|\Pr(p_0(X^1, \dots, X^\ell) \le 0) - \Pr(p_0(Y^1, \dots, Y^\ell) \le 0)| \\ &\quad \le O_{d,m,k}(\varepsilon^{-k} T + \varepsilon N \log(\varepsilon^{-1})^{dm/2+1}).\end{aligned}$$

*If furthermore, $p$ is a degree-$d$ polynomial so that for some parameters $\delta, \eta > 0$*

$$\Pr(|p(X) - p_0(X)| < \delta |p|_2), \qquad \Pr(|p(Y) - p_0(Y)| < \delta |p|_2) \ge 1 - \eta$$

*then*

$$\begin{aligned}&|\Pr(p(X^1, \dots, X^\ell) \le 0) - \Pr(p(Y^1, \dots, Y^\ell) \le 0)| \\ &\quad \le O_{d,m,k}(\varepsilon^{-k} T + \varepsilon N \log(\varepsilon^{-1})^{dm/2+1} + \delta^{1/d} + \eta).\end{aligned}$$

When considering Proposition 21, it might be useful to keep the intended applications in mind. In Section 5, we will consider the case where the $X^j$ are chosen from $dk$-independent families of Gaussians. In Section 6, we will consider the case where the $X^j$ are Bernoulli random variables. Finally, in Section 8, we will consider the case where the $X^j$ are chosen from $4d$-independent families of random Bernoullis.

PROOF. We begin by proving the first of the two bounds, and will then use it to prove the second. By rescaling $p_0$, we may assume that $|p_0|_2 = 1$. Let $X = (X^1, \ldots, X^\ell)$ and $Y = (Y^1, \ldots, Y^\ell)$. Let $q$ denote the vector valued polynomial $(q_1, \ldots, q_m)$. We will show that

$$\Pr\bigl(p_0(X) \le 0\bigr) \le \Pr\bigl(p_0(Y) \le 0\bigr) + O_{d,m,k}\bigl(\varepsilon^{-k}T + \varepsilon N \log\bigl(\varepsilon^{-1}\bigr)^{dm/2+1}\bigr).$$

The other inequality will follow analogously.

By Proposition 17, there exists a function $f : \mathbb{R}^m \to [0, 1]$ so that:

1. $f(x) = 1$ for all $x$ where $h(x) \le 0$.
2. $\mathbb{E}[f(q(Y))] = \Pr(p_0(Y) \le 0) + O_{d,m}(\varepsilon N \log(\varepsilon^{-1})^{dm/2+1})$.
3. $|f^{(k)}|_\infty = O_{m,k}(\varepsilon^{-k})$.

We note that

$$\Pr\bigl(p_0(X) \le 0\bigr) \le \mathbb{E}\bigl[f\bigl(q(X)\bigr)\bigr]$$

and that

$$\mathbb{E}\bigl[f\bigl(q(Y)\bigr)\bigr] \le \Pr\bigl(p_0(Y) \le 0\bigr) + O_{d,m}\bigl(\varepsilon N \log\bigl(\varepsilon^{-1}\bigr)^{dm/2+1}\bigr).$$

Hence, it suffices to prove that

$$\bigl|\mathbb{E}\bigl[f\bigl(q(X)\bigr)\bigr] - \mathbb{E}\bigl[f\bigl(q(Y)\bigr)\bigr]\bigr| = O_{d,m,k}\bigl(\varepsilon^{-k}T\bigr). \tag{9}$$

For $0 \le j \le \ell$, let

$$Z^{(j)} := \bigl(Y^1, \ldots, Y^j, X^{j+1}, \ldots, X^\ell\bigr).$$

In particular, $Z^{(0)} = X$, $Z^{(\ell)} = Y$, and $Z^{(j)}$ is obtained from $Z^{(j-1)}$ by changing the $j$th block of coordinates from $X^j$ to $Y^j$. We will attempt to bound the left-hand side of equation (9) by bounding

$$\bigl|\mathbb{E}\bigl[f\bigl(q\bigl(Z^{(j)}\bigr)\bigr)\bigr] - \mathbb{E}\bigl[f\bigl(q\bigl(Z^{(j-1)}\bigr)\bigr)\bigr]\bigr| \tag{10}$$

for each $j$.

Consider the expression in equation (10) for fixed values of $Y^1, \ldots, Y^{j-1}$, $X^{j+1}, \ldots, X^\ell$. We approximate $f(q(Z^{(j)}))$ and $f(q(Z^{(j-1)}))$ by Taylor expanding $f$ about $(\overline{q_1}, \ldots, \overline{q_m})$ where

$$\begin{aligned}
&\overline{q_i}\bigl(Y^1, \ldots, Y^{j-1}, Z, X^{j+1}, \ldots, X^\ell\bigr) \\
&\quad := Q_{i,j}\bigl(Y^1, \ldots, Y^{j-1}, X^{j+1}, \ldots, X^\ell\bigr) \\
&\quad = \mathbb{E}\bigl[q_i\bigl(Z^{(j)}\bigr)\bigr] \\
&\quad = \mathbb{E}\bigl[q_i\bigl(Z^{(j-1)}\bigr)\bigr].
\end{aligned}$$

Thus, for some polynomial $g$ of degree $k-1$ we have that

$$\begin{aligned} f\big(q(Z)\big) &= g\big(q(Z)\big) + O\Bigg(\sum_{i_1,\dots,i_k} \frac{\partial^k f}{\partial q_{i_1}\cdots\partial q_{i_k}} \prod_{a=1}^{k} \big(q_{i_a}(Z) - \overline{q_{i_a}}(Z)\big)\Bigg) \\ &= g\big(q(Z)\big) + O_{d,m,k}\Bigg(\sum_{i_1,\dots,i_k} \varepsilon^{-k} \sum_{a=1}^{k} \big|q_{i_a}(Z) - \overline{q_{i_a}}(Z)\big|^k\Bigg) \\ &= g\big(q(Z)\big) + O_{d,m,k}\Bigg(\varepsilon^{-k} \sum_{i=1}^{m} \big|q_i(Z) - \overline{q_i}(Z)\big|^k\Bigg). \end{aligned}$$

By assumption,

$$\mathbb{E}\big[g\big(q\big(Z^{(j)}\big)\big)\big] = \mathbb{E}\big[g\big(q\big(Z^{(j-1)}\big)\big)\big].$$

Thus, the expression in equation (10) is at most

$$\begin{aligned} &\varepsilon^{-k} O_{d,m,k}\Bigg(\sum_{i=1}^{m} \mathbb{E}\big[\big|q_i\big(Z^{(j)}\big) - \overline{q_i}\big(Z^{(j)}\big)\big|^k\big] + \mathbb{E}\big[\big|q_i\big(Z^{(j-1)}\big) - \overline{q_i}\big(Z^{(j-1)}\big)\big|^k\big]\Bigg) \\ &\quad = O_{d,m,k}\Bigg(\varepsilon^{-k} \sum_{i=1}^{m} T_{i,j}\Bigg). \end{aligned}$$

Summing over $j$ yields equation (9), proving the first part of this proposition.

Changing our normalization so that $|p|_2 = 1$, we have that

$$\Pr\big(p(X) \le 0\big) \le \Pr\big(p_0(X) - \delta \le 0\big) + O(\eta).$$

Notice that $p_0 - \delta$ has the diffuse decomposition $(h - \delta, q_1, \dots, q_m)$. Therefore, applying our previous result to this decomposition of $p_0 - \delta$, we have that

$$\Pr\big(p_0(X) - \delta \le 0\big) \le \Pr\big(p_0(Y) - \delta \le 0\big) + O_{d,m,k}\big(\varepsilon^{-k} T + \varepsilon N \log\big(\varepsilon^{-1}\big)^{dm/2+1}\big).$$

On the other hand, we have that

$$\Pr\big(p_0(Y) - \delta \le 0\big) \le \Pr\big(p(Y) - 2\delta \le 0\big) + O(\eta).$$

Finally, by Lemma 2 we have that

$$\Pr\big(p(Y) - 2\delta \le 0\big) \le \Pr\big(p(Y) \le 0\big) + O\big(d\delta^{1/d}\big).$$

Combining the above inequalities, we find that

$$\Pr\big(p(X) \le 0\big) \le \Pr\big(p(Y) \le 0\big) + O_{d,m,k}\big(\varepsilon^{-k} T + \varepsilon N \log\big(\varepsilon^{-1}\big)^{dm/2+1} + \delta^{1/d} + \eta\big).$$

The other direction of the inequality follows analogously, and this completes our proof. $\square$

**5. Application to PRGs for PTFs with Gaussian inputs.** In [13], the author introduced a new pseudorandom generator for polynomial threshold functions of Gaussian inputs. In particular, for appropriately chosen parameters $N$ and $k$ he lets

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

where the $X^i$ are independently chosen from $k$-independent families of Gaussians. He shows that for some $k = O(d/c)$ and $N = 2^{O_c(d)}\varepsilon^{-4-c}$ that for any such $X$, if $Y$ is a random Gaussian and $f$ any degree-$d$ polynomial threshold function then

$$\left|\mathbb{E}\left[f(X)\right] - \mathbb{E}\left[f(Y)\right]\right| < \varepsilon. \tag{11}$$

The proof of this is by the replacement method. In particular, $f$ is replaced by a smooth approximation $g$, and bounds are proved on the change in the expectation of $g(X)$ as the $X^i$ are replaced by random Gaussians one at a time. The power of this method is highly dependent on one's ability to find a $g$ that is close to $f$ with high probability and yet has relatively small higher derivatives. If $f(x) = \operatorname{sgn}(p(x))$, a naive attempt to use the replacement method would use $g = \rho(p(x))$ for $\rho$ a smooth approximation to the sign function. Unfortunately, this approach will have difficulty proving equation (11) unless $N > \varepsilon^{-2d}$. In [13], the author uses a version of Proposition 6 and constructs a $g$ which approximates $f$ as long as an appropriate analogue of

$$\left|g(x)\right| \geq \varepsilon\left|\nabla_{i_1} g(x)\right|_2 \geq \varepsilon^2\left|\nabla_{i_1}\nabla_{i_2} g(x)\right|_2 \geq \cdots$$

holds. The analysis of this is somewhat complicated, involving the development of the theory of the so-called "noisy derivative." Furthermore, for technical reasons this method has difficulty dealing with $N$ smaller than $\varepsilon^{-4}$. As a first application of our theory of diffuse decompositions, we provide a relatively simple analysis of this generator that works with $N$ as small as $\varepsilon^{-2-c}$. In particular, we show the following.

THEOREM 22. *Given, an integer $d > 0$ and real numbers $1 > c, \varepsilon > 0$, there exist integers $k = O(d/c)$ and $N = O_{c,d}(\varepsilon^{-2-c})$ so that for any random variable*

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

*where the $X^i$ are chosen independently from $k$-independent distributions of Gaussians, and for any degree-$d$ polynomial threshold function $f$,*

$$\left|\mathbb{E}\left[f(X)\right] - \mathbb{E}_{Y\sim\mathcal{N}}\left[f(Y)\right]\right| < \varepsilon.$$

PROOF. We begin by making a few reductions to produce a more amenable case. We assume throughout that $\varepsilon$ is sufficiently small. Note that it is sufficient to prove that for $N = \varepsilon^{-2-c}$ that the error is $O_{c,d}(\varepsilon^{1-2c})$, since making appropriate changes to $c$ and $\varepsilon$ will yield the necessary result. Secondly, we may let $f(x) = \operatorname{sgn}(p(x))$ for $p$ a degree-$d$ polynomial with $|p|_2 = 1$.

By Theorem 1, there exists a degree-$d$ polynomial $p_0$ with $|p - p_0|_2 = O_{c,d}(\varepsilon^{d+1})$, and so that $p_0$ has an $(\varepsilon, \varepsilon^{-c/2})$-diffuse decomposition $(h, q_1, \ldots, q_m)$. It should be noted that by $2d$-independence,

$$\mathbb{E}\big[|p(X) - p_0(X)|^2\big] = \mathbb{E}\big[|p(Y) - p_0(Y)|^2\big] = O_{c,d}\big(\varepsilon^{2d+2}\big).$$

Therefore, by the Markov bound we have with probability at least $1 - \varepsilon^2$ that

$$|p(X) - p_0(X)|, |p(Y) - p_0(Y)| < \varepsilon^d.$$

We note that we may write $Y = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} Y^j$, where the $Y^j$ are independent Gaussians. We define the polynomial

$$p'\big(Y^1, \ldots, Y^N\big) := p\left(\frac{1}{\sqrt{N}} \sum_{j=1}^{N} Y^j\right).$$

We note that

$$p(X) = p'\big(X^1, \ldots, X^N\big)$$

and

$$p(Y) = p'\big(Y^1, \ldots, Y^N\big).$$

It is clear that if we define $p_0'$ and $q_i'$ analogously, that $p_0'$ has an $(\varepsilon, \varepsilon^{c/2})$-diffuse decomposition $(h, q_1', \ldots, q_m')$, and that with probability at least $1 - \varepsilon^2$ that

$$|p'(X^i) - p_0'(X^i)|, |p'(Y^i) - p_0'(Y^i)| < \varepsilon^d.$$

We may thus apply Proposition 21 to $p'$, $p_0'$ with $\eta = \varepsilon^2$ and $\delta = \varepsilon^d$.

Let $K$ be an even integer less than $k/d$ and more than $6/c$. By $k$-independence of the $X^j$, any polynomial $g$ of degree less than $K$ in the $q_i'$ will have the same expectation evaluated at $X^1, \ldots, X^N$ as at $Y^1, \ldots, Y^N$. Hence, by Proposition 21,

$$\begin{aligned} &\big|\mathbb{E}\big[f(X)\big] - \mathbb{E}\big[f(Y)\big]\big| \\ &\quad = 2\big|\Pr\big(p\big(X^1, \ldots, X^N\big) \le 0\big) - \Pr\big(p\big(Y^1, \ldots, Y^N\big) \le 0\big)\big| \\ &\quad = O_{d,m,c}\big(\varepsilon^{1-c} \log\big(\varepsilon^{-1}\big)^{dm/2+1} + \varepsilon^{-K} T + \varepsilon\big). \end{aligned} \tag{12}$$

Where by the $dK$-independence of $X$, the $T$ above is

$$2 \sum_{i=1}^{m} \sum_{j=1}^{N} \mathbb{E}\big[\big(q_i(Y) - \mathbb{E}_{Y^j}\big[q_i'\big(Y^1, \ldots, Y^N\big)\big]\big)^K\big].$$

By Lemma 3, this is

$$O_{c,d}\left(\sum_{i=1}^{m}\sum_{j=1}^{N}\mathbb{E}\big[(q_i(Y)-\mathbb{E}_{Yj}[q_i'(Y^1,\ldots,Y^N)])^2\big]^{K/2}\right).$$

Letting $Z=\frac{1}{\sqrt{N-1}}\sum_{i\neq j}Y^i$ (which is a random Gaussian), the expectations in question are

$$\mathbb{E}_Z\left[\mathrm{Var}_Y\left(q_i\left(\sqrt{\frac{N-1}{N}}Z+\frac{1}{\sqrt{N}}Y\right)\right)\right].$$

This in turn is at most

$$\mathbb{E}\left[\left(q_i\left(\sqrt{\frac{N-1}{N}}Z+\frac{1}{\sqrt{N}}Y\right)-q_i(Z)\right)^2\right].$$

We bound this with the following lemma, which follows immediately from Claim 4.1 of [5].

LEMMA 23. *For $q$ any degree-$d$ polynomial, we have that*

$$\mathbb{E}\left[\left|q(Z)-q\left(\sqrt{\frac{N-1}{N}}Z+\frac{1}{\sqrt{N}}Y\right)\right|^2\right]=O\big(d^2|q|_2^2/N\big).$$

Thus, $T$ is at most

$$\begin{aligned}O_{c,d}\left(\sum_{i=1}^{m}\sum_{j=1}^{N}N^{-K/2}\right)&=O_{c,d,m}\big(N^{-K/2+1}\big)\\&=O_{c,d,m}\big(\varepsilon^{K-2+Kc/2-c}\big)\\&=O_{c,d,m}\big(\varepsilon^{K+1-c}\big).\end{aligned}$$

Thus, by equation (12),

$$\big|\mathbb{E}\big[f(X)\big]-\mathbb{E}\big[f(Y)\big]\big|\leq O_{c,d,m}\big(\varepsilon^{1-2c}\big),$$

as desired. □

**6. The diffuse invariance principle and regularity lemma.** While the case of Gaussian inputs is very convenient for proving theorems such as the decomposition theorem, many interesting questions involve evaluation of polynomials on random variables from other distributions. Perhaps the most studied of these is the Bernoulli, or hypercube distribution.

DEFINITION 6. The $n$-dimensional Bernoulli distribution is the probability distribution on $\mathbb{R}^n$ where each coordinate is randomly and uniformly chosen from the set $\{-1,1\}$. Equivalently, it is the uniform distribution on the set $\{-1,1\}^n$.

As we have been using $X, Y, Z$, etc. to represent Gaussian random variables, we will attempt to use $A, B$, etc. for Bernoulli random variables.

A powerful tool for dealing with Bernoulli variables is the use of *invariance principles*. These are theorems which state that if $p$ is a sufficiently regular polynomial (for some definition of regularity) that the distributions of $p(X)$ and $p(B)$ are similar to each other (generally that they are close in c.d.f. distance). This allows one to make use of results in the Gaussian setting and apply them to the Bernoulli setting (at least for sufficiently regular polynomials). Since not all polynomials will be regular, in order to make use of this idea in a more general context, one also needs a *regularity lemma*. These are structural results that allow us to write arbitrary polynomials of Bernoulli random variables in terms of regular ones.

In this section, we will discuss some of the existing invariance principles and regularity lemmas, and make use of the theory of diffuse decompositions to provide some new ones that will deal better with high degree polynomials. In Section 6.1, we discuss some background information about polynomials of Bernoulli random variables and give a brief overview of existing invariance principles and regularity lemmas. In Section 6.2, we state and prove the diffuse invariance principle, and in Section 6.3 prove the corresponding regularity lemma.

### 6.1. *Basic facts about Bernoulli random variables.*

6.1.1. *Multilinear polynomials.* For a Bernoulli random variable $B$, we have that any coordinate, $b_i$, satisfies $b_i^2 = 1$ with probability 1. This, of course, does not hold in the Gaussian case. Thus, if there is going to by any hope of comparing polynomials on Gaussian and Bernoulli inputs, we must restrict ourself to polynomials that have no term that is degree more than 1 in any variable. In particular, we must restrict ourselves to the case of multilinear polynomials.

DEFINITION 7. A polynomial $p : \mathbb{R}^n \to \mathbb{R}$ is multilinear if its degree with respect to any coordinate variable is at most 1.

To clarify the relationship between general polynomials and multilinear polynomials, we mention the following lemma.

LEMMA 24. *For every polynomial $p : \mathbb{R}^n \to \mathbb{R}$, there exists a unique multilinear polynomial $q : \mathbb{R}^n \to \mathbb{R}$ so that $q$ agrees with $p$ on $\{-1, 1\}^n$. Furthermore,* $\deg(q) \leq \deg(p)$.

PROOF. To prove the existence of $q$, it suffices to show that the result holds for every monomial $p = \prod x_i^{\alpha_i}$. It is clear that this monomial agrees on the hypercube with the multilinear monomial $\prod x_i^{\alpha_i \pmod 2}$.

Uniqueness will follow from the fact that any nonzero multilinear polynomial on $\mathbb{R}^n$ is nonvanishing on the hypercube. This follows from the fact that the map

from a multilinear polynomial to its vector of values on $\{-1,1\}^n$ is a surjective linear map of vector spaces of dimension $2^n$. $\square$

DEFINITION 8. For any polynomial $p(x)$, let $L(p(x))$ be the corresponding multilinear polynomial as described by Lemma 24.

6.1.2. *$L^p$ norms and hypercontractivity.* As the $L^p$ norms for polynomials of Gaussians have been useful to us, the corresponding norms for the Bernoulli distribution will also be useful.

DEFINITION 9. Let $p:\mathbb{R}^n \to \mathbb{R}$ we for $t \geq 1$, we define $|p|_{B,t}$ as

$$|p|_{B,t} = \left(\mathbb{E}_B\left[|p(B)|^t\right]\right)^{1/t},$$

where above $B$ is an $n$-dimensional Bernoulli random variable.

We also have the analogue of Lemma 3. In particular, we have the following.

LEMMA 25 (Bonami [2]). *For $p:\mathbb{R}^n \to \mathbb{R}$ a degree-$d$ polynomial, and $t \geq 2$ we have that*

$$|p|_{B,t} \leq \sqrt{t-1}^d |p|_{B,2}.$$

From this we derive the following corollary.

COROLLARY 26. *For $p:\mathbb{R}^n \to \mathbb{R}$ a degree-$d$ polynomial $N > 0$, then*

$$\Pr_B\left(|p(B)| > N|p|_{B,2}\right) = O\left(2^{-(N/2)^{2/d}}\right).$$

The proof is analogous to that of Corollary 5.
We will also need a result combining Lemmas 3 and 25.

LEMMA 27. *Let $p$ be a degree-$d$ polynomial, $B$ a Bernoulli random variable, $G$ a Gaussian random variable, and $t \geq 2$ a real number. Then*

$$\mathbb{E}\left[|p(G,B)|^t\right] \leq (t-1)^{td/2}\mathbb{E}\left[p(G,B)^2\right]^{t/2}.$$

PROOF. For integers $N$, let $G^N$ be a random variable defined by $G^N = \frac{1}{\sqrt{N}}\sum_{j=1}^N A^j$ where the $A^j$ are independent Bernoulli random variables. Clearly, the coordinates of $G^N$ are independent and by the central limit theorem, as $N \to \infty$, their distributions converge to Gaussians in c.d.f. distance. This implies that for and $\varepsilon > 0$ and for sufficiently large $N$ that we can have correlated copies of the random variables $G$ and $G^N$ so that $|G - G^N| < \varepsilon$ with probability $1-\varepsilon$. Furthermore, with probability $1-\varepsilon$, $|G| = O_n(\log(\varepsilon^{-1}))$ (here $n$ in the number of

coordinates of $G$). Therefore, for sufficiently large $N$ we have that with probability $1 - O(\varepsilon)$ that $|p(G, B) - p(G^N, B)| = O_p(\varepsilon \log(\varepsilon^{-1})^d)$ (this follows from considering every possible value of $B$ separately). Therefore, the $p(G^N, B)$ converge in law to $p(G, B)$. Furthermore, since $\mathbb{E}[|p(G^N, B)|^{2\lfloor t \rfloor}]$ is uniformly bounded [expand out $p(G^N, B)^{2\lfloor t \rfloor}$ and note that each monomial has uniformly bounded expectation], this implies that

$$\lim_{N\to\infty} \mathbb{E}\big[|p(G^N, B)|^t\big] = \mathbb{E}\big[|p(G, B)|^t\big],$$

and

$$\lim_{N\to\infty} \mathbb{E}\big[|p(G^N, B)|^2\big] = \mathbb{E}\big[|p(G, B)|^2\big].$$

The result now follows from applying Lemma 25 to $p(G^N, B)$ and taking a limit as $N \to \infty$. □

We also note the following relationship between the Gaussian and Bernoulli norms.

LEMMA 28. *If $p : \mathbb{R}^n \to \mathbb{R}$ is a multilinear polynomial then $|p|_2 = |p|_{B,2}$.*

PROOF. This follows immediately after noting that the basis $\prod x_i^{\alpha_i}$ for $\alpha \in \{0, 1\}^n$ is an orthonormal basis of the set of multilinear polynomials with respect to both the Bernoulli and Gaussian measures. □

6.1.3. *Influence and regularity.* The primary obstruction to a multilinear polynomial behaving similarly when evaluated at Bernoulli inputs rather than Gaussian inputs is when some single coordinate has undo effect on the output value of the polynomial. In such a case, the fact that this coordinate is distributed as a Bernoulli rather than a Gaussian may cause significant change to the resulting distribution. In order to quantify the extent to which this can happen, we define the $i$th influence of a coordinate as follows.

DEFINITION 10. For $p : \mathbb{R}^n \to \mathbb{R}$, a function we define the $i$th influence of $p$ to be

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

It should be noted that for multilinear polynomials $p$, this is equivalent to the more standard definition

$$\mathrm{Inf}_i(p) = \mathbb{E}_A\big[\mathrm{Var}_{a_i}\big(p(A)\big)\big].$$

This is the expectation over uniform independent $\{-1, 1\}$ choices for the coordinates other than the $i$th coordinate of the variance of the resulting function over a Bernoulli choice of the $i$th coordinate. Equivalently, it is

$$\frac{1}{4}\mathbb{E}\big[\big|p(a_1,\dots,a_{i-1},-1,a_{i+1},\dots,a_n)-p(a_1,\dots,a_{i-1},1,a_{i+1},\dots,a_n)\big|^2\big].$$

We now prove some basic facts about the influence.

LEMMA 29. *If $p:\mathbb{R}^n\to\mathbb{R}$ is a polynomial* $\mathrm{Inf}_i(p)$ *is* $\sum_a a_i|c_a(p)|^2$, *where* $c_a(p)$ *are the Hermite coefficients of* $p$.

PROOF. Recall that

$$p(x)=\sum_a c_a(p)h_a(x).$$

Therefore, we have that

$$\frac{\partial p}{\partial x_i}=\sum_a \sqrt{a_i}c_a(p)h_{a-e_i}(x).$$

Thus,

$$\left|\frac{\partial p}{\partial x_i}\right|_2^2=\sum_a a_i|c_a(p)|^2.$$

□

From this, we have the following.

COROLLARY 30. *For $p$ a degree-$d$ polynomial in $n$ variables,*

$$\sum_{i=1}^n \mathrm{Inf}_i(p)=\sum_{k=1}^d k\big|p^{[k]}\big|_2^2=\Theta_d\big(\mathrm{Var}\big(p(X)\big)\big).$$

We now make the following definition [which agrees with the standard ones up to changing $\tau$ by a factor of $\Theta_d(1)$].

DEFINITION 11. Let $p$ be a degree-$d$ multilinear polynomial. We say that $p$ is $\tau$-regular if for each $i$

$$\mathrm{Inf}_i(p)\le\tau\,\mathrm{Var}_A(p).$$

In terms of this notion of regularity, the standard invariance principle, proved in [18], can be stated as follows.

THEOREM 31 (The invariance principle (Mossel, O'Donnell, and Oleszkiewicz)). *If $p$ is a $\tau$-regular, degree-$d$ multilinear polynomial, $A$ and $X$ are Bernoulli and Gaussian random variables respectively and $t \in \mathbb{R}$, then*

$$|\Pr(p(X) \le t) - \Pr(p(A) \le t)| = O(d\tau^{1/(8d)}).$$

It should be noted that the dependence on $\tau^{1/d}$ in the error of Theorem 31 is necessary. In particular, if $d$ is even and $N$ is a sufficiently large integer consider the polynomial $p : \mathbb{R}^{n+1} \to \mathbb{R}$ defined by

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

Let $q = L(p)$. It is not hard to see that by making $N$ sufficiently large, one can make $|p - q|_2$ arbitrarily small, and thus, by Lemma 2 and Corollary 5, we can make the probability distributions for $p(X)$ and $q(X)$ arbitrarily close. It is also not hard to see that $q$ is $\Theta_d(\tau^2)$ regular. This is because $\mathrm{Inf}_0(q) = \tau^2$, $\mathrm{Inf}_i(q) = O_d(N^{-1})$ for $i \neq 0$, and $\mathrm{Var}_A(q(A)) = \Theta_d(1)$. On the other hand, it is clear that for Bernoulli input $A$ we have that

$$q(A) = p(A) \ge -\tau.$$

On the other hand, considering the distribution of values of $p(X)$ [which as stated can be arbitrarily close to that of $q(X)$], if we let $y = \frac{1}{\sqrt{N}} \sum_{i=2}^{N+1} x_i$, we note that $x_0$ and $y$ are independent Gaussians. Thus, with probability $\Theta(\tau^{1/d})$ we have that $x_0 < -2$ and $|y| \le \tau^{1/d}$. If these occur, then $p(X) < -\tau$. Thus, for $N$ sufficiently large the difference between the probabilities that $q(A) < -\tau$ and that $q(X) < -\tau$ can be as large as $\Omega(\tau^{1/d})$.

The essential problem in the above example is that although the first coordinate has low influence, there is a reasonable probability that the size of $q(X)$ will be comparable to $\tau$, and in the case when $|q(X)|$ is small, the relative effect of the first coordinate is much larger. We get around this problem by introducing a new concept of regularity involving the idea of a diffuse decomposition. The problem above came from the fact that the probability distribution of $q(X)$ was too clustered near 0. Since the analogue of this cannot happen for a diffuse set of polynomials, we expect to obtain better bounds.

DEFINITION 12. For $p$ a degree-$d$ multilinear polynomial, we say that $p$ has a $(\tau, N, m, \varepsilon)$-*regular decomposition* if there exists a polynomial $p_0$ of degree-$d$ so that:

- $|p - p_0|_{B,2}^2 \le \varepsilon^2 \mathrm{Var}(p_0(X))$.
- $p_0$ has a $(\tau^{1/5}, N)$-diffuse decomposition of size $m$, $(h, q_1, \dots, q_m)$ so that $q_i$ is multilinear for each $i$ and $\mathrm{Inf}_j(q_i) \le \tau$ for each $i, j$.

THEOREM 32 (The diffuse invariance principle). *If $p$ is a degree-$d$ multilinear polynomial that has a $(\tau, N, m, \varepsilon)$-regular decomposition for $1/2 > \varepsilon, \tau > 0$, $A$ and $X$ and random Bernoulli and Gaussian variables, respectively, and $t$ is a real number, then*

$$\begin{aligned}&\bigl|\Pr\bigl(p(A) \le t\bigr) - \Pr\bigl(p(X) \le t\bigr)\bigr| \\ &\quad = O_{d,m}\bigl(\tau^{1/5} N \log\bigl(\tau^{-1}\bigr)^{dm/2+1} + \varepsilon^{1/d} \log\bigl(\varepsilon^{-1}\bigr)^{1/2}\bigr).\end{aligned}$$

REMARK 3. We can derive a statement very similar to that of Theorem 31 from Theorem 32. In particular, if $p$ is multilinear, and $\tau$-regular, we may normalize $p$ so that $\mathbb{E}_X[p(X)] = 0, \mathbb{E}_X[p(X)^2] = 1$. Then by Lemma 2, we have that for $h = \mathrm{Id}$ and $q = p$, $(h, q)$ is a $(\tau^{1/5}, O(d\tau^{(1/d-1)/5}))$-diffuse decomposition of $p$. Furthermore, by assumption $q$ is multilinear and has all influences at most $\tau$. Therefore, this is a $(\tau, O(d\tau^{(1/d-1)/5}), 1, 0)$-regular decomposition of $p$. Thus, we obtain

$$\bigl|\Pr\bigl(p(A) \le t\bigr) - \Pr\bigl(p(X) \le t\bigr)\bigr| = O_d\bigl(\tau^{1/(5d)} \log\bigl(\tau^{-1}\bigr)^{d/2+1}\bigr).$$

Neither invariance principle on its own is very useful for dealing with general polynomial threshold functions which might not satisfy the necessary regularity conditions. Fortunately, in both cases if regularity fails it will be because some small number of coordinates have undo effect on the value of the polynomial. If this is the case, we can hope to make things better by fixing the values of these coordinates and considering the resulting polynomial over the remaining coordinates, hoping that it is regular. Theorems confirming this intuition have been known as *regularity lemmas*.

Ideally, one would like a regularity lemma to say that for some small set $S$ of coordinates, if one takes a random restriction over the coordinates of $S$ that with high probability the resulting polynomial is either regular or nearly constant. Unfortunately, existing techniques are insufficient to prove such a result where the coordinates of $S$ are picked ahead of time. Instead most results instead use the idea of a low depth decision tree. In particular, when we say that we write $f$ as a decision tree of depth $D$ with nodes given by functions $f_\rho$, we are specifying $f$ by considering its restrictions on sets of at most $D$ coordinates at a time, but rather than declaring the coordinates to be fixed ahead of time, we allow them to be chosen adaptively. In particular, if we are restricting on coordinates $x_{i_1}, x_{i_2}, \ldots, x_{i_D}$ we allow $i_j$ to depend on the $(\pm 1)$ values assigned to $x_{i_1}, \ldots, x_{i_{j-1}}$.

Making use of these ideas, several regularity lemmas have appear for the standard notion of regularity, for example, in [6] and [4] as well as other places. As an example, [6] proved the following.

THEOREM 33 (Diakonikolas, Servedio, Tan, Wan). *Let $f(x) = \mathrm{sign}(p(x))$ be any degree-$d$ PTF. Fix any $\tau > 0$. Then $f$ is equivalent to a decision tree $T$, of*

*depth*

$$\operatorname{depth}(d,\tau)=\frac{1}{\tau}\cdot\left(d\log\left(\tau^{-1}\right)\right)^{O(d)}$$

*with variables at the internal nodes and a degree-d PTF* $f_\rho=\operatorname{sgn}(p_\rho)$ *at each leaf* $\rho$, *with the following property*: *with probability at least* $1-\tau$, *a random path from the root reaches a leaf* $\rho$ *such that* $f_\rho$ *is* $\tau$*-close to some* $\tau$*-regular degree-d PTF.*

Along similar lines, we prove the following.

THEOREM 34 (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*

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

*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* $\rho$ *so that the corresponding polynomial* $p_\rho$ *either satisfies* $\operatorname{Var}(p_\rho)<\tau^M|p_\rho|_2^2$ *or* $p_\rho$ *has an* $(\tau,\tau^{-c},O_{c,d,M}(1),O_{c,d,M}(\tau^M))$*-regular decomposition.*

6.2. *The diffuse invariance principle.* In this section, we prove Theorem 32. We begin with the following proposition.

PROPOSITION 35. *Let* $p$ *be a degree-d polynomial with a* $(\tau^{1/5},N)$*-diffuse decomposition (for* $1/2>\tau>0$*)* $(h,q_1,\dots,q_m)$ *with* $q_i$ *multilinear so that* $\operatorname{Inf}_i(q_j)\le\tau$ *for all* $i,j$. *Then if* $A$ *is a Bernoulli random variable,* $X$ *a Gaussian random variable and* $t$ *a real number then*

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

PROOF. It suffices to prove this statement for $t=0$. We proceed via Proposition 21. We note that for each $i$ the first three moments of $A_i$ agree with the corresponding moments of $X_i$. Therefore, since the $q_i$ are multilinear, any degree-3 polynomial in the $q_i$ has the same expectation under $A$ as under $X$. Thus, we may apply Proposition 21 with $k=4$. We have that

$$\left|\Pr\left(p(A)\le 0\right)-\Pr\left(p(X)\le 0\right)\right|=O_{d,m}\left(N\tau^{1/5}\log\left(\tau^{-1}\right)^{dm/2+1}+\tau^{-4/5}T\right). \tag{13}$$

Recall that $T_{i,j}$ is

$$\begin{aligned}\mathbb{E}\big[\big(q_i(X_1,\dots,X_{j-1},A_j,\dots,A_n)-\mathbb{E}_Y\left[q_i(X_1,\dots,X_{j-1},Y,A_{j+1},\dots,A_n)\right]\big)^4\big]\\ +\mathbb{E}\big[\big(q_i(X_1,\dots,X_j,A_{j+1},\dots,A_n)\\ -\mathbb{E}_Y\left[q_i(X_1,\dots,X_{j-1},Y,A_{j+1},\dots,A_n)\right]\big)^4\big].\end{aligned}$$

By Lemma 27, this is at most

$$O_d\big(\mathbb{E}_{X_1,\ldots,X_{j-1},A_{j+1},\ldots,A_n}\big[\mathrm{Var}_Y\big(q_i(X_1,\ldots,X_{j-1},Y,A_{j+1},\ldots,A_n)\big)\big]^2\big).$$

Since the polynomial in expectation is at most quadratic in each $X_i$, this is

$$O_d\big(\mathbb{E}_A\big[\mathrm{Var}_Y\big(q_i(A_1,\ldots,A_{j-1},Y,A_{j+1},\ldots,A_n)\big)\big]^2\big) = O_d\big(\mathrm{Inf}_j(q_i)^2\big).$$

Thus,

$$\begin{aligned} T &= \sum_{i=1}^{m}\sum_{j=1}^{n} T_{i,j} \\ &= O_d\left(\sum_{i=1}^{m}\sum_{j=1}^{n} \mathrm{Inf}_j(q_i)^2\right) \\ &\le O_d\left(\sum_{i=1}^{m}\sum_{j=1}^{n} \tau\,\mathrm{Inf}_j(q_i)\right) \\ &= O_d\left(\sum_{i=1}^{m}\tau\right) \\ &= O_{d,m}(\tau). \end{aligned}$$

Thus, by equation (13),

$$\big|\Pr\big(p(A)\le 0\big) - \Pr\big(p(X)\le 0\big)\big| = O_{d,m}\big(N\tau^{1/5}\log\big(\tau^{-1}\big)^{dm/2+1}\big),$$

as desired. □

Proposition 35 is the main analytic tool used in our proof of Theorem 32. From it, we can quickly derive the following theorem.

THEOREM 36. *Let $p$ be a degree-$d$ multilinear polynomial with a $(\tau, N, m, \varepsilon)$-regular decomposition (for $1/2 > \varepsilon, \tau > 0$) given by $(h, q_1, \ldots, q_m)$. Let $p_0(x) := h(q_1(x), \ldots, q_m(x))$. Let $A$ be a Bernoulli random variable, $X$ a Gaussian random variable, and $t$ a real number. Then*

$$\begin{aligned} &\big|\Pr\big(p(A)\le t\big) - \Pr\big(p_0(X)\le t\big)\big| \\ &\quad = O_{d,m}\big(N\tau^{1/5}\log\big(\tau^{-1}\big)^{dm/2+1} + \varepsilon^{1/d}\log\big(\varepsilon^{-1}\big)^{1/2}\big). \end{aligned}$$

REMARK 4. For most applications, Theorem 36 will be as good as Theorem 32 as it shows that the regular polynomial of Bernoullis behaves similarly to a polynomial of Gaussians. As we shall see later, it will take some work to show that it will necessarily behave like the same polynomial of Gaussians. This is because although $|p - p_0|_{2,B}$ is small, this does not immediately imply that $|p - p_0|_2$ is sufficiently small for the proof to work.

PROOF. As in the proof of Proposition 35, we may assume that $t = 0$ and prove the inequality

$$\begin{aligned}&\Pr\bigl(p(A) \le 0\bigr)\\ &\quad\le \Pr\bigl(p_0(X) \le 0\bigr) + O_{d,m}\bigl(N\tau^{1/5}\log\bigl(\tau^{-1}\bigr)^{dm/2+1} + \varepsilon^{1/d}\log\bigl(\varepsilon^{-1}\bigr)^{1/2}\bigr).\end{aligned}$$

By Corollary 26, we have with probability $1-\varepsilon$ that

$$\bigl|p(A) - p_0(A)\bigr| \le O\bigl(\varepsilon\log\bigl(\varepsilon^{-1}\bigr)^{d/2}\bigr)\sqrt{\operatorname{Var}\bigl(p_0(X)\bigr)} \le O\bigl(\varepsilon\log\bigl(\varepsilon^{-1}\bigr)^{d/2}\bigr)|p_0|_2.$$

Thus, we have that

$$\begin{aligned}&\Pr\bigl(p(A) \le 0\bigr)\\ &\quad\le \varepsilon + \Pr\bigl(p_0(A) \le O\bigl(\varepsilon\log\bigl(\varepsilon^{-1}\bigr)^{d/2}\bigr)|p_0|_2\bigr)\\ &\quad\le O_{d,m}\bigl(N\tau^{1/5}\log\bigl(\tau^{-1}\bigr)^{dm/2+1} + \varepsilon\bigr) + \Pr\bigl(p_0(X) \le O\bigl(\varepsilon\log\bigl(\varepsilon^{-1}\bigr)^{d/2}\bigr)|p_0|_2\bigr)\\ &\quad\le O_{d,m}\bigl(N\tau^{1/5}\log\bigl(\tau^{-1}\bigr)^{dm/2+1} + \varepsilon^{1/d}\log\bigl(\varepsilon^{-1}\bigr)^{1/2}\bigr) + \Pr\bigl(p_0(X) \le 0\bigr).\end{aligned}$$

The second line above is by Proposition 35 and the third is by Lemma 2.

The lower bound on $\Pr(p(A) \le 0)$ is proved analogously. □

In order to complete the proof of Theorem 32, we need the following.

PROPOSITION 37. *If $p$ is a degree-$d$ polynomial with a $(\tau, N, m, \varepsilon)$-regular decomposition (for $1/2 > \varepsilon, \tau > 0$) given by $p_0(x) = h(q_1(x), \dots, q_m(x))$, then for $X$ a Gaussian random variable, and $t$ a real number,*

$$\begin{aligned}&\bigl|\Pr\bigl(p(X) \le t\bigr) - \Pr\bigl(p_0(X) \le t\bigr)\bigr|\\ &\quad\le O_{d,m}\bigl(\tau^{1/4}N\log\bigl(\tau^{-1}\bigr)^{d(m+1)/2+1} + \varepsilon^{1/d}\log\bigl(\varepsilon^{-1}\bigr)^{1/2}\bigr).\end{aligned}$$

The biggest difficulty with proving this proposition will be dealing with the discrepancy between $p_0$ and $L(p_0)$. To deal with this, we make the following definition.

DEFINITION 13. Let $p_1, \dots, p_k$ be multilinear polynomials. Define

$$A(p_1, \dots, p_k) = \sum_{S\subseteq\{1,2,\dots,k\}} (-1)^{|S|}\Biggl(\prod_{i\in S} p_i\Biggr)L\Biggl(\prod_{i\notin S} p_i\Biggr).$$

We note the following.

LEMMA 38. *Let $q_1, \dots, q_m$ be multilinear polynomials and let $h$ be a degree-$d$ polynomial in $m$ variables then $L(h(q_1(x), \dots, q_m(x)))$ is*

$$\sum_{k=0}^{d}\sum_{i_1,\dots,i_k=1}^{m} \frac{\partial^k h}{\partial q_{i_1}\cdots\partial q_{i_k}} A(q_{i_1}, \dots, q_{i_k}).$$

PROOF. As the above expression is linear in $h$, we may assume that $h$ is a monomial of degree $d$. In particular, we may assume that $h = q_{i_1} q_{i_2} \cdots q_{i_d}$ (note that some of the indices $i_j$ might coincide). The expression in question then becomes:

$$\begin{aligned}
&\sum_{T=\{t_1,\ldots,t_k\}\subseteq\{1,\ldots,d\}} \Bigl(\prod_{j\notin T} q_{i_j}\Bigr) A(q_{i_{t_1}},\ldots,q_{i_{t_k}}) \\
&\quad = \sum_{T\subseteq\{1,\ldots,d\}} \Bigl(\prod_{j\notin T} q_{i_j}\Bigr) \sum_{S\subseteq T} (-1)^{|S|} \Bigl(\prod_{i\in S} q_i\Bigr) L\Bigl(\prod_{i\in T\setminus S} q_i\Bigr) \\
&\quad = \sum_{S\subseteq T\subseteq\{1,\ldots,d\}} \Bigl(\prod_{j\notin T\setminus S} q_{i_j}\Bigr)(-1)^{|S|} L\Bigl(\prod_{j\in T\setminus S} q_{i_j}\Bigr).
\end{aligned}$$

Letting $R = T\setminus S$, this is

$$\begin{aligned}
&\sum_{R\subseteq\{1,\ldots,d\}} L\Bigl(\prod_{j\in R} q_{i_j}\Bigr)\Bigl(\prod_{j\notin R} q_{i_j}\Bigr) \sum_{S\in\{1,\ldots,d\}\setminus R} (-1)^{|S|} \\
&\quad = \sum_{R=\{1,\ldots,d\}} L\Bigl(\prod_{j\in R} q_{i_j}\Bigr)\Bigl(\prod_{j\notin R} q_{i_j}\Bigr) \\
&\quad = L\Bigl(\prod_{j\in\{1,\ldots,d\}} q_{i_j}\Bigr) \\
&\quad = L(h),
\end{aligned}$$

as desired. □

To control the discrepancy between $p_0$ and $L(p_0)$ it now suffices to prove the following.

PROPOSITION 39. *Let $p_1,\ldots,p_k$ be multilinear, degree at most $d$ polynomials with $\mathrm{Inf}_i(p_j) \le \tau$ for all $i$, $j$, $|p_j|_2 \le 1$ for all $j$. Then*

$$\bigl|A(p_1,\ldots,p_k)\bigr|_2 = O_{k,d}\bigl(\tau^{k/4}\bigr).$$

PROOF. We proceed by bounding the expected value of $A(p_1,\ldots,p_k)^2$. In particular, we show that if $p_j$ are multilinear degree-$d$ polynomials of norm at most 1 with all influences at most $\tau$ then

$$\mathbb{E}\bigl[A(p_1,\ldots,p_k)(X)A(p_{k+1},\ldots,p_{2k})(X)\bigr] = O_{k,d}\bigl(\tau^{k/2}\bigr).$$

We note that the above expression is linear in the $p_j$. We may therefore rewrite it as a sum over sequences of monomials $m_1,\ldots,m_{2k}$ where $m_j$ is a monomial of $p_j$, of

$$\mathbb{E}\bigl[A(m_1,\ldots,m_k)(X)A(m_{k+1},\ldots,m_{2k})(X)\bigr].$$

To each such sequence of monomials $m_1, \dots, m_{2k}$ we associate a *repeat pattern*, which is the multiset of nonempty subsets of $\{1, 2, \dots, 2k\}$ whose elements correspond to $\{j : x_i \text{ appears in monomial } m_j\}$ for all $i$ so that $x_i$ appears in any of the monomials $m_j$. We break up the above sum into parts based on the repeat pattern satisfied by $m_1, \dots, m_{2k}$, since there are $O_{k,d}(1)$ such possible patterns, it suffices to prove our bound for the sum of all terms coming from each such pattern. In particular, we need to show that for any repeat pattern $P$ that

$$\sum_{\substack{m_j \text{ a monomial from } p_j \\ (m_1,\dots,m_{2k}) \text{ has repeat pattern } P}} \mathbb{E}\big[A(m_1, \dots, m_k)(X)A(m_{k+1}, \dots, m_{2k})(X)\big] \tag{14}$$

$$= O_{k,d}\big(\tau^{k/2}\big). \tag{15}$$

Note that if the repeat pattern contains any subset of odd size that the resulting sum will be 0. This is because for any $m_1, \dots, m_{2k}$ with this repeat pattern, there will be some $x_i$ appearing in an odd number of the $m_j$. This means that the product of the $m_j$ will be an odd function of $x_i$. Since $L$ of an odd polynomial is still odd, this means that $A(m_1, \dots, m_k)A(m_{k+1}, \dots, m_{2k})$ will be an odd function of $x_i$ and thus, have expectation 0.

Furthermore, suppose that given $P$, there is some $1 \le j \le 2k$ so that $j$ does not appear in any element of $P$ of size greater than 2. We claim again that for any $m_1, \dots, m_{2k}$ satisfying $P$ that

$$\mathbb{E}\big[A(m_1, \dots, m_k)(X)A(m_{k+1}, \dots, m_{2k})(X)\big] = 0.$$

To show this, we assume without loss of generality that $j = 1$. We expand out the $A$'s to get that the expression in question is the expectation of

$$\sum_{S\subseteq\{1,2,\dots,k\}} \sum_{T\subseteq\{k+1,\dots,2k\}} (-1)^{|S|+|T|}\Big(\prod_{j\in S\cup T} p_j\Big) \times L\Big(\prod_{j\in\{1,\dots,k\}\setminus S} p_j\Big)L\Big(\prod_{j\in\{k+1,\dots,2k\}\setminus T} p_j\Big).$$

We claim that if we toggle whether 1 is in $S$ in the above sum, it has no effect on the expectation of the resulting product other than to negate the $(-1)^{|S|+|T|}$ term. This is because adding 1 to $S$ can only have the effect of removing some $x_i^2$ terms from the resulting monomial. On the other hand, since $\mathbb{E}[1] = \mathbb{E}[X_i^2]$, this does not effect the resulting expectation. Thus, the expectations of the terms with 1 in $S$ cancel the expectations of the terms with 1 not in $S$, leaving us with expectation 0.

It thus suffices to consider equation (14) when all elements of $P$ have even order and so that for each $1 \le j \le 2k$ there is some element of $P$ of order at least 4 containing $j$. For such $P$, we upper bound the left-hand side of equation (14) by

$$\sum_{\substack{m_j \text{ a monomial from } p_j \\ (m_1,\dots,m_{2k}) \text{ has repeat pattern } P}} O_{k,d}\Bigg(\prod_{j=1}^{2k} |m_j|_2\Bigg). \tag{16}$$

We will now prove the following statement, which will imply our desired bound. Let $p_1, \dots, p_{2k}$ be multilinear polynomials with $|p_j| \le 1$ and $T \subseteq \{1, 2, \dots, 2k\}$ some set so that $\mathrm{Inf}_i\, p_j \le \tau$ for all $i$ and all $j \in T$. Furthermore, let $P$ be a repeat pattern all of whose elements have even order and so that each element of $T$ appears in some element of $P$ of order at least 4, then the expression in equation (16) is at most $O_{k,d}(\tau^{|T|/4})$. We prove this by induction on $|P|$. The base case where $|P| = 0$ is trivial since then we are considering only the term where all of the $m_j$ are constants.

If $|P| > 0$, we consider an element of $P$ of maximal size. In particular, if $T \neq \varnothing$, this implies that this element is of size at least 4. Without loss of generality, this element is $\{1, 2, \dots, 2\ell\}$. We break our sum into pieces based on which coordinate is shared by all of $m_1, \dots, m_{2\ell}$ (if more than one coordinate is shared by each of these elements we will count all of them leading to a strictly larger sum). If we wish to compute the sum over all terms where they share a coordinate $x_i$, we find that it is

$$\sum_{\substack{m_j \text{ a monomial from } p'_j \\ (m_1,\dots,m_{2k}) \text{ has repeat pattern } P'}} O_{k,d}\left(\prod_{j=1}^{2} k|m_j|_2\right).$$

Where above $p'_j = p_j$ for $j > 2\ell$ and for $j \le 2\ell$, $p'_j$ consists of the sum of the monomials in $p_j$ containing $x_i$ divided by $x_i$, and $P'$ is $P$ minus $\{1, 2, \dots, 2\ell\}$. Furthermore, note that $|p'_j|_2 = \sqrt{\mathrm{Inf}_i(p_j)}$ for $j \le 2\ell$. Letting $p''_j$ be the normalized version of $p'_j$, the above is at most

$$\prod_{j=1}^{2\ell} \sqrt{\mathrm{Inf}_i(p_j)} \sum_{\substack{m_j \text{ a monomial from } p''_j \\ (m_1,\dots,m_{2k}) \text{ has repeat pattern } P'}} O_{k,d}\left(\prod_{j=1}^{2} k|m_j|_2\right).$$

Letting $T' = T \setminus \{1, 2, \dots, 2\ell\}$, we note that this sum is of the form specified for the value $T'$, hence we have by the inductive hypothesis that the above sum is

$$O_{k,d}\left(\tau^{|T'|/4} \prod_{j=1}^{2\ell} \sqrt{\mathrm{Inf}_i(p_j)}\right).$$

It thus suffices to prove that

$$\sum_i \prod_{j=1}^{2\ell} \sqrt{\mathrm{Inf}_i(p_j)} = O_{k,d}\left(\tau^{(|T|-|T'|)/4}\right) = O_{k,d}\left(\tau^{(|T\cap\{1,2,\dots,2\ell\}|)/4}\right).$$

We assume without loss of generality that $T \cap \{1, 2, \dots, 2\ell\} = \{1, 2, \dots, a\}$. We note by Cauchy–Schwarz that

$$\sum_i \prod_{j=1}^{2\ell} \sqrt{\mathrm{Inf}_i(p_j)} \le \left(\prod_{j=1}^{2\ell-2} \max_i \sqrt{\mathrm{Inf}_i(p_j)}\right)\left(\prod_{j=2\ell-1}^{2\ell} \sum_i \mathrm{Inf}_i(p_j)\right)^{1/2}.$$

We note that for each of the last two terms that

$$\sum_i \mathrm{Inf}_i(p_j) = O_d\big(|p_j|_2^2\big) = O_d(1).$$

Furthermore, we have that

$$\prod_{j=1}^{2\ell-2} \max_i \sqrt{\mathrm{Inf}_i(p_j)} \leq \prod_{j=1}^{\min(a,2\ell-2)} \tau^{1/2} \prod_{j=a+1}^{2\ell-2} 1 = \tau^{\min(a,2\ell-2)/2}.$$

Thus, we have that

$$\sum_i \prod_{j=1}^{2\ell} \sqrt{\mathrm{Inf}_i(p_j)} \leq O_d\big(\tau^{\min(a,2\ell-2)/2}\big) = O_d\big(\tau^{a/4}\big).$$

With the last step following from the observation that either $a = 0$ or $\ell \geq 2$. This completes our inductive step and proves our proposition. □

We are now prepared to prove Proposition 37, and thus, Theorem 32.

PROOF. We may clearly assume that $t = 0$. We will give a series of high probability statements that together imply that

$$\mathrm{sgn}\big(p(X)\big) = \mathrm{sgn}\big(p_0(X)\big).$$

Let $V = \mathrm{Var}(p_0)$.

First, note that by assumption

$$\big|p - L(p_0)\big|_2^2 = |p - p_0|_{2,B}^2 \leq \varepsilon V.$$

Thus, by Corollary 5 we have for some sufficiently large $C$ that with probability $1 - \varepsilon$ that

$$\big|p(X) - L(p_0)(X)\big| \leq C\varepsilon \log\big(\varepsilon^{-1}\big)^{d/2} V.$$

Additionally, by Lemma 2, we have with probability $1 - O(d\varepsilon^{1/d}\log(\varepsilon^{-1})^{1/2})$ that

$$\big|p_0(X)\big| \geq 2C\varepsilon \log\big(\varepsilon^{-1}\big)^{d/2} |p_0|_2 \geq 2C\varepsilon \log\big(\varepsilon^{-1}\big)^{d/2} \sqrt{V}.$$

By Proposition 39 and Corollary 5, we have that for $C$ a sufficiently large number given $d$ that with probability $1 - O_{d,m}(\tau)$ that for all $1 \leq i_1, i_2, \ldots, i_k \leq m$ for $k \leq d$ that

$$\big|A(q_{i_1}, \ldots, q_{i_k})(X)\big| \leq C\tau^{k/4} \log\big(\tau^{-1}\big)^{dk/2}.$$

Finally, by Lemma 18 we have that with probability

$$1 - O_{d,m}\big(\tau^{1/4} N \log\big(\tau^{-1}\big)^{d(m+1)/2+1}\big)$$

that letting $Q = (q_1(X), \ldots, q_m(X))$ that

$$\begin{aligned}|h(Q)| &\geq 3Cm\tau^{1/4}\log(\tau^{-1})^{d/2}|\nabla_{i_1}h(Q)|_2\\ &\geq 3^2C^2m^2\tau^{2/4}\log(\tau^{-1})^{d2/2}|\nabla_{i_1}\nabla_{i_2}h(Q)|_2\\ &\geq \cdots \geq 3^dC^dm^d\tau^{d/4}\log(\tau^{-1})^{d^2/2}|\nabla_{i_1}\cdots\nabla_{i_d}h(Q)|_2.\end{aligned}$$

Assuming that all of the above hold, then

$$\begin{aligned}|p(X) - p_0(X)| &\leq |p(X) - L(p_0)(X)| + |p_0(X) - L(p_0)(X)|\\ &\leq |p_0(X)|/2 + |p_0(X) - L(p_0)(X)|.\end{aligned}$$

By Lemma 38, we have that letting $Q = (q_1(X), \ldots, q_m(X))$

$$\begin{aligned}|L(p_0)(X) - p_0(X)| &= \left|\sum_{k=1}^{d}\sum_{i_1,\ldots,i_k=1}^{m} A(q_{i_1},\ldots,q_{i_k})(Q)\nabla_{i_1}\cdots\nabla_{i_k}h(Q)\right|\\ &\leq \sum_{k=1}^{d}\sum_{i_1,\ldots,i_k=1}^{m} C\tau^{k/4}\log(\tau^{-1})^{dk/2}|\nabla_{i_1}\cdots\nabla_{i_k}h(Q)|_2\\ &\leq \sum_{k=1}^{d}\sum_{i_1,\ldots,i_k=1}^{m} 3^{-k}m^{-k}|h(Q)|\\ &\leq \sum_{k=1}^{d} 3^{-k}|p_0(X)|\\ &< |p_0(X)|/2.\end{aligned}$$

Combining this with the above, we find that

$$|p(X) - p_0(X)| < |p_0(X)|/2 + |p_0(X)|/2 = |p_0(X)|.$$

Thus, with probability at least

$$1 - O_{d,m}(\tau^{1/4}N\log(\tau^{-1})^{d(m+1)/2+1} + \varepsilon^{1/d}\log(\varepsilon^{-1})^{1/2})$$

that $\operatorname{sgn}(p(X)) = \operatorname{sgn}(p_0(X))$. □

6.3. *The regularity lemma.* In this section, we will prove Theorem 34. Much of it will be along the lines of the proof of Theorem 1 with some extra work being done to ensure that the resulting $q_i$ are regular. We begin with a lemma on the regularity of restrictions of polynomials.

LEMMA 40. *Let $p$ be a degree-$d$ multilinear polynomial with $|p|_2 \leq 1$. Let $1/2 > \varepsilon > 0$ be a real number. Then there exists an $M = O_d(\varepsilon^{-1}\log(\varepsilon^{-1})^d)$ so*

*that for any set $S$ of coordinates containing the $M$ coordinates of highest influence for $p$, if we let $A$ be a random Bernoulli variable over the coordinates in $S$ and let $p_A$ be the polynomial over the remaining coordinates upon plugging these values into the coordinates of $S$ then with probability $1-\varepsilon$*

$$\max_i\bigl(\mathrm{Inf}_i(p_A)\bigr) \le \varepsilon.$$

PROOF. We assume throughout that $\varepsilon$ is sufficiently small. Note that the sum of the influences of $p$ is $O_d(1)$, therefore, if $M$ is a sufficiently large multiple of $\varepsilon^{-1}\log(\varepsilon^{-1})^d$, we have that the largest influence of a coordinate not in $S$ is at most a small constant times $\varepsilon\log(\varepsilon^{-1})^{-d}$. Note that for each $i \notin S$, there is a polynomial $p_i$ of degree at most $d$ so that $\mathrm{Inf}_i(p_A) = p_i(A)^2$. Furthermore, it is easy to check that $\mathbb{E}[p_i(A)^2] = \mathrm{Inf}_i(p)$. Applying Corollary 5, we find that if $M$ were chosen to be sufficiently large, then with probability at most $\varepsilon^4/2$ is any given $\mathrm{Inf}_i(p_A)$ more than $\varepsilon$. Taking a union bound over $i$, we find that with probability at most $\varepsilon/2$ is some $\mathrm{Inf}_i(p_A) > \varepsilon$ for any $i$ with $\mathrm{Inf}_i(p) > d\varepsilon^3$. Consider the polynomial

$$q(A) = \sum_{j:\mathrm{Inf}_j(p)\le d\varepsilon^3} \mathrm{Inf}_j(p_A)^2.$$

Note that $|\mathrm{Inf}_j(p_A)^2|_2 = O_d(\mathrm{Inf}_j(p)^2)$ by Lemma 3. Thus,

$$|q|_2 \le O_d(1) \sum_{j:\mathrm{Inf}_j(p)\le d\varepsilon^3} \mathrm{Inf}_j(p)^2 \le O_d\bigl(\varepsilon^3\bigr) \sum_{j:\mathrm{Inf}_j(p)\le d\varepsilon^3} \mathrm{Inf}_j(p) = O_d\bigl(\varepsilon^3\bigr).$$

Thus, by Corollary 5, $q(A) > \varepsilon^2$ with probability at most $\varepsilon/2$. On the other hand, if $q(A) \le \varepsilon^2$, it implies that $\mathrm{Inf}_j(p_A) \le \varepsilon$ for all $j$ so that $\mathrm{Inf}_j(p) \le d\varepsilon^3$. Thus, with probability at most $\varepsilon$ is any $\mathrm{Inf}_j(p_A)$ more than $\varepsilon$. □

LEMMA 41. *Let $p$ be a degree-$d$ multilinear polynomial. Let $S$ be a set of coordinates and $A$ a Bernoulli random variable over those coordinates. Let $p_A$ be the restricted polynomial when the coordinates of $A$ are plugged into $p$. Then*

$$\Pr\bigl(|p_A|_2 \ge N|p|_2\bigr) = O_d\bigl(2^{-(N/2)^{1/d}}\bigr).$$

PROOF. Note that $|p_A|_2^2$ is a polynomial in $A$ of degree at most $2d$. Note that the squared $L_2$ norm of this polynomial is

$$\mathbb{E}_A\bigl[\mathbb{E}_B\bigl[p(A,B)^2\bigr]^2\bigr] \le \mathbb{E}_{A,B}\bigl[p(A,B)^4\bigr] = |p|_{4,B}^4 \le O_d\bigl(|p|_2^4\bigr).$$

The result now follows from Corollary 26. □

The main parts of the proof of Theorem 34 are contained in the following proposition.

PROPOSITION 42. *Let $p$ be a degree-$d$ multilinear polynomial and let $\varepsilon, c, M > 0$ for $1/2 > \varepsilon$. Then $p$ can be written as a decision tree of depth*

$$O_{c,d,M}\bigl(\varepsilon^{-1}\log(\varepsilon^{-1})^{O(d)}\bigr)$$

*with coordinate variables for internal nodes and polynomials for leaves so that for a random leaf $p_\rho$ we have with probability $1 - O_{c,d,M}(\varepsilon)$ that there exists a $p_0$ with $|p - p_0|_{2,B} \le \varepsilon^N |p|_2$ and so that $p_0$ has an $(\varepsilon, \varepsilon^{-c})$-diffuse decomposition $(h, q_1, \ldots, q_m)$ with $m = O_{c,d,N}(1)$, $q_i$ multilinear and so that $\operatorname{Inf}_j(q_i) \le \varepsilon$ for each $i, j$.*

PROOF. The proof is along the same lines as the proof of Theorem 1, with some extra work done to ensure that the influences can be controlled. We assume that $|p|_2 = 1$ and assume throughout that $\varepsilon$ is sufficiently small.

We define: a *partial decomposition* of our polynomial $p$ to be a set of the following data:

- A positive integer $m$.
- A polynomial $h : \mathbb{R}^m \to \mathbb{R}$.
- A sequence of multilinear polynomials $(q_1, \ldots, q_m)$ each on $\mathbb{R}^n$ with $|q_i|_2 = 1$ for each $i$.
- A sequence of integers $(a_1, \ldots, a_m)$ with $a_i$ between 0 and $4 \cdot 3^i(N+1)/c - 1$.

Furthermore, we require that each $q_i$ is nonconstant, and that for any monomial $\prod x_i^{\alpha_1}$ appearing in $h$ that $\sum \alpha_i \deg(q_i) \le d$.

We say that such a partial decomposition has complexity at most $C$ if the following hold:

- $m \le C$.
- $|h|_2 \le C\varepsilon^{-1+C^{-1}}$.
- $|p(A) - h(\varepsilon^{a_i c/(2\cdot 3^i)} q_i(A))|_{2,B} \le C\varepsilon^{N+1}\log(\varepsilon^{-1})^C$.

We define the weight of a partial decomposition as follows. First, we define the polynomial:

$$w(x) = \sum_{i=1}^{m} x^{\deg(q_i)}\bigl(4 \cdot 3^i(N+1)/c - a_i\bigr).$$

We then let the weight of the decomposition be $w(\omega)$.

We prove by ordinal induction on $w$ that if $p$ has a partial decomposition of weight $w$ and complexity $C$, then there is a decision tree of depth $O_{c,C,d,N,w}(\varepsilon^{-1}\log(\varepsilon^{-1})^{O(d)})$ so that with probability $1 - O_{c,C,d,w,N}(\varepsilon)$ a random leaf has such a $p_0$ with a diffuse decomposition into multilinear polynomials whose influences are at most $\varepsilon$.

Again the idea of the proof is to show that after a decision tree of appropriate depth and with appropriate probability, that we either have such a $p_0$ or that we

have a partial decomposition with smaller weight. By Lemma 40, if we restrict to random values of the $O_d(\varepsilon^{-1}\log(\varepsilon^{-1})^{O(d)})$ highest influence coordinates of each of the $q_i$, we will have all influences of all of the $q_i$ at most $\varepsilon$ with probability $1 - O_{d,m}(\varepsilon)$. Applying Lemma 41 to the $q_i$ and $p - h(q_1, \dots, q_m)$, we find that with probability $1 - O_{d,m}(\varepsilon)$ that the restricted values of $q_i$ have norm at most $\log(\varepsilon^{-1})^{O(d)}$ and that the $L^2$ norm of $p - h(q_1, \dots, q_m)$ increased by at most a similar factor. Thus, rescaling the $q_i$ and modifying $h$ appropriately, we find that with probability $1 - O_{d,m}(\varepsilon)$ over our restrictions, we have a new partial decomposition of weight $w$ and complexity $O_C(1)$ so that $\mathrm{Inf}_i(q_j) \le \varepsilon$ for each $i$ and $j$. As in Lemma 16, we show that either $(h, q_1, \dots, q_m)$ is an $(\varepsilon, \varepsilon^{-c})$-diffuse set or that we have a partial decomposition of strictly smaller weight and with complexity $O_{c,C,d,N}(1)$. The proof follows through identically to the proof in Lemma 16 with the additional caveat that the $A_\ell$, $B_\ell$ can be chosen to be multilinear. This is because the $q_i$ are multilinear, so keeping only the multilinear parts of the $A_\ell$, $B_\ell$ only reduces the error produced by the approximation. This completes the proof. □

We will need one more lemma about decision trees of polynomials before we proceed.

LEMMA 43. *Let $p$ be a multilinear, degree-$d$ polynomial. Let $T$ be some decision tree over it's coordinates. If $T$ is evaluated making random, independent choices at each step, and the restricted function is called $p_\rho$, then with probability at least $2^{O(d)}$ over these choices we have that*

$$|p_\rho|_2 \ge |p|_2/2.$$

PROOF. Given a partially filled-in decision tree $T'$ define $V(T') = \mathbb{E}[p(A)^2|T']$. It is clear that $V$ is a martingale. Therefore, $V^2$ is a submartingale. In particular, this means that the expectation of $V^2$ over some decision tree is at most the expectation over an extended decision tree that eventually decides values for all coordinates. This latter expectation is $|p|_{4,B}^4 = 2^{O(d)}|p|_2^4$. Therefore, the expectation over fills of $T$ of $V$ is $|p|_2^2$ and the expectation of $V^2$ is at most $2^{O(d)}|p|_2^4$. Therefore, by the Paley–Zygmund inequality with probability at least $2^{O(d)}$, we have that $V \ge |p|_2^2/4$, proving our lemma. □

We are now prepared to prove Theorem 34.

PROOF. We claim that for $\tau$ sufficiently small that a correctly constructed decision tree of depth $O_{c,d,M}(\tau^{-1}\log(\tau^{-1})^{O(d)})$ yields a restriction with the desired property with probability at least $2^{O(d)}$. Repeating this process up to $2^{O(d)}\log(\tau^{-1})$ many times upon failure will guarantee an aggregate success probability of $1 - \tau$.

To do this, we construct the decision tree given by Proposition 42 for $N = M + d + 2$ and $\varepsilon = \tau$. We claim that if the restricted polynomial has $L^2$ norm at least $|p|_2/2$ (which happens with probability $2^{O(d)}$ by Lemma 43), then the resulting polynomial has the desired property.

Let $P$ be the resulting polynomial. We have a polynomial $p_0$ with an appropriate diffuse decomposition into multilinear polynomials with sufficiently small influences and so that $|P - p_0|_{2,B} = O_{c,d,M}(\tau^{M+d+2})|P|_2$. If $\mathrm{Var}(p_0) \geq \tau^{M+d}|P|_2^2$, we have an appropriate regular decomposition. Otherwise, $\mathrm{Var}(p_0) \leq \tau^{M+d}|P|_2^2$. This implies that for some $\mu$ that $|p_0 - \mu|_2^2 \leq \tau^{M+d}|P|_2^2$. Thus, by Lemma 20 we have that $|h - \mu|_2^2 \leq O_{c,d,M}(\tau^M)|P|_2^2$. From this, it is easy to see that the sum of the squares of the coefficients of $h - \mu$ is $O_{c,d,M}(\tau^M)|P|_2^2$. From this, it is easy to verify that the variance of $p_0$ over Bernoulli inputs is $O_{c,d,M}(\tau^M)|P|_2^2$. Therefore, due to the small difference between $p$ and $p_0$ under Bernoulli inputs, we have that $\mathrm{Var}(P) \leq O_{c,d,M}(\tau^M)|P|_2^2$, which satisfies one of the necessary conditions. □

## 7. Application to noise sensitivity of polynomial threshold functions.

### 7.1. *Background of noise sensitivity results.*

7.1.1. *Definitions.* If $f : \mathbb{R}^n \to \{-1, 1\}$ is a boolean function, the noise sensitivity of $f$ is a measure of the likelihood that a small change in the input value to $f$ changes the output. There are several different notions of noise sensitivity, suitable for slightly different contexts. We present their definitions here.

DEFINITION 14. For $f : \mathbb{R}^n \to \{-1, 1\}$ a Boolean function, we define its *average sensitivity* (also known as the total influence) to be

$$\mathbb{AS}(f) := \sum_{i=1}^{n} \mathrm{Pr}_{A \sim_u \{-1,1\}^n}\big(f(A) \neq f\big(A^{(i)}\big)\big),$$

where $A^{(i)}$ is obtained from $A$ by flipping the sign of the $i$th coordinate. In other words, the average sensitivity is the expected number of coordinates of $A$ that could be changed in order to change the value of $f$.

We also define the average sensitivity in the Gaussian setting.

DEFINITION 15. For $f : \mathbb{R}^n \to \{-1, 1\}$ a boolean function, we define its *Gaussian average sensitivity* to be

$$\mathbb{GAS}(f) := \sum_{i=1}^{n} \mathrm{Pr}\big(f(X) \neq f\big(X^{(i)}\big)\big),$$

where above $X$ is a Gaussian random variable and $X^{(i)}$ is obtained from $X$ by replacing the $i$th coordinate by an independent random Gaussian.

A related notion is that of noise sensitivity in the Bernoulli or Gaussian context. Whereas average sensitivity counts the expected number of coordinates that could be changed to alter the sign of $f$, noise sensitivity measures the probability that the sign of $f$ changes if each coordinate is changed by a small amount. In particular, we define the following.

DEFINITION 16. For $f : \mathbb{R}^n \to \{-1, 1\}$ a Boolean function, and $1 \geq \delta \geq 0$ we define the *noise sensitivity of $f$ with parameter $\delta$* to be

$$\mathbb{NS}_\delta(f) := \Pr\big(f(A) \neq f(B)\big),$$

where $A$ and $B$ are Bernoulli random variables with $B$ obtained from $A$ by flipping the sign of each coordinate randomly and independently with probability $\delta$.

DEFINITION 17. For $f : \mathbb{R}^n \to \{-1, 1\}$ a Boolean function, and $1 \geq \delta \geq 0$ we define the *Gaussian noise sensitivity of $f$ with parameter $\delta$* to be

$$\mathbb{GNS}_\delta(f) := \Pr\big(f(X) \neq f(Y)\big),$$

where $X$ and $Y$ are Gaussian random variables that together form a joint Gaussian with

$$\operatorname{Cov}(X_i, Y_j) = \begin{cases} (1-\delta), & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases}$$

7.1.2. *Previous work.* The main conjecture about the noise sensitivity of polynomial threshold functions was given in [9].

CONJECTURE 44 (Gotsman–Linial). *Let $f$ be a degree-$d$ polynomial threshold function in $n$ variables, then*

$$\mathbb{AS}(f) \leq 2^{-n+1} \sum_{k=0}^{d-1} \binom{n}{\lfloor (n-k)/2 \rfloor} (n - \lfloor (n-k)/2 \rfloor).$$

REMARK 5. It should be noted that the upper bound conjectured above is actually obtainable. In particular, if $f$ is the polynomial threshold function associated to the polynomial

$$\prod_{i=1}^{d} \left( \sum_{j=1}^{n} A_j - d + 2i - 1/2 \right)$$

achieves this bound.

In particular, Conjecture 44 implies that

$$\mathbb{AS}(f) = O(d\sqrt{n}).$$

By the work of [11], this implies bounds on other notions of sensitivity. In particular, it would imply that

$$\mathbb{NS}_\delta(f) = O(d\sqrt{\delta})$$

and

$$\mathbb{GNS}_\delta(f) = O(d\sqrt{\delta}).$$

Furthermore, this would imply the following bound on the Gaussian average sensitivity:

$$\mathbb{GAS}(f) = O(d\sqrt{n}).$$

In particular, we have the following.

LEMMA 45. *The largest Gaussian average sensitivity of any degree-d polynomial threshold function in n variables is at most the largest average sensitivity of a degree-d polynomial threshold function in n variables.*

PROOF. We will show that if $f$ is a degree-$d$ PTF in $n$ variables, then $\mathbb{GAS}(f)$ can be written as an expectation over the average sensitivities of certain other degree-$d$ PTFs in $n$ variables. The key to this argument is to produce the correct distribution on pairs of Gaussians that differ in exactly one coordinate in an unusual way. In particular, we define $n$-variable Gaussians $Z$ and $Z'$ as follows:

$$Z_i = \frac{1}{\sqrt{2}}(X_i + A_i Y_i), \qquad Z'_i = \frac{1}{\sqrt{2}}(X_i + B_i Y_i),$$

where $X_i, Y_i$ are independent Gaussian random variables, and $A = (A_1, \ldots, A_n)$, $B = (B_1, \ldots, B_n)$ are Bernoulli random variables that differ only in a single random coordinate and are independent of $X$ and $Y$. It is clear that $Z$ and $Z'$ are random Gaussians that agree in all but one of their coordinates, and that they are independent in the coordinate on which they differ. Thus,

$$\mathbb{GAS}(f) = \Pr\bigl(f(Z) \neq f\bigl(Z'\bigr)\bigr).$$

On the other hand, after fixing values of $X$ and $Y$, we may define a new degree-$d$ PTF $f_{X,Y}$ by

$$f_{X,Y}(A) := f\left(\frac{1}{\sqrt{2}}(X_i + A_i Y_i)\right).$$

Therefore, we have that

$$\begin{aligned}\mathbb{GAS}(f) &= \Pr\bigl(f(Z) \neq f\bigl(Z'\bigr)\bigr)\\ &= \mathbb{E}_{X,Y}\bigl[\Pr\bigl(f_{X,Y}(A) \neq f_{X,Y}(B)\bigr)\bigr]\\ &= \mathbb{E}_{X,Y}\bigl[\mathbb{AS}(f_{X,Y})\bigr].\end{aligned}$$

This is at most the maximum possible average sensitivity of a degree-$d$ PTF in $n$ variables. $\square$

Proving the conjectured bounds for the various notions of sensitivity has proved to be quite difficult. The degree-1 case of Conjecture 44 was known to Gotsman and Linial. The first nontrivial bounds for higher degrees were obtained independently by [11] and [5], who later combined their papers into [4]. They essentially proved bounds on average sensitivities of $O_d(n^{1-1/O(d)})$ and bounds on noise sensitivities of $O_d(\delta^{1/O(d)})$. For the special case of Gaussian noise sensitivity, the author proved essentially optimal bounds in [12] of $O(d\sqrt{\delta})$. Only recently were better bounds obtained for the other cases. In this paper, we prove a bound on $\mathbb{AS}(f)$ of $O_{c,d}(n^{5/6+c})$, though note that this bound has been superseded by [14], which improved the bound to $\sqrt{n}\log(n)^{O(d\log(d))}2^{O(d^2\log(d))}$.

In this section, we show how the theory of diffuse decompositions can be used to obtain the bound $\mathbb{AS}(f) = O_{c,d}(n^{5/6+c})$. Our basic technique will be to compare $\mathbb{NS}_\delta(f)$ to $\mathbb{GNS}_{2\delta}(f)$ using an appropriate invariance principle. It should be noted that this idea could have been applied using traditional means, but that the bound obtained would not have been better than $\delta^{1-O(1/d)}$.

7.2. *Noise sensitivity bounds.* In this section, we prove the following three theorems.

THEOREM 46. *If $f$ is a degree-$d$ polynomial threshold function, and if $c, \delta > 0$, then*

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

THEOREM 47. *If $f$ is a degree-$d$ polynomial threshold function in $n$ variables, and if $c > 0$, then*

$$\mathbb{AS}(f) = O_{c,d}\big(n^{5/6+c}\big).$$

THEOREM 48. *For $f$ a degree-$d$ polynomial threshold function in $n$ variables and $c > 0$,*

$$\mathbb{GAS}(f) = O_{c,d}\big(n^{5/6+c}\big).$$

We begin with the proof of Theorem 46 in the case of regular polynomial threshold function.

PROPOSITION 49. *Let $f = \operatorname{sgn} \circ p$ be a polynomial threshold function for $p$ a degree-$d$ polynomial with a $(\tau, N, m, \varepsilon)$-regular decomposition for $1/2 > \varepsilon, \tau > 0$. Let $1 > \delta > 0$, then*

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

The proof of Proposition 49 will be to use the replacement method to show that $\mathbb{NS}_\delta(f)$ is approximately $\mathbb{GNS}_{2\delta}(f)$, which we bound using the main theorem of [12]. Unfortunately, we will not be able to apply Proposition 21 directly, but many of the techniques will be similar.

PROOF. Let $A^1, A^2$ be a pair of Bernoulli random variables so that for each coordinate $i$, $A^1_i$ and $A^2_i$ are equal with probability $1-\delta$ independently over different $i$. $\mathbb{NS}_\delta(f) = \Pr(f(A^1) \neq f(A^2)) = 2\Pr(f(A^1) = 1, f(A^2) = -1)$. We wish to bound this later probability.

Let $X^1$ and $X^2$ be Gaussian random variables so that the joint distribution $(X^1, X^2)$ is a Gaussian with

$$\operatorname{Cov}(X^1_i, X^2_j) = \begin{cases} 1-2\delta, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases}$$

Note that all of the first three moments of $(A^1, A^2)$ are identical to the corresponding moments of $(X^1, X^2)$.

We are given that there exists a polynomial $p_0$ with $|p - p_0|^2_{2,B} < \varepsilon \operatorname{Var}(p_0)$ so that $p_0$ has a $(\tau^{1/5}, N)$-diffuse decomposition $(h, q_1, \ldots, q_m)$ with $q_i$ multilinear and $\operatorname{Inf}_i(q_j) \leq \tau$ for all $i, j$. After rescaling these polynomials, we may assume that $\operatorname{Var}(p_0) \leq |p_0|^2_2 = 1$. Note that by Corollary 26 that with probability $1 - O(\varepsilon)$ that $|p(A^i) - p_0(A^i)| < \varepsilon^{1/2}\log(\varepsilon^{-1})^d$ for each of $i = 1, 2$. By Proposition 17, there exist functions $f^1, f^2 : \mathbb{R}^m \to [0,1]$ so that:

- $f^1(x) = 1$ if $h(x) + \varepsilon^{1/2}\log(1+\varepsilon^{-1})^d > 0$.
- $f^2(x) = 1$ if $h(x) - \varepsilon^{1/2}\log(1+\varepsilon^{-1})^d < 0$.
- $$\begin{aligned} &|\mathbb{E}[f^1(q_1(X^1), \ldots, q_m(X^1))] \\ &\quad - \mathbb{E}[I_{(0,\infty)}(h(q_1(X^1), \ldots, q_m(X^1)) + \varepsilon^{1/2}\log(1+\varepsilon^{-1})^d)]| \\ &\quad = O_{d,m}(N\tau^{1/5}\log(\tau^{-1})^{dm/2+1}). \end{aligned}$$
- $$\begin{aligned} &|\mathbb{E}[f^2(q_1(X^2), \ldots, q_m(X^2))] \\ &\quad - \mathbb{E}[I_{(-\infty,0)}(h(q_1(X^1), \ldots, q_m(X^1)) - \varepsilon^{1/2}\log(1+\varepsilon^{-1})^d)]| \\ &\quad = O_{d,m}(N\tau^{1/5}\log(\tau^{-1})^{dm/2+1}). \end{aligned}$$
- $|(f^i)^{(k)}|_\infty = O_m(\tau^{-k/5})$ for $1 \leq k \leq 4$.

We then have that

$$\begin{aligned} \mathbb{NS}_\delta(f) &= 2\Pr(f(A^1) = 1, f(A^2) = -1) \\ &\leq 2\mathbb{E}[f^1(q_1(A^1), \ldots, q_m(A^1))f^2(q_1(A^2), \ldots, q_m(A^2))] + O(\varepsilon). \end{aligned}$$

We would like to relate

$$\mathbb{E}[f^1(q_1(A^1),\dots,q_m(A^1))f^2(q_1(A^2),\dots,q_m(A^2))]$$

to

$$\mathbb{E}[f^1(q_1(X^1),\dots,q_m(X^1))f^2(q_1(X^2),\dots,q_m(X^2))].$$

In particular, we have that with respect to the Gaussian distribution, $f^i(q_1(X),\dots,q_m(X))$ differs from $I_{(0,\infty)}(\pm(p_0(X)-\varepsilon^{1/2}\log(\varepsilon^{-1})^d))$ with $L^1$ error at most $O_{d,m}(N\tau^{1/5}\log(\tau^{-1})^{dm/2+1})$. This in turn differs from $I_{(0,\infty)}(\pm p_0(X))$ with probability at most $O(d\varepsilon^{1/2d}\log(\varepsilon^{-1}))$ by Lemma 2. Hence, we have that

$$\begin{aligned}
&\mathbb{E}[f^1(q_1(X^1),\dots,q_m(X^1))f^2(q_1(X^2),\dots,q_m(X^2))]\\
&\quad= O\big(d\varepsilon^{1/2d}\log(\varepsilon^{-1})\big)+O_{d,m}\big(N\tau^{1/5}\log\big(\tau^{-1}\big)^{dm/2+1}\big)\\
&\qquad+\mathbb{E}[I_{(0,\infty)}(p(X^1))I_{(-\infty,0)}(p(X^2))]\\
&\quad= O\big(d\varepsilon^{1/2d}\log(\varepsilon^{-1})\big)+O_{d,m}\big(N\tau^{1/5}\log\big(\tau^{-1}\big)^{dm/2+1}\big)+\mathbb{GNS}_{2\delta}(f)\\
&\quad= O\big(d\varepsilon^{1/2d}\log(\varepsilon^{-1})\big)+O_{d,m}\big(N\tau^{1/5}\log\big(\tau^{-1}\big)^{dm/2+1}\big)+O(d\sqrt{\delta}),
\end{aligned}$$

where the bound on the Gaussian noise sensitivity comes from the main theorem of [12].

Thus, we are left with the task of bounding the difference between $\mathbb{E}[f^1(q_i(A^1))f^2(q_i(A^2))]$ and $\mathbb{E}[f^1(q_i(X^1))f^2(q_i(X^2))]$. We do this with the replacement method. We let $Z^{i,\ell}$ be the random vector whose $j$th component is $A^i_j$ if $j>\ell$ and $X^i_j$ otherwise. We note that $Z^{i,0}=A^i$ and $Z^{i,n}=X^i$. We proceed to bound the difference

$$|\mathbb{E}[f^1(q_i(Z^{1,j-1}))f^2(q_i(Z^{2,j-1}))]-\mathbb{E}[f^1(q_i(Z^{1,j}))f^2(q_i(Z^{2,j}))]|. \tag{17}$$

We note that $Z^{i,j-1}$ and $Z^{i,j}$ agree in all but the $j$th coordinate. Thus, in bounding the difference above we may consider all but the $j$th coordinate fixed. We then approximate the resulting function of $Z^1_j$, $Z^2_j$ by it's Taylor series. In particular, if we let $z_i=Z^i_j$, then for appropriate functions $g_1$ and $g_2$ (depending on the other coordinates of $Z$) we need to consider $\mathbb{E}[g_1(z_1)g_2(z_2)]$. Taylor expanding about $(0,0)$, we have that $g_1(z_1)g_2(z_2)$ equals a degree 3 polynomial in $z_1$ and $z_2$ plus an error of at most

$$\begin{aligned}
&z_1^4g_1''''(t_1)g_2(0)/24+z_1^3z_2g_1'''(t_2)g_2'(t_3)/6+z_1^2z_2^2g_1''(t_4)g_2''(t_5)/4\\
&\qquad+z_1z_2^3g_1'(t_6)g_2'''(t_7)/6+z_2^4g_1(0)g_2''''(t_8)/24
\end{aligned}$$

for some points $t_i$. Since the expectations of the degree 3 polynomials in $z_1$ and $z_2$ are the same in the Bernoulli and Gaussian case, and since the fourth moments are bounded, we have that the difference in equation (17) is

$$O\big(\mathbb{E}[|g_1''''|_\infty+|g_1'''g_2'|_\infty+|g_1''g_2''|_\infty+|g_1'''g_2'|_\infty+|g_2''''|_\infty]\big).$$

Now the $k$th derivative of $g_i$ can be written as

$$\sum_{i_1,\dots,i_k=1}^{m} \frac{\partial^k f^i}{\partial q_{i_1}\cdots\partial q_{i_k}} \prod_{\ell=1}^{k} \frac{\partial q_{i_\ell}(Z^i)}{\partial z_j}.$$

On the other hand, by assumption, this partial derivative of $f^i$ is at most $\tau^{-k/5}$, and the product is at most

$$\left(\max_{\ell} \frac{\partial q_\ell}{\partial x_j}\right)^k.$$

Thus, the total error in equation (17) is at most

$$O\left(m^4\tau^{-4/5}\mathbb{E}\left[\sum_{\ell=1}^{m}\sum_{i=1}^{2}\left(\frac{\partial q_\ell(Z^i)}{\partial z_j}\right)^4\right]\right).$$

It is clear that

$$\mathbb{E}\left[\left(\frac{\partial q_\ell(Z^i)}{\partial z_j}\right)^2\right] = \mathrm{Inf}_j(q_\ell).$$

Thus, $\frac{\partial q_\ell(Z^i)}{\partial z_j}$ is a polynomial in independent Bernoulli and Gaussian random variables with second moment $\mathrm{Inf}_j(q_\ell)$. Therefore, by Lemma 27 its fourth moment is $O_d(\mathrm{Inf}_j(q_\ell)^2)$. Therefore, we have that the expression in equation (17) is at most

$$O_{d,m}\left(\tau^{-4/5}\sum_{\ell}\mathrm{Inf}_j^2(q_\ell)\right).$$

Therefore, summing this over $j$, we get that

$$\left|\mathbb{E}\left[f^1(q_i(A^1))f^2(q_i(A^2))\right] - \mathbb{E}\left[f^1(q_i(X^1))f^2(q_i(X^2))\right]\right|$$

is at most

$$O_{d,m}\left(\tau^{-4/5}\sum_{j,\ell}\mathrm{Inf}_j^2(q_\ell)\right).$$

On the other hand, for fixed $\ell$ we have that $\sum_j \mathrm{Inf}_j(q_\ell) = O_d(1)$ by Corollary 30 and that for each $j$ that $\mathrm{Inf}_j(q_\ell) \le \tau$. Therefore, $\sum_j \mathrm{Inf}_j^2(q_\ell) = O_d(\tau)$. Thus, we have that

$$\left|\mathbb{E}\left[f^1(q_i(A^1))f^2(q_i(A^2))\right] - \mathbb{E}\left[f^1(q_i(X^1))f^2(q_i(X^2))\right]\right| = O_{d,m}(\tau^{1/5}).$$

Recall though that

$$\mathbb{NS}_\delta \le \mathbb{E}\left[f^1(q_i(A^1))f^2(q_i(A^2))\right] + O(\varepsilon)$$

and that

$$\mathbb{E}[f^1(q_i(X^1))f^2(q_i(X^2))] \\ = O(d\varepsilon^{1/2d}\log(1+\varepsilon^{-1})) + O_{d,m}(N\tau^{1/5}\log(\tau^{-1})^{dm/2+1}) + O(d\sqrt{\delta}).$$

Combining these yields our result. □

We are now prepared to prove Theorem 46.

PROOF. Write $f = \operatorname{sgn} \circ p$ for $p$ a degree-$d$ polynomial. We will reduce to the case of Proposition 49 by use of Theorem 34. In particular, we may write $p$ as a decision tree of depth $O_{c,d}(\delta^{-5/6}\log(\delta^{-1})^{O(d)})$ so that a $1-\delta^{5/6}$ fraction of the leaves are polynomials with either a $(\delta^{5/6}, \delta^{-c/2}, O_{c,d}(1), \delta^{2d})$-regular decomposition or with variance less than $\delta^2$ times their squared mean.

Consider $A^1$ and $A^2$ random Bernoulli variables that differ in each coordinate independently with probability $\delta$. Consider the path on the decision tree above followed by $A^1$. With probability at least $1-\delta^{5/6}$, the resulting leaf satisfies one of the two cases specified by Theorem 34. Furthermore, with probability at least $1 - O_{c,d}(\delta^{1/6}\log(\delta^{-1})^{O(d)})$, $A^2$ agrees with $A^1$ on all coordinates queried by the decision tree. Conditioned on this occurrence, the probability that $p(A^1)$ and $p(A^2)$ have different signs is equal to the noise sensitivity with parameter $\delta$ of the polynomial threshold function defined by the leaf. If the leaf has a $(\delta^{5/6}, \delta^{-c/2}, O_{c,d}(1), \delta^{2d})$-regular decomposition, this is $O_{c,d}(\delta^{1/6-c})$ by Proposition 49. If this polynomial has low variance compared to its mean, then both $p(A^1)$ and $p(A^2)$ are the same sign as the mean of $p$ with high probability by Corollary 26. Thus, we have that

$$\mathbb{NS}_\delta(f) \le \delta^{5/6} + O_{c,d}(\delta^{1/6}\log(\delta^{-1})^{O(d)}) + O_{c,d}(\delta^{1/6-c}) = O_{c,d}(\delta^{1/6-c}). \quad \square$$

Theorem 47 now follows immediately by Lemma 8.1 of [11], and Theorem 48 follows from Theorem 47 and Lemma 45.

**8. Application to PRGs for PTFs with Bernoulli inputs.** In [17], Meka and Zuckerman developed a relatively small pseudo-random generator of polynomial threshold functions with Bernoulli inputs. Their generator was defined as follows. Let $h : [n] \to [a]$ be a hash function picked from a 2-independent family. Let $A^1, \dots, A^a : [n] \to \{-1, 1\}$ be chosen independently from a $k$-independent hash family. Meka and Zuckerman's generator is given by $A_i = A_i^{h(i)}$. Meka and Zuckerman show that for appropriate chosen $m = \tilde{O}(\varepsilon^{-2})$ and $a = O(\varepsilon^{-O(d)})$ that this generator fools all degree-$d$ polynomial threshold functions to within $\varepsilon$.

Meka and Zuckerman's proof is essentially to think of $h$ as constant and to use the replacement method to bound the expected errors as the $A^i$ are replaced

by random Gaussians vectors one at a time. If the polynomial in question is sufficiently regular, then these errors will be small, and thus, the expected value of the PTF in question over the PRG will be close to the expected value of the PTF over random Gaussian inputs, and by the invariance principle, the expected value at random Bernoulli inputs will also be close. Unfortunately, this technique had been limited by the classical invariance principle and regularity lemma, and thus, could not produce a PRG of seed length less than $\varepsilon^{-O(d)}$. In this section, we will show how our diffuse invariance principle and regularity lemma can improve this to produce a PRG of seed length $O_d(\log(n)\varepsilon^{-O(1)})$.

We begin by producing a pseudorandom generator that works in the case of regular polynomials, and then reducing the general case to this one.

8.1. *The regular case.*

PROPOSITION 50. *Let $p$ be a degree-$d$ polynomial in $n$ variables with a $(\tau, N, m, \varepsilon)$-regular decomposition. Let $a$ be a positive integer. Let $h : [n] \to [a]$ be picked randomly from a 2-independent hash family and for each $h$ let $A^1, \dots, A^a : [n] \to \{-1, 1\}$ be picked independently from $4d$-independent hash families. Define the $n$-variable function $A$ in terms of $h$ and $A^i$ as $A_i = A_i^{h(i)}$. Then if $B$ is a Bernoulli random variable, $|\mathbb{E}[\operatorname{sgn}(p(A))] - \mathbb{E}[\operatorname{sgn}(p(B))]|$ is at most*

$$O_{d,m}\bigl(N\tau^{1/5}\log(1+\tau^{-1})^{dm/2+1}\bigr) + O\bigl(d\varepsilon^{1/d}\log(\varepsilon^{-1})^{1/2}\bigr) + O\bigl(a^{-1}\tau^{-1}\bigr).$$

We begin by showing that a similar statement holds for an appropriate choice of $h$.

LEMMA 51. *Let $p$ and $p_0$ be degree-$d$ polynomials with $|p - p_0|_{2,B}^2 \le \varepsilon^2 \operatorname{Var}(p_0)$ so that $p_0$ has a $(\tau, N)$-diffuse decomposition $(g, q_1, \dots, q_m)$ with $q_i$ multilinear $(1/2 > \varepsilon, \tau > 0)$. Suppose furthermore that $h : [n] \to [a]$ is a function so that*

$$\sum_{j=1}^{a}\Biggl(\sum_{\ell:h(\ell)=j}\sum_{i}\operatorname{Inf}_\ell(q_i)\Biggr)^2 \le \tau.$$

*Let $A^1, \dots, A^a : [n] \to \{-1, 1\}$ be picked independently from a $4d$-independent hash family. Define the random variable $A$ so that its $i$th coordinate is the $i$th coordinate of $A^{h(i)}$. Then for $G$ a random Gaussian we have that*

$$\bigl|\mathbb{E}\bigl[\operatorname{sgn}\bigl(p(A)\bigr)\bigr] - \mathbb{E}\bigl[\operatorname{sgn}\bigl(p_0(G)\bigr)\bigr]\bigr| \le O_{d,m}\bigl(N\tau^{1/5}\log\bigl(\tau^{-1}\bigr)^{dm/2+1}\bigr) + O\bigl(d\varepsilon^{1/2d}\bigr).$$

PROOF. We show that

$$\Pr\bigl(p(A) \le 0\bigr) \le \Pr\bigl(p_0(G) \le 0\bigr) + O_{d,m}\bigl(N\tau^{1/5}\log\bigl(\tau^{-1}\bigr)^{dm/2+1}\bigr) + O\bigl(d\varepsilon^{1/2d}\bigr).$$

The other direction will follow analogously.

First, we note that by Corollary 26 that with probability $1 - O(\varepsilon)$ that $|p(A) - p_0(A)| < \varepsilon^{1/2}\sqrt{\mathrm{Var}(p_0)} \le \varepsilon^{1/2}|p_0|_2$. Therefore,

$$\Pr\big(p(A) \le 0\big) \le \Pr\big(p_0(A) \le -\varepsilon^{1/2}|p_0|_2\big) + O(\varepsilon).$$

On the other hand,

$$\Pr\big(p_0(G) \le -\varepsilon^{1/2}|p_0|_2\big) = \Pr\big(p_0(G) \le 0\big) + O\big(d\varepsilon^{1/2d}\big)$$

by Lemma 2. Hence, it will suffice to prove that

$$\begin{aligned}&\Pr\big(p_0(A) \le -\varepsilon^{1/2}|p_0|_2\big)\\ &\quad\le \Pr\big(p_0(G) \le -\varepsilon^{1/2}|p_0|_2\big) + O_{d,m}\big(N\tau^{1/5}\log\big(\tau^{-1}\big)^{dm/2+1}\big).\end{aligned}$$

Modifying $p_0$ by $\varepsilon^{1/2}|p_0|_2$, it suffices to prove under the same hypothesis that

$$\Pr\big(p_0(A) \le 0\big) \le \Pr\big(p_0(G) \le 0\big) + O_{d,m}\big(N\tau^{1/5}\log\big(\tau^{-1}\big)^{dm/2+1}\big).$$

The proof is by Proposition 21. Let $B^i$ be the vector of entries $A^i_j$ of $A^i$ for which $h(j) = i$. Reordering, the coordinate variables we can make it so that $A = (B^1, \dots, B^a)$. Similarly, let $G = (G^1, \dots, G^a)$. Note that since the $q_i$ are multilinear and degree at most $d$, that any degree-3 polynomial in the $q_i$ has the same expectation under the $B^i$ as under the $G^i$. We may thus apply Proposition 21 with $k = 4$. We have that

$$\begin{aligned}&\big|\Pr\big(p_0(A) \le 0\big) - \Pr\big(p_0(G) \le 0\big)\big|\\ &\qquad = O_{d,m}\big(N\tau^{1/5}\log\big(\tau^{-1}\big)^{dm/2+1} + \tau^{-4/5}T\big).\end{aligned} \tag{18}$$

Recall that $T$ above is

$$\sum_{i,j} T_{i,j},$$

where $T_{i,j}$ is

$$\begin{aligned}&\mathbb{E}\big[\big(q_i\big(B^1, \dots, B^j, G^{j+1}, \dots, G^a\big) - \mathbb{E}_Y\big[q_i\big(B^1, \dots, B^{j-1}, Y, G^{j+1}, \dots, G^a\big)\big]\big)^4\big]\\ &\quad + \mathbb{E}\big[\big(q_i\big(B^1, \dots, B^{j-1}, G^j, \dots, G^a\big)\\ &\quad - \mathbb{E}_Y\big[q_i\big(B^1, \dots, B^{j-1}, Y, G^{j+1}, \dots, G^a\big)\big]\big)^4\big].\end{aligned}$$

By the $4d$-independence of the $B^i$, this expectation is the same as it would be if they were fully independent Bernoulli variables. Thus, by Lemma 27, this is at most

$$\begin{aligned}&O_d\big(\mathbb{E}\big[\big(q_i\big(B^1, \dots, B^j, G^{j+1}, \dots, G^a\big)\\ &\qquad - \mathbb{E}_Y\big[q_i\big(B^1, \dots, B^{j-1}, Y, G^{j+1}, \dots, G^a\big)\big]\big)^2\big]\big)^2\\ &\qquad + O_d\big(\mathbb{E}\big[\big(q_i\big(B^1, \dots, B^{j-1}, G^j, \dots, G^a\big)\\ &\qquad - \mathbb{E}_Y\big[q_i\big(B^1, \dots, B^{j-1}, Y, G^{j+1}, \dots, G^a\big)\big]\big)^2\big]\big)^2.\end{aligned}$$

Since the terms in the expectations above are at most quadratic in any coordinate, the expectation is unchanged by replacing Gaussian inputs with Bernoullis, and hence

$$T_{i,j} = O_d\big(\mathbb{E}_{B^1,\ldots,\hat{B^j},\ldots,B^a}\big[\mathrm{Var}_{B^j}\big(q_i(B)\big)\big]^2\big).$$

The variance above is clearly the sum of the squares of the coefficients of the nonconstant terms of the polynomial obtained by substituting the values of $B^1, \ldots, \hat{B^j}, \ldots, B^a$ into $q_i$. The expectation of this is easily seen to be the sum of the squares of the coefficients of the monomials in $q_i$ containing at least one of the $B^j$ variables. This in turn is clearly at most $\sum_{\ell:h(\ell)=j} \mathrm{Inf}_\ell(q_i)$. Thus,

$$\begin{aligned} T &= \sum_{i=1}^{m}\sum_{j=1}^{a} T_{i,j} \\ &\le \sum_{i=1}^{m}\sum_{j=1}^{a} O_d\Big(\sum_{\ell:h(\ell)=j} \mathrm{Inf}_\ell(q_i)\Big)^2 \\ &\le \sum_{j=1}^{a}\Big(\sum_{\ell:h(\ell)=j}\sum_{i} \mathrm{Inf}_\ell(q_i)\Big)^2 \\ &\le \tau. \end{aligned}$$

Thus, by equation (18),

$$\big|\Pr\big(p_0(A)\le 0\big) - \Pr\big(p_0(G)\le 0\big)\big| = O_{d,m}\big(N\tau^{1/5}\log\big(\tau^{-1}\big)^{dm/2+1}\big),$$

completing our proof. □

We can now prove Proposition 50.

PROOF. Let $q_1, \ldots, q_m$ be as given in the $(\tau, N, m, \varepsilon)$-regular decomposition of $p$.

By the above lemma, it suffices to prove that with probability $1 - O(a^{-1}\tau^{-1})$ over $h$ that

$$\sum_{j=1}^{a}\Big(\sum_{i:h(i)=j}\sum_{\ell} \mathrm{Inf}_i(q_\ell)\Big)^2 = O_{d,m}(\tau).$$

On the other hand, this is at most

$$m\sum_{\ell}\sum_{i} \mathrm{Inf}_i(q_\ell)^2 + m\sum_{\ell}\sum_{i\neq i':h(i)=h(i')} \mathrm{Inf}_i(q_\ell)\,\mathrm{Inf}_{i'}(q_\ell).$$

Since $\sum_i \mathrm{Inf}_i(q_\ell) = O_d(1)$ for each $\ell$ and since each $\mathrm{Inf}_i(q_\ell)$ is at most $\tau$, the first term above is $O_{d,m}(\tau)$. The expectation of the latter term above is

$$m/a \sum_\ell \sum_{i \neq i'} \mathrm{Inf}_i(q_\ell)\,\mathrm{Inf}_{i'}(q_\ell) \leq O_m\left(a^{-1}\sum_\ell \left(\sum_i \mathrm{Inf}_i(q_\ell)\right)^2\right) = O_{d,m}(a^{-1}).$$

Our result follows from the Markov bound on this random variable. □

8.2. *The general case.* We are now prepared to state our conclusions in the general case.

THEOREM 52. *Let $A$ be a random variable defined as follows. Let $h : [n] \to [a]$ be picked randomly from a 2-independent hash family for $a = \varepsilon^{-6}$. Let $A^1, \ldots, A^a : [n] \to \{-1, 1\}$ be picked independently from $k$-independent hash families for $k = \varepsilon^{-5} + 4d$. Let $A_i = A_i^{h(i)}$ for $1 \leq i \leq n$. Note that $A$ can be generated from a seed of length $O(\log(n)\varepsilon^{-11})$. Let $B$ be a random $n$-dimensional Bernoulli random variable, and let $f$ be any degree-$d$ polynomial threshold function in $n$ variables. Then for any $c > 0$*

$$\left|\mathbb{E}\big[f(A)\big] - \mathbb{E}\big[f(B)\big]\right| = O_{c,d}\big(\varepsilon^{1-c}\big).$$

REMARK 6. Note that by changing the values of $a$ and $k$ above we can find a PRG with seed length $O_{c,d}(\log(n)\varepsilon^{-11-c})$ that fools degree-$d$ PTFs to within $\varepsilon$.

PROOF. Note that the coordinates of $A$ are $k$-independent (since they are for each possible value of $h$). Assume that $\varepsilon$ is sufficiently small (since otherwise there is nothing to prove). By Theorem 34, we know that $f$ can be written as a decision tree of depth $\varepsilon^{-5}$ so that with probability $1 - O(\varepsilon)$ a randomly chosen leaf is of the form $\mathrm{sgn} \circ p$ where either $\mathrm{Var}(p(B)) < \varepsilon^2 |\mathbb{E}[p(B)]|$ or $p$ has an $(\varepsilon^5, \varepsilon^{-c/5}, O_{c,d}(1), \varepsilon^{2d})$-regular decomposition. For each such decision-tree path, condition on $A$ and $B$ on having the appropriate values on the appropriate $\varepsilon^{-5}$ coordinates defining this branch of the decision tree. Note that the conditional distribution on $A$ can be written in the same form as $A$ was originally written only with the $A^i$ perhaps only being $4d$-independent.

There is a probability of $1 - O(\varepsilon)$ that $p$ satisfies one of the two conditions outlined above. If the former condition holds, both $p(A)$ and $p(B)$ have the same sign as $\mathbb{E}[p(B)]$ with probability $1 - O(\varepsilon)$. In the latter case, by Proposition 50, we have that for an appropriate $p_0$

$$\mathbb{E}\big[\mathrm{sgn}\big(p(A)\big)\big] = \mathbb{E}\big[\mathrm{sgn}\big(p_0(G)\big)\big] + O_{d,m}\big(\varepsilon^{1-c}\big) = \mathbb{E}\big[\mathrm{sgn}\big(p(B)\big)\big] + O_{d,m}\big(\varepsilon^{1-c}\big)$$

(since $B$ is also of the form specified in Proposition 50). This completes our proof. □

**9. Conclusion.** We have introduced the notion of a diffuse decomposition of a polynomial and proved that they exist for reasonable parameters. This in turn has allowed us to make improvements on known bounds for several major problems relating to polynomial threshold functions. There are several directions in which this work might be expanded. Perhaps most importantly is that the theory introduced in this paper may well have applications to other problems of interest in the field. On the other hand, Theorem 1 still has room for improvement. In particular, I believe that such a diffuse decomposition should exist with size merely polynomial in $dN/c$. Producing such a technical improvement, would allow one to noticeably improve the $d$-dependence in all of the applications presented in this paper.

**Acknowledgments.** I would like the thank the anonymous reviewers for their useful feedback.

## REFERENCES

[1] BOGACHEV, V. I. (1998). *Gaussian Measures*. *Mathematical Surveys and Monographs* **62**. Amer. Math. Soc., Providence, RI. MR1642391

[2] BONAMI, A. (1970). Étude des coefficients de Fourier des fonctions de $L^p(G)$. *Ann. Inst. Fourier* (*Grenoble*) **20** 335–402 (1971). MR0283496

[3] CARBERY, A. and WRIGHT, J. (2001). Distributional and $L^q$ norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Math. Res. Lett.* **8** 233–248.

[4] DIAKONIKOLAS, I., HARSHA, P., KLIVANS, A., MEKA, R., RAGHAVENDRA, P., SERVEDIO, R. A. and TAN, L.-Y. (2010). Bounding the average sensitivity and noise sensitivity of polynomial threshold functions. In *Proceedings of the* 42*nd ACM Symposium on Theory of Computing* (*STOC*).

[5] DIAKONIKOLAS, I., RAGHAVENDRA, P., SERVEDIO, R. A. and TAN, L.-Y. (2014). Average sensitivity and noise sensitivity of polynomial threshold functions. *SIAM J. Comput.* **43** 231–253. MR3164562

[6] DIAKONIKOLAS, I., SERVEDIO, R., TAN, L.-Y. and WAN, A. (2010). A Regularity Lemma, and Low-Weight Approximators, for Low-Degree Polynomial Threshold Functions. In 25*th Conference on Computational Complexity* (*CCC*).

[7] FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications*. *Vol. II*, 2nd ed. Wiley, New York. MR0270403

[8] FULTON, W. (1998). *Intersection Theory*, 2nd ed. Springer, Berlin. MR1644323

[9] GOTSMAN, C. and LINIAL, N. (1994). Spectral properties of threshold functions. *Combinatorica* **14** 35–50. MR1273199

[10] GREEN, B. and TAO, T. (2009). The distribution of polynomials over finite fields, with applications to the Gowers norms. *Contrib. Discrete Math.* **4** 1–36. MR2592422

[11] HARSHA, P., KLIVANS, A. and MEKA, R. (2014). Bounding the sensitivity of polynomial threshold functions. *Theory Comput.* **10** 1–26. MR3206267

[12] KANE, D. M. (2010). The Gaussian surface area and noise sensitivity of degree-d polynomial threshold functions. In 25*th Annual IEEE Conference on Computational Complexity—CCC* 2010 205–210. IEEE Computer Soc., Los Alamitos, CA. MR2932357

[13] KANE, D. M. (2011). A small PRG for polynomial threshold functions of Gaussians. In 2011 *IEEE* 52*nd Annual Symposium on Foundations of Computer Science—FOCS* 2011 257–266. IEEE Computer Soc., Los Alamitos, CA. MR2932701

[14] KANE, D. M. (2013). The correct exponent for the Gotsman–Linial conjecture. In 2013 *IEEE Conference on Computational Complexity—CCC* 2013 56–64. IEEE Computer Soc., Los Alamitos, CA. MR3306974

[15] KAUFMAN, T. and LOVETT, S. (2008). Worst Case to Average Case Reductions for Polynomials, In *The* 49*th Annual IEEE Symposium on Foundations of Computer Science* (*FOCS* 2008).

[16] LINDEBERG, J. W. (1922). Eine neue herleitung des exponential-gesetzes in der wahrscheinlichkeitsrechnung. *Math. Z.* **15** 211–235.

[17] MEKA, R. and ZUCKERMAN, D. (2010). Pseudorandom generators for polynomial threshold functions. In *STOC'*10*—Proceedings of the* 2010 *ACM International Symposium on Theory of Computing* 427–436. ACM, New York. MR2743291

[18] MOSSEL, E., O'DONNELL, R. and OLESZKIEWICZ, K. (2010). Noise stability of functions with low influences: Invariance and optimality. *Ann. of Math.* (2) **171** 295–341.

[19] NELSON, E. (1973). The free Markoff field. *J. Funct. Anal.* **12** 211–227. MR0343816

[20] PALEY, R. E. A. C. and ZYGMUND, A. (1932). A note on analytic functions in the unit circle. *Math. Proc. Cambridge Philos. Soc.* **28** 266–272.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
AND
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, SAN DIEGO
9500 GILMAN DRIVE # 0112
LA JOLLA, CALIFORNIA 92093
USA
E-MAIL: dankane@math.stanford.edu