Bounding the Average Sensitivity and Noise Sensitivity of Polynomial Threshold Functions

Ilias Diakonikolas
Columbia University
ilias@cs.columbia.edu

Prahladh Harsha
Tata Institute of Fundamental Research
prahladh@tifr.res.in

Adam Klivans
University of Texas at Austin
klivans@cs.utexas.edu

Raghu Meka
University of Texas at Austin
raghu@cs.utexas.edu

Prasad Raghavendra
Microsoft Research, New England
prasad@cs.washington.edu

Rocco A. Servedio
Columbia University
rocco@cs.columbia.edu

Li-Yang Tan
Columbia University
liyang@cs.columbia.edu

ABSTRACT

We give the first non-trivial upper bounds on the average sensitivity and noise sensitivity of degree-d polynomial threshold functions (PTFs). These bounds hold both for PTFs over the Boolean hypercube $\{-1,1\}^n$ and for PTFs over $\mathbb{R}^n$ under the standard n-dimensional Gaussian distribution $\mathcal{N}(0, I_n)$. Our bound on the Boolean average sensitivity of PTFs represents progress towards the resolution of a conjecture of Gotsman and Linial [17], which states that the symmetric function slicing the middle d layers of the Boolean hypercube has the highest average sensitivity of all degree-d PTFs. Via the L_1 polynomial regression algorithm of Kalai et al. [22], our bounds on Gaussian and Boolean noise sensitivity yield polynomial-time agnostic learning algorithms for the broad class of constant-degree PTFs under these input distributions.

The main ingredients used to obtain our bounds on both average and noise sensitivity of PTFs in the Gaussian setting are tail bounds and anti-concentration bounds on low-degree polynomials in Gaussian random variables [20, 7]. To obtain our bound on the Boolean average sensitivity of PTFs, we generalize the "critical-index" machinery of [37] (which in that work applies to halfspaces, i.e. degree-1 PTFs) to general PTFs. Together with the "invariance principle" of [30], this lets us extend our techniques from the Gaussian setting to the Boolean setting. Our bound on Boolean noise sensitivity is achieved via a simple reduction from upper bounds on average sensitivity of Boolean PTFs to corresponding bounds on noise sensitivity.

STOC'10, June 5–8, 2010, Cambridge, Massachusetts, USA.
Copyright 2010 ACM 978-1-4503-0050-6/10/06 ...$10.00.

Categories and Subject Descriptors

F.2.2 [**Nonnumerical Algorithms and Problems**]: Computations on Discrete Structures; I.2.6 [**Learning**]: [Concept Learning]

General Terms

Theory

Keywords

Boolean function; Fourier analysis; Average Sensitivity; Noise Sensitivity; Polynomial Threshold Function

1. INTRODUCTION

A degree-d polynomial threshold function (PTF) over a domain $X \subseteq \mathbb{R}^n$ is a Boolean-valued function $f : X \to \{-1,+1\}$,

$$f(x) = \text{sign}(p(x_1, \dots, x_n))$$

where $p : X \to \mathbb{R}$ is a degree-d polynomial with real coefficients. When $d = 1$ polynomial threshold functions are simply linear threshold functions (also known as halfspaces or LTFs), which play an important role in complexity theory, learning theory, and other fields such as voting theory. Low-degree PTFs (where d is greater than 1 but is not too large) are a natural generalization of LTFs which are also of significant interest in these fields.

Over more than twenty years much research effort in the study of Boolean functions has been devoted to different notions of the "sensitivity" of a Boolean function to small perturbations of its input, see e.g. [21, 6, 5, 14, 2, 38, 29, 30, 32, 33] and many other works. In this work we focus on two natural and well-studied measures of this sensitivity, the "average sensitivity" and the "noise sensitivity." As our main results, we give the first non-trivial upper bounds on average sensitivity and noise sensitivity of low-degree PTFs. These bounds have several applications in learning theory and complexity theory as we describe later in this introduction.

We now define the notions of average and noise sensitivity in the setting of Boolean functions $f : \{-1,1\}^n \to \{-1,1\}$. (Our paper also deals with average sensitivity and noise sensitivity of functions $f : \mathbb{R}^n \to \{-1,1\}$ under the Gaussian distribution, but the precise definitions are more involved than in the Boolean case so we defer them until later.)

1.1 Average Sensitivity and Noise Sensitivity

The *sensitivity* of a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ on an input $x \in \{-1,1\}^n$, denoted $s_f(x)$, is the number of Hamming neighbors $y \in \{-1,1\}^n$ of x (i.e. strings which differ from x in precisely one coordinate) for which $f(x) \neq f(y)$. The *average sensitivity* of f, denoted $\mathrm{AS}(f)$, is simply $\mathbf{E}[s_f(x)]$ (where the expectation is with respect to the uniform distribution over $\{-1,1\}^n$). An alternate definition of average sensitivity can be given in terms of the influence of individual coordinates on f. For a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ and a coordinate index $i \in [n]$, the *influence of coordinate* i *on* f is the probability that flipping the i-th bit of a uniform random input $x \in \{-1,1\}^n$ causes the value of f to change, i.e. $\mathrm{Inf}_i(f) = \mathbf{Pr}[f(x) \neq f(x^{\oplus i})]$ (where the probability is with respect to the uniform distribution over $\{-1,1\}^n$). The sum of all n coordinate influences, $\sum_{i=1}^n \mathrm{Inf}_i(f)$, is called the *total influence* of f; it is easily seen to equal $\mathrm{AS}(f)$. Bounds on average sensitivity have been of use in the structural analysis of Boolean functions (see e.g. [21, 14, 38]) and in developing computationally efficient learning algorithms (see e.g. [6, 33]).

The average sensitivity is a measure of how f changes when a single coordinate is perturbed. In contrast, the noise sensitivity of f measures how f changes when a random collection of coordinates are all perturbed simultaneously. More precisely, given a noise parameter $0 \leq \epsilon \leq 1$ and a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$, the *noise sensitivity of* f *at noise rate* ϵ is defined to be

$$\mathrm{NS}_\epsilon(f) = \mathbf{Pr}_{x,y}[f(x) \neq f(y)]$$

where x is uniform from $\{-1,1\}^n$ and y is obtained from x by flipping each bit independently with probability ϵ. Noise sensitivity has been studied in a range of contexts including Boolean function analysis, percolation theory, and computational learning theory [2, 25, 29, 36, 26].

1.2 Main Results: Upper Bounds on Average Sensitivity and Noise Sensitivity

1.2.1 Boolean PTFs

In 1994 Gotsman and Linial [17] conjectured that the symmetric function slicing the middle d layers of the Boolean hypercube has the highest average sensitivity among all degree-d PTFs. Since this function has average sensitivity $\Theta(d\sqrt{n})$ for every $1 \leq d \leq \sqrt{n}$, this conjecture implies (and is nearly equivalent to) the conjecture that every degree-d PTF f over $\{-1,1\}^n$ has $\mathrm{AS}(f) \leq d\sqrt{n}$.

Our first main result is an upper bound on average sensitivity which makes progress toward this conjecture:

THEOREM 1.1. *For any degree-d PTF f over $\{-1,1\}^n$, we have* $\mathrm{AS}(f) \leq 2^{O(d)} \cdot \log n \cdot n^{1-1/(4d+2)}$.

Using a completely different set of techniques, we also prove a different bound which improves on Theorem 1.1 for $d \leq 4$:

THEOREM 1.2. *For any degree-d PTF f over $\{-1,1\}^n$, we have* $\mathrm{AS}(f) \leq 2n^{1-1/2^d}$.

We give a simple reduction which translates any upper bound on average sensitivity for degree-d PTFs over Boolean variables into a corresponding upper bound on their noise sensitivity. Combining this reduction with Theorems 1.1 and 1.2, we establish:

THEOREM 1.3. *For any degree-d PTF f over $\{-1,1\}^n$ and any $0 \leq \epsilon \leq 1$, we have*

$$\begin{aligned} \mathrm{NS}_\epsilon(f) &\leq 2^{O(d)} \cdot \epsilon^{1/(4d+2)} \log(1/\epsilon) \\ \mathrm{NS}_\epsilon(f) &\leq O(\epsilon^{1/2^d}). \end{aligned}$$

1.2.2 Gaussian PTFs

Looking beyond the Boolean hypercube, there are well-studied notions of average sensitivity and noise sensitivity for Boolean-valued functions over $\mathbb{R}^n$, where we view $\mathbb{R}^n$ as endowed with the standard multivariate Gaussian distribution $\mathcal{N}(0, I_n)$ [4, 30]. Given $f : \mathbb{R}^n \to \mathbb{R}$ that is square-integrable under the Gaussian measure $\mathcal{N}(0,1)$ and $i \in [n]$, the *Gaussian influence of co-ordinate* i *on* f is defined to be $\mathrm{GI}_i(f) = \mathbf{E}_{x_{-i} \sim \mathcal{N}^{n-1}}[\mathrm{Var}_{x_i \sim \mathcal{N}}[f]]$where x_{-i} denotes all but the i^{th} coordinate of x. The *Gaussian average sensitivity of* f is defined as $\mathrm{GAS}(f) = \sum_{i \in [n]} \mathrm{GI}_i(f)$. The *Gaussian noise sensitivity of* f *at noise rate* $\epsilon \in [0,1]$ is defined to be $\mathrm{GNS}_\epsilon(f) = \mathbf{Pr}_{x,z}[f(x) \neq f(y)]$ where $x \sim \mathcal{N}^n$ and $y \overset{\text{def}}{=} (1-\epsilon)\,x + \sqrt{2\epsilon - \epsilon^2}\,z$ for an independent Gaussian *noise vector* $z \sim \mathcal{N}^n$. These are natural analogues of their uniform-distribution Boolean hypercube counterparts defined above.) We prove upper bounds on Gaussian average sensitivity and Gaussian noise sensitivity of low-degree PTFs:

THEOREM 1.4. *For any degree-d PTF f over $\mathbb{R}^n$, we have* $\mathrm{GAS}(f) \leq O(d^2 \cdot \log n \cdot n^{1-1/2d})$.

THEOREM 1.5. *For any degree-d PTF f over $\mathbb{R}^n$ and any $0 \leq \epsilon \leq 1$, we have* $\mathrm{GNS}_\epsilon(f) \leq O(d \cdot \log^{1/2}(1/\epsilon) \cdot \epsilon^{1/2d})$.

We note that in subsequent work D. Kane [24] has given an optimal upper bound $\mathrm{GNS}_\epsilon(f) \leq O(d\sqrt{\epsilon})$ on the Gaussian noise sensitivity of any degree-d PTF.

1.3 Application: agnostically learning constant-degree PTFs in polynomial time

Our bounds on noise sensitivity, together with machinery developed in [25, 22, 26], yield the first efficient agnostic learning algorithms for low-degree polynomial threshold functions. In this section we state our new learning results; details are given in the full version.

We begin by briefly reviewing the fixed-distribution agnostic learning framework that has been studied in several recent works, see e.g. [22, 26, 3, 15, 23, 39]. Let $\mathcal{D}_X$ be a (fixed, known) distribution over an example space X such as the uniform distribution over $\{-1,1\}^n$ or the standard multivariate Gaussian distribution $\mathcal{N}(0, I_n)$ over $\mathbb{R}^n$. Let $\mathcal{C}$ denote a class of Boolean functions, such as the class of all degree-d PTFs. An algorithm A is said to be an *agnostic learning algorithm for* $\mathcal{C}$ *under distribution* $\mathcal{D}_X$ if it has the following property: Let $\mathcal{D}$ be any distribution over $X \times \{-1,1\}$ such that the marginal of $\mathcal{D}$ over X is $\mathcal{D}_X$. Then if A is run on a sample of labeled examples drawn independently from $\mathcal{D}$, with high probability A outputs a hypothesis $h : X \to \{-1,1\}$ such that $\mathbf{Pr}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] \leq \mathsf{opt} + \epsilon$, where $\mathsf{opt} = \min_{f \in \mathcal{C}} \mathbf{Pr}_{(x,y)\sim\mathcal{D}}[f(x) \neq y]$. In words, A's hypothesis is nearly as accurate as the best hypothesis in the class $\mathcal{C}$.

Kalai et al. [22] gave an agnostic learning algorithm based on L_1 polynomial regression. More precisely, they showed that for a class $\mathcal{C}$ of functions and a distribution $\mathcal{D}$, if every function in $\mathcal{C}$ has a low-degree polynomial approximator (in the L_2 norm) under the marginal distribution $\mathcal{D}_X$, then the L_1 polynomial regression algorithm is an efficient agnostic learning algorithm for $\mathcal{C}$ under $\mathcal{D}_X$. Together with the existence of low-degree polynomial approximators for halfspaces (under the uniform distribution on $\{-1,1\}^n$ and the standard Gaussian distribution $\mathcal{N}(0, I_n)$ on $\mathbb{R}^n$), the L_1 polynomial regression algorithm yields a $n^{O(1/\epsilon^4)}$-time agnostic learning algorithm for halfspaces under these distributions.

Using ingredients from [25], upper bounds on Boolean noise sensitivity (such as Theorem 1.3) imply the existence of low-degree L_2-norm polynomial approximators under the uniform distribution on $\{-1,1\}^n$. Hence we obtain the following agnostic learning result:

Theorem 1.6. *The class of degree-d PTFs is agnostically learnable under the uniform distribution on $\{-1,1\}^n$ in time $n^{2^{O(d^2)}(\log 1/\epsilon)^{4d+2}/\epsilon^{8d+4}}$. For $d \leq 4$, this bound can be improved to $n^{O(1/\epsilon^{2^{d+1}})}$.*

Similarly, using ingredients from [26], upper bounds on Gaussian noise sensitivity (such as Theorem 1.5) imply the existence of low-degree L_2-norm polynomial approximators under $\mathcal{N}(0, I_n)$. This lets us obtain

Theorem 1.7. *The class of degree-d PTFs is agnostically learnable under any n-dimensional Gaussian distribution in time $n^{(d/\epsilon)^{O(d)}}$.*

For ϵ constant, these results are the first polynomial-time agnostic learning algorithms for constant-degree PTFs.

1.4 Other applications

The results and approaches of this paper have found other recent applications beyond the agnostic learning results presented above; we describe two of these below.

Gopalan and Servedio [16] have combined the average sensitivity bound given by Theorem 1.1 with techniques from [27] to give the first sub-exponential time algorithms for learning AC^0 circuits augmented with a small (but superconstant) number of arbitrary threshold gates, i.e. gates that compute arbitrary LTFs which may have weights of any magnitude. (Previous work using different techniques [19] could only handle AC^0 circuits augmented with majority gates.)

In other recent work Diakonikolas et al. [10] and Harsha et al. [18] have refined the approach used to prove Theorem 1.1 to establish a "regularity lemma" for low-degree polynomial threshold functions. Roughly speaking, this lemma says that any degree-d PTF can be decomposed into a constant number of subfunctions, almost all of which are "regular" degree-d PTFs. [10] apply this regularity lemma to extend the positive results on the existence of low-weight approximators for LTFs, proved in [37], to low-degree PTFs.

1.5 Techniques

In this section we give a high-level overview of how Theorems 1.1, 1.4 and 1.5 are proved. (As mentioned earlier, Theorem 1.2 is proved using completely different techniques; see Section 5.) The arguments are simpler for the Gaussian setting so we begin with these.

1.5.1 The Gaussian case

We sketch the argument for the Gaussian average sensitivity bound Theorem 1.4; the Gaussian noise sensitivity bound Theorem 1.5 follows along similar lines.

Let $f = \operatorname{sign}(p)$ where $p : \mathbb{R}^n \to \mathbb{R}$ is a degree-d polynomial. The Gaussian average sensitivity $\mathrm{GAS}(f)$ of f is equal to the sum of individual Gaussian influences $\mathrm{GI}_i(f) = 2\,\mathbf{Pr}_{x,x^i}[f(x) \neq f(x^i)]$, where $x \sim \mathcal{N}^n$ and x^i is obtained by replacing the i^{th} coordinate of x by an independent random sample from $\mathcal{N}$. Central to the proof of Theorem 1.4 is a bound on $\mathrm{GI}_i(f)$ by $\mathrm{GI}_i(p)$, the influence of variable i in the polynomial p. Let $i = 1$, and express $p(x)$ as a univariate polynomial in x_1 as follows: $p(x) = p(x_1, \ldots, x_n) = \sum_{i=0}^{d} p_i(x_2, \ldots, x_n) \cdot h_i(x_1)$, where $h_i(x_1)$ is the univariate degree-i Hermite polynomial. Intuitively, the event $f(x) \neq f(x^1)$ can only take place if either:

- $|p_0(g_2, \ldots, g_n)|$ is "small", or
- $|p_i(g_2, \ldots, g_n)|$ is "large" for some $i \in [d]$.

We use an anti-concentration result for polynomials in Gaussian random variables, due to Carbery and Wright [7], to show that $|p_0(g_2, \ldots, g_n)|$ is "small" only with low probability. For the second bullet, we apply tail bounds for low-degree polynomials in independent Gaussian random variables [20] to show, for each $i \in [d]$, that $|p_i(g_2, \ldots, g_n)|$ is "large" only with low probability. We can thus argue that $\mathbf{Pr}_{x,x^1}[f(x) \neq f(x^1)]$ is low, bounding the Gaussian influence of variable 1 on f in terms of $\mathrm{GI}_1(p)$. By normalizing $\mathrm{Var}[p] = 1$ and applying a convexity argument, we see that $\mathrm{GI}(f)$ is maximized when $\mathrm{GI}_i(p) = d/n$ for each $i \in [n]$, establishing Theorem 1.4.

1.5.2 The Boolean case

One advantage of working over the Boolean domain $\{-1,1\}^n$ is that without loss of generality we may consider only *multilinear* PTFs, where $f = \operatorname{sign}(p(x))$ for p a multilinear polynomial. However, this advantage is offset by the fact that the uniform distribution on $\{-1,1\}^n$ is less symmetric than the Gaussian distribution; for example, every degree-1 PTF under the Gaussian distribution $\mathcal{N}^n$ is equivalent simply to $\operatorname{sign}(x_1 - \theta)$, but this is of course not true for degree-1 PTFs over $\{-1,1\}^n$. Our upper bound on Boolean average sensitivity uses ideas from the Gaussian setting but also requires significant additional ingredients.

An important notion in the Boolean case is that of a "regular" PTF; this is a PTF $f = \operatorname{sign}(p)$ where every variable *in the polynomial p* has low influence. (See Section 2 for a definition of the influence of a variable on a real-valued function; note that the definition from Section 1.1 applies only for Boolean-valued functions.) If f is a regular PTF, then the "invariance principle" of [30] tells us that $p(x)$ (where x is uniform from $\{-1,1\}^n$) behaves much like $p(\mathcal{G})$ (where $\mathcal{G}$ is drawn from $\mathcal{N}(0, I_n)$), and essentially the arguments from the Gaussian case can be used.

It remains to handle the case where f is not a regular PTF, i.e. some variable has high influence in p. To accomplish this, we generalize the notion of the "critical-index" of a halfspace (see [37, 11]) to apply to PTFs. We show that a carefully chosen random restriction (one which fixes only the variables up to the critical index – very roughly speaking, only the highest-influence variables – and leaves the other ones free) has non-negligible probability of causing f to collapse down to a regular PTF. This lets us give a recursive bound on average sensitivity which ends up being not much worse than the bound that can be obtained for the regular case; see Section 4.2 for a detailed explanation of the recursive argument.

1.6 Organization

Due to space constraints, this proceedings version contains only a selection of our results with high-level arguments. Full proofs are provided in [8, 18].

Formal definitions of average sensitivity and noise sensitivity (especially in the Gaussian case), along with mathematical tools we use such as tail bounds and anticoncentration results for low degree polynomials, are presented in Section 2.

In Section 3, we outline the proof of an upper bound on the Gaussian average sensitivity of PTFs (Theorem 1.4); the missing details can be found in the full version. The main result of the paper – a bound on the Boolean average sensitivity (Theorem 1.1) – is outlined in Section 4 (again details are in the full version). In Section 5, an alternate bound is established for Boolean average sensitivity that is better than Theorem 1.1 for degrees $d \leq 4$ (Theorem 1.2). This is followed by a reduction from Boolean average sensitivity bounds to corresponding noise sensitivity bounds (Theorem 6.1) in Section 6.

2. DEFINITIONS AND BACKGROUND

2.1 Basic Definitions

In this subsection we record the basic notation and definitions used throughout the paper. For $n \in \mathbb{N}$, we denote by $[n]$ the set $\{1, 2, \ldots, n\}$. We write $\mathcal{N}$ to denote the standard univariate Gaussian distribution $\mathcal{N}(0, 1)$.

For a degree-d polynomial $p : X \to \mathbb{R}$ we denote by $\|p\|_2$ its l_2 norm, $\|p\|_2 = \mathbf{E}_x[p(x)^2]^{1/2}$, where the intended distribution over $x \in \mathbb{R}^n$ (which will always be either uniform over $\{-1,1\}^n$, or the $\mathcal{N}^n$ distribution) will always be clear from context. We note that for multilinear p the two notions are always equal (see e.g. Proposition 3.5 of [30]).

We now proceed to define the notion of influence for real-valued functions in a product probability space. Throughout this paper we consider either the uniform distribution on the hypercube $\{\pm 1\}^n$ or the standard n-dimensional Gaussian distribution in $\mathbb{R}^n$. However, for the sake of generality, we adopt this more general setting.

Let $(\Omega_1, \mu_1), \ldots, (\Omega_n, \mu_n)$ be probability spaces and let $(\Omega = \otimes_{i=1}^n \Omega_i, \mu = \otimes_{i=1}^n \mu_i)$ denote the corresponding product space. Let $f : \Omega \to \mathbb{R}$ be any square integrable function on (Ω, μ), i.e. $f \in L^2(\Omega, \mu)$. The influence of the ith coordinate on f [30] is

$$\mathrm{Inf}_i^{\mu}(f) \stackrel{\mathrm{def}}{=} \mathbf{E}_{\mu}[\mathrm{Var}_{\mu_i}[f]]$$

and the total influence of f is $\mathrm{Inf}^{\mu}(f) \stackrel{\mathrm{def}}{=} \sum_{i=1}^n \mathrm{Inf}_i^{\mu}(f)$.

For a function $f : \{-1,1\}^n \to \mathbb{R}$ over the Boolean hypercube endowed with the uniform distribution, the influence of variable i on f can be expressed in terms of the Fourier coefficients of f as $\mathrm{Inf}_i(f) = \sum_{S \ni i} \widehat{f}(S)^2$, and as mentioned in the introduction it is easily seen that $\mathrm{AS}(f) = \mathrm{Inf}(f)$ for Boolean-valued functions $f : \{-1,1\}^n \to \{-1,1\}$.

In this paper we are concerned with variable influences for functions defined over $\{-1,1\}^n$ under the uniform distribution, and over $\mathbb{R}^n$ under $\mathcal{N}(0, I_n)$; we shall adopt the convention that $\mathrm{Inf}_i(f)$ denotes the former and $\mathrm{GI}_i(f)$ the latter. We also denote by $\mathrm{GAS}(f) = \sum_{i \in [n]} \mathrm{GI}_i(f)$ the Gaussian average sensitivity.

Note that for a function $f : \mathbb{R}^n \to \{-1,1\}$, the Gaussian influence $\mathrm{GI}_i(f)$ can be equivalently written as: $\mathrm{GI}_i(f) = 2\,\mathbf{Pr}_{x,x^i}[f(x) \neq f(x^i)]$, where $x \sim \mathcal{N}^n$ and x^i is obtained by replacing the i^{th} coordinate of x by an independent random sample from $\mathcal{N}$.

Fourier and Hermite Analysis. We assume familiarity with the basics of Fourier analysis over the Boolean hypercube $\{-1,1\}^n$. We will also require similar basics of Hermite analysis over the space $\mathbb{R}^n$ equipped with the standard n-dimensional Gaussian distribution $\mathcal{N}^n$; a brief review is provided in the full version.

2.2 Probabilistic Facts

In this subsection, we record the basic probabilistic tools we use in our proofs.

We first recall the following well-known consequence of hypercontractivity (see e.g. Lecture 16 of [31] for the boolean setting and [4] for the Gaussian setting):

THEOREM 2.1. *Let $p : X \to \mathbb{R}$ be a degree-d polynomial, where X is either $\{-1,1\}^n$ under the uniform distribution or $\mathbb{R}^n$ under $\mathcal{N}^n$, and fix $q > 2$. Then*

$$\|p\|_q^2 \leq (q-1)^d \|p\|_2^2.$$

We will need a concentration bound for low-degree polynomials over independent random signs or standard Gaussians. It can be proved (in both cases) using Markov's inequality and hypercontractivity, see e.g. [20, 31, 1].

THEOREM 2.2 ("DEGREE-d CHERNOFF BOUND"). *Let $p(x)$ be a degree-d polynomial. Let x be drawn either from the uniform distribution in $\{-1,1\}^n$ or from $\mathcal{N}^n$. For any $t > e^d$, we have*

$$\mathbf{Pr}_x[|p(x)| \geq t\|p\|_2] \leq \exp(-\Omega(t^{2/d})).$$

The second fact is a powerful anti-concentration bound for low-degree polynomials over Gaussian random variables. (We note that this result does not hold in the Boolean setting.)

THEOREM 2.3 ([7]). *Let $0 \neq p : \mathbb{R}^n \to \mathbb{R}$ be a degree-d polynomial. Then for all $\epsilon > 0$, we have*

$$\mathbf{Pr}_{x \sim \mathcal{N}^n}[|p(x)| \leq \epsilon\|p\|_2] \leq O(d\epsilon^{1/d}).$$

We also make essential use of a (weak) anti-concentration property of low-degree polynomials over the hypercube $\{-1,1\}^n$:

THEOREM 2.4 ([12, 1]). *Let $p : \{-1,1\}^n \to \mathbb{R}$ be a degree-d polynomial with $\mathrm{Var}[p] \equiv \sum_{0<|S|\leq d} \widehat{p}(S)^2 = 1$ and $\mathbf{E}[p] = \widehat{p}(\emptyset) = 0$. Then we have*

$$\mathbf{Pr}[p(x) > 1/2^{O(d)}] > 1/2^{O(d)}$$

and hence,

$$\mathbf{Pr}[|p(x)| \geq 1/2^{O(d)}] > 1/2^{O(d)}.$$

The following is a restatement of the invariance principle, specifically Theorem 3.19 under hypothesis **H4** in [30].

THEOREM 2.5 ([30]). *Let $p(x) = \sum_{|S| \le d} \widehat{p}(S) x_S$ be a degree-d multilinear polynomial with $\sum_{0<|S|\le d} \widehat{p}(S)^2 = 1$. Suppose each variable $i \in [n]$ has low influence $\mathrm{Inf}_i(p) \le \tau$, i.e. $\sum_{S \ni i} \widehat{p}(S)^2 \le \tau$. Let x be drawn uniformly from $\{-1,1\}^n$ and $\mathcal{G} \sim \mathcal{N}^n$. Then,*

$$\sup_{t \in \mathbb{R}} |\, \mathbf{Pr}[p(x) \le t] - \mathbf{Pr}[p(\mathcal{G}) \le t]| \le O(d\tau^{1/(4d+1)}).$$

3. GAUSSIAN AVERAGE SENSITIVITY

The following lemma, which relates the influence of a variable on $f = \mathrm{sign}(p)$ to its influence on the polynomial p, is central to the proof of Theorem 1.4.

LEMMA 3.1. *Let $p : \mathbb{R}^n \to \mathbb{R}$ be a degree-d polynomial over Gaussian inputs with $\|p\|_2 = 1$ and let $f = \mathrm{sign}(p)$. Then for each $i \in [n]$, we have $\mathrm{GI}_i(f) \le O(d^2 \cdot \mathrm{GI}_i(p)^{1/(2d)} \cdot \log(1/\mathrm{GI}_i(p)))$.*

PROOF OF LEMMA 3.1. Let $p(x)$ be a degree-d polynomial with $\|p\|_2 = 1$. For notational convenience let us fix $i = 1$ and let $\tau = \mathrm{GI}_1(p)$. We may assume that $\tau < 1/4$ since otherwise the claimed bound holds trivially. We express $p(x)$ as a univariate polynomial in x_1 as follows,

$$p(x) = p(x_1, \dots, x_n) = \sum_{i=0}^{d} p_i(x_2, \dots, x_n) \cdot h_i(x_1)$$

where $h_i(x_1)$ is the univariate degree-i Hermite polynomial. Note that for any multi-index $S = (S_2, \dots, S_n) \in \mathbb{N}^{n-1}$ and $0 \le i \le d$, we have $\widehat{p_i}(S) = \widehat{p}(S')$ where $S' = (i, S_2, \dots, S_n) \in \mathbb{N}^n$. As a result, using Parseval's identity for the Hermite basis, we have that $\|p\|^2 = \sum_{i=0}^{d} \|p_i\|^2$.

We further have

$$\|p_i\|^2 = \sum_{S \in \mathbb{N}^{n-1}} \widehat{p_i}(S)^2 \quad \text{and} \quad \mathrm{GI}_i(p) = \sum_{S : S_i > 0} \widehat{p}(S)^2.$$

Consequently the 2-norms of $p_1, \dots, p_d$ are "small" and the 2-norm of p_0 is "large":

$$\sum_{i=1}^{d} \|p_i\|^2 = \sum_{S : S_1 > 0} \widehat{p}(S)^2 = \tau, \text{ and } \|p_0\|^2 = 1 - \tau > 1/2.$$

Let $t = C^{d/2} \tau^{1/2} \log^{d/2}(1/\tau)$ and $\gamma = d^2 \cdot \tau^{1/2d} \log(1/\tau)$ where C is an absolute constant. We can assume that $\gamma < 1/10$ since otherwise the bound of Lemma 3.1 holds trivially. For these values of t and γ, the proof strategy is as follows:

- We use the anti-concentration bound Theorem 2.3 to argue that with high probability $p_0(g_2, \dots, g_n)$ is not too small: more precisely, $\mathbf{Pr}_{\mathcal{N}^{n-1}}[|p_0(g_2, \dots, g_n)| \le td(2ed\log(1/\gamma))^{d/2}] \le O(\gamma)$.
- We use the "degree-d Chernoff bound" (Theorem 2.2) to argue that with high probability each $p_i(g_2, \dots, g_n)$, $i \in [d]$, is not too large: more precisely, $\mathbf{Pr}_{N^{n-1}}[|p_i(g_2, \dots, g_n)| \ge t] \le O(\gamma)$.
- We use elementary properties of the $\mathcal{N}(0,1)$ distribution to argue that if $|a| \ge td(2ed\log(1/\gamma))^{d/2}$ and $|b_i| \le t$, then the function $\mathrm{sign}(a + \sum_{i=1}^{d} b_i h_i(g_1))$ (a function of one $\mathcal{N}(0,1)$ random variable g_1) is $O(\gamma)$-close to the constant function $\mathrm{sign}(a)$.
- Thus with probability at least $1 - O(\gamma)$ over the choice of $g_2, \dots, g_n$, we have $\mathrm{Var}_{g_1}[\mathrm{sign}(p(g_1, \dots, g_n))] \le O(\gamma(1-\gamma)) \le O(\gamma)$. For the remaining (at most) $O(\gamma)$ fraction of outcomes for $g_2, \dots, g_n$ we always have $\mathrm{Var}_{g_1}[\mathrm{sign}(p(g_1, \dots, g_n))] \le 1$, so overall we get $\mathrm{GI}_1(\mathrm{sign}(p)) \le O(\gamma)$.

Proofs of the three aforementioned claims are presented in the full version. □

We now sketch the proof of Theorem 1.4 using Lemma 3.1. The idea is simple: by normalizing we may assume that $\mathrm{Var}[p] = 1$, and consequently the total influence $\mathrm{GI}(p) = \sum_{i=1}^{n} \mathrm{GI}_i(p)$ of p is at most d since p is a degree-d polynomial. Intuitively, Lemma 3.1 tells us that the largest possible total value of $\mathrm{GI}(f)$ is obtained if each $\mathrm{GI}_i(p)$ equals d/n; a convexity argument, presented in the full version, makes this precise and establishes Theorem 1.4.

We remark that in the case of degree-d *multilinear* PTFs it is possible to obtain a slightly stronger bound of $\mathrm{GAS}(f) \le O(d \cdot \log n \cdot n^{1-1/2d})$ using our approach; we omit the details.

4. BOOLEAN AVERAGE SENSITIVITY

Let $\mathrm{AS}(n, d)$ denote the maximum possible average sensitivity of any degree-d PTF over n Boolean variables. In this section we outline the proof of the claimed bound in Theorem 1.1:

$$\mathrm{AS}(n, d) \le 2^{O(d)} \cdot \log n \cdot n^{1-1/(4d+2)}. \tag{1}$$

For $d = 1$ (linear threshold functions) it is well known that $\mathrm{AS}(n, 1) = 2^{-n}\binom{n}{n/2} = \Theta(\sqrt{n})$. Also, notice that the RHS of (1) is larger than n for $d = \omega(\sqrt{\log n})$, yielding a trivial bound of $\mathrm{AS}(n, d) \le n$. Therefore throughout this section we shall assume d satisfies $2 \le d \le O(\sqrt{\log n})$.

4.1 Regularity and the critical index of polynomials

The proof of Theorem 1.1 is a combination of a case analysis and a recursive bound. The case analysis is based on the notion of critical index for polynomials that we define below.

In [37] a notion of the "critical index" of a linear form was defined and subsequently used in [34, 9, 11]. This notion is generalized below to the case of polynomials.

DEFINITION 1. *Let $0 \ne p : \{-1,1\}^n \to \mathbb{R}$ and $\tau > 0$. Assume the variables are ordered such that $\mathrm{Inf}_i(f) \ge \mathrm{Inf}_{i+1}(f)$ for all $i \in [n-1]$. The* τ-critical index *of f is the least i such that:*

$$\mathrm{Inf}_{i+1}(p) \le \tau \cdot \sum_{j=i+1}^{n} \mathrm{Inf}_j(p). \tag{2}$$

If (2) does not hold for any i we say that the τ-critical index of p is $+\infty$. If p is has τ-critical index 0, we say that p is τ-regular.

The following simple lemma will be useful for us. It says that the total influence $\sum_{i=j+1}^{n} \mathrm{Inf}_i(p)$ goes down exponentially as a function of j prior to the critical index:

LEMMA 4.1. *Let $p : \{-1,1\}^n \to \mathbb{R}$ and $\tau > 0$. Let k be the τ-critical index of p. For $0 \le j \le k$ we have $\sum_{i=j+1}^{n} \mathrm{Inf}_i(p) \le (1-\tau)^j \cdot \mathrm{Inf}(p)$.*

PROOF. The lemma trivially holds for $j = 0$. In general, since j is at most k, we have that $\mathrm{Inf}_j(p) \geq \tau \cdot \sum_{i=j}^{n} \mathrm{Inf}_i(p)$, or equivalently $\sum_{i=j+1}^{n} \mathrm{Inf}_i(p) \leq (1-\tau)\cdot\sum_{i=j}^{n} \mathrm{Inf}_i(p)$ which yields the claimed bound. □

4.2 Overview of proof

The high-level approach to proving Theorem 1.1 is a combination of a case analysis and a recursive bound.

By the invariance principle (see Section 2.2), the behaviour of τ-regular PTFs in the boolean and Gaussian settings is nearly the same. Therefore, we can argue directly that the average sensitivity is small using arguments similar to the Gaussian case (in fact simpler since now the polynomial is multilinear). In particular, we show the following lemma:

LEMMA 4.2. *Fix $\tau = n^{-\Theta(1)}$. Let f be a τ-regular degree-d PTF. Then,* $\mathrm{AS}(f) \leq O(d \cdot n \cdot \tau^{1/(4d+1)})$.

The following claim is a direct consequence of the above lemma.

CLAIM 4.3. *Suppose $f = \mathrm{sign}(p)$ is a τ-regular degree-d PTF where $\tau \overset{def}{=} n^{-(4d+1)/(4d+2)}$. Then,*

$$\mathrm{AS}(f) \leq O(d \cdot n^{1-1/(4d+2)}).$$

For PTFs that are not τ-regular, we show that there is a not-too-large value of k (at most $K \overset{\mathrm{def}}{=} 2d\log n/\tau$), and a collection of k variables (the variables whose influence in p are largest), such that the following holds: if we consider all 2^k subfunctions of f obtained by fixing the variables in all possible ways, a "large" (at least $1/2^{O(d)}$) fraction of the restricted functions have low average sensitivity. More precisely, we show:

CLAIM 4.4. *Let $K \overset{def}{=} 2d\log n/\tau$ where $\tau \overset{def}{=} n^{-(4d+1)/(4d+2)}$. Suppose $f = \mathrm{sign}(p)$ is a degree-d PTF that is not τ-regular. Then for some $1 \leq k \leq K$, there is a set of k variables with the following property: for at least a $1/2^{O(d)}$ fraction of all 2^k assignments ρ to those k variables, we have*

$$\mathrm{AS}(f_\rho) \leq O(d \cdot (\log n)^{1/4} \cdot n^{1-1/(4d+2)}).$$

The proof of Claim 4.4 is given in Section 4.3.3. We do this by generalizing the "critical index" case analysis from [37]. We define a notion of the τ-critical index of a degree-d polynomial; a τ-regular polynomial p is one for which the τ-critical index is 0. If the τ-critical index of p is some value $k \leq 2d\log n/\tau$, we restrict the k largest-influence variables (see Section 4.3.1). If the τ-critical index is larger than $2d\log n/\tau$, we restrict the $k = 2d\log n/\tau$ largest-influence variables in p (see Section 4.3.2).

4.2.1 Proof of main result (Theorem 1.1) assuming Claim 4.3 and Claim 4.4

Given these two claims it is not difficult to obtain the final result. In Claim 4.4, we note that the k restricted variables may each contribute at most 1 to the average sensitivity of f (recall that average sensitivity is equal to the sum of influences of each variable), and that the total influence of the remaining variables on f is equal to the expected average sensitivity of f_ρ, where the expectation is taken over all 2^k restrictions ρ. Since each function f_ρ is itself a degree-d PTF over at most n variables, we have the following recursive constraint on $\mathrm{AS}(n,d)$:

$$\mathrm{AS}(n,d) \leq \max\{O(d \cdot n^{1-1/(4d+2)}),$$

$$\max_{\substack{1\leq k\leq K\\ 1/2^{O(d)}\leq\alpha\leq 1}} \{k+\alpha\cdot O(d\cdot(\log n)^{1/4}\cdot n^{1-1/(4d+2)})+(1-\alpha)\mathrm{AS}(n,d)\}\}.$$

It is easy to see that the maximum possible value of $\mathrm{AS}(n,d)$ subject to the above constraint is at most the maximum possible value of $\mathrm{AS}'(n,d)$ that satisfies the following weaker constraint:

$$\mathrm{AS}'(n,d) \leq K + \left(1 - \frac{1}{2^{O(d)}}\right)\mathrm{AS}'(n,d)$$

which is satisfied by $\mathrm{AS}'(n,d) \leq 2^{O(d)} \cdot \log n \cdot n^{1-1/(4d+2)}$.

4.3 Non τ-regular PTFs

The proof of Claim 4.4 is divided into two cases based on the value of the critical index.

4.3.1 The small critical index case

Let $f = \mathrm{sign}(p)$ be such that the τ-critical index of p is some value k between 1 and $K = 2d\log n/\tau$. By definition, the sequence of influences $\mathrm{Inf}_{k+1}(p),\ldots,\mathrm{Inf}_n(p)$ is τ-regular. We essentially reduce this case to the regular case for a regularity parameter τ' somewhat larger than τ.

Consider a random restriction ρ of all the variables up to the critical index. We will show the following:

LEMMA 4.5. *For a $1/2^{O(d)}$ fraction of restrictions ρ, the sequence of influences $\mathrm{Inf}_{k+1}(p_\rho),\ldots,\mathrm{Inf}_n(p_\rho)$ is τ'-regular, where $\tau' \overset{def}{=} (3\log n)^d \cdot \tau$.*

By our choice of $\tau = n^{-(4d+1)/(4d+2)}$, we have that $\tau' = n^{-\Theta(1)}$, and so we may apply Lemma 4.2 to these restrictions to conclude that the associated PTFs have average sensitivity at most $O(d \cdot n \cdot (\tau')^{1/(4d+1)})$.

PROOF. Since the sequence of influences $\mathrm{Inf}_{k+1}(p),\ldots,\mathrm{Inf}_n(p)$ is τ-regular, we have

$$\mathrm{Inf}_i(p) \leq \tau \cdot \sum_{j=k+1}^{n} \mathrm{Inf}_j(p)$$

for all $i \in [k+1,n]$.

We want to prove that for a $1/2^{O(d)}$ fraction of all 2^k restrictions ρ to $x_1,\ldots,x_k$ we have

$$\mathrm{Inf}_i(p_\rho) \leq \tau' \cdot \sum_{j=k+1}^{n} \mathrm{Inf}_j(p_\rho) \qquad (3)$$

for all $i \in [k+1,n]$.

To do this we proceed as follows: First we show that for a low-degree polynomial, a random restriction with very high probability does not cause any variable's influence to increase by more than a polylog(n) factor. Formally, we prove the following lemma in the full version:

LEMMA 4.6. *Let $p(x_1,\ldots,x_n)$ be a degree-d polynomial. Let ρ be a randomly chosen assignment to the variables $x_1,\ldots,x_k$. Fix any $t > e^{2d}$ and any $\ell \in [k+1,n]$. With probability at least $1-\exp(-\Omega(t^{1/d}))$ over the choice of ρ, we have*

$$\mathrm{Inf}_\ell(p_\rho) \leq t \cdot 3^d \mathrm{Inf}_\ell(p).$$

In particular, for $t = \log^d n$, we have that with probability at least $1-n^{-\omega(1)}$, every variable $\ell \in [k+1,n]$ has $\mathrm{Inf}_\ell(p_\rho) \leq (3\log n)^d \cdot \mathrm{Inf}_\ell(p)$.

Lemma 4.6 implies that, with very high probability over the random restrictions, we have $\mathrm{Inf}_i(p_\rho) \le (3\log n)^d \cdot \mathrm{Inf}_i(p)$, for all $i \in [k+1, n]$. We need to show that for a $1/2^{O(d)}$ fraction of all restrictions the sum on the RHS of (3) is at least $\sum_{j=k+1}^{n} \mathrm{Inf}_j(p)$ (its expected value). The lemma then follows by a union bound.

We consider the degree-$2d$ polynomial $A(\rho_1, \ldots, \rho_k) \stackrel{\mathrm{def}}{=} \sum_{j=k+1}^{n} \mathrm{Inf}_j(p_\rho)$ in variables $\rho_1, \ldots, \rho_k$. The expected value of A is $\mathbf{E}_\rho[A] = \sum_{j=k+1}^{n} \mathrm{Inf}_j(p) = \widehat{A}(\emptyset)$. We apply Theorem 2.4 to the polynomial $B = (A - \widehat{A}(\emptyset))/\operatorname{Var}[A]$. We get $\mathbf{Pr}_\rho[B > 0] > 1/2^{O(d)}$, which implies $\mathbf{Pr}_\rho[A > \mathbf{E}_\rho[A]] > 1/2^{O(d)}$ and we are done. □

4.3.2 The large critical index case

Finally we consider PTFs $f = \mathrm{sign}(p)$ with τ-critical index greater than $K = 2d\log n/\tau$. Let ρ be a restriction of the first K variables $\mathcal{H} = \{1, \ldots, K\}$; we call these the "head" variables. We will show the following:

Lemma 4.7. *For a $1/2^{O(d)}$ fraction of restrictions ρ, the function $\mathrm{sign}(p_\rho(x))$ is a constant function.*

Proof. By Lemma 4.1, the surviving variables $x_{K+1}, \ldots, x_n$ have very small total influence in p:

$$\sum_{i=K+1}^{n} \mathrm{Inf}_i(p) = \sum_{i=K+1}^{n} \sum_{S \ni i} \widehat{p}(S)^2 \le (1-\tau)^K \cdot \mathrm{Inf}(p) \le d/n^{2d}. \tag{4}$$

Therefore, if we let p' be the truncation of p comprising only the monomials with all variables in $\mathcal{H}$,

$$p'(x_1, \ldots, x_k) = \sum_{S \subset \mathcal{H}} \widehat{p}(S) x_S,$$

we know that almost all of the original Fourier weight of p is on the coefficients of p':

$$1 \ge \sum_{\substack{S \subset \mathcal{H} \\ |S|>0}} \widehat{p}(S)^2 \ge 1 - \sum_{i=K+1}^{n} \mathrm{Inf}_i(p) \ge 1 - d/n^{2d}.$$

Applying Theorem 2.4 to p' [1] we get $\mathbf{Pr}_{x \in \{-1,1\}^K}[|p'(x)| \ge 1/2^{O(d)}] \ge 1/2^{O(d)}$. In words, for a $1/2^{O(d)}$ fraction of all restrictions ρ to $x_1, \ldots, x_K$, the value $p'(\rho)$ has magnitude at least $1/2^{O(d)}$.

For any such restriction, if the function $f_\rho(x)$ is not a constant function it must necessarily be the case that:

$$\sum_{\emptyset \ne S \subseteq \{K+1, \ldots, n\}} |\widehat{p_\rho}(S)| \ge 1/2^{O(d)}.$$

As noted in (4), each tail variable $\ell > K$ has very small influence in p: $\mathrm{Inf}_\ell(p) \le \sum_{i=K+1}^{n} \mathrm{Inf}_i(p) = d/n^{2d}$.

Applying Lemma 4.6, we get that for the overwhelming majority of the $1/2^{O(d)}$ fraction of restrictions mentioned above, the influence of ℓ in p_ρ is not much larger than the influence of ℓ in p:

$$\mathrm{Inf}_\ell(p_\rho) \le (3\log n)^d \cdot \mathrm{Inf}_\ell(p) \le d \cdot (3\log n)^d / n^{2d} \tag{5}$$

Using Cauchy-Schwarz, we have

$$\begin{aligned} & \sum_{S \ni \ell, S \subseteq \{x_{K+1}, \ldots, x_n\}} |\widehat{p_\rho}(S)| \\ \le \quad & n^{d/2} \cdot \sqrt{\sum_{S \ni \ell, S \subseteq \{x_{K+1}, \ldots, x_n\}} \widehat{p_\rho}(S)^2} \\ = \quad & n^{d/2} \sqrt{\mathrm{Inf}_\ell(p_\rho)} \le n^{-\Omega(1)} \end{aligned}$$

where we have used (5) (and our upper bound on d). From this we easily get that

$$\sum_{0 < |S| \subseteq \{x_{K+1}, \ldots, x_n\}} |\widehat{p_\rho}(S)| \le n^{-\Omega(1)} \ll 1/2^{O(d)}.$$

We have established that for a $1/2^{O(d)}$ fraction of all restrictions to $x_1, \ldots, x_K$, the function $f_\rho = \mathrm{sign}(p_\rho)$ is a constant function, and the lemma is proved. □

4.3.3 Proof of Claim 4.4

If f is a degree-d PTF that is not τ-regular, then its τ-critical index is either in the range $\{1, \ldots, K\}$ or it is greater than K.

In the first case (small critical index case), as shown in Section 4.3.1, we have that for a $1/2^{O(d)}$ fraction of restrictions ρ to variables $x_1, \ldots, x_k$, the total influence of $f_\rho = \mathrm{sign}(p_\rho)$ is at most

$$O(d \cdot n \cdot (\tau')^{1/(4d+1)}) = O(d \cdot (\log n)^{1/4} \cdot n^{1-1/(4d+2)}),$$

so the conclusion of Claim 4.4 holds in this case.

In the second case (large critical index case), as shown in Section 4.3.2, for a $1/2^{O(d)}$ fraction of restrictions ρ to $x_1, \ldots, x_K$ the function f_ρ is constant and hence has zero influence, so the conclusion of Claim 4.4 certainly holds in this case as well. □

5. BOOLEAN AVERAGE SENSITIVITY: A FOURIER-ANALYTIC BOUND

In this section, we present a simple proof of the following upper bound on the average sensitivity of a degree-d PTF (Theorem 1.2): $\mathrm{AS}(n, d) \le 2n^{1-1/2^d}$.

We recall here the definition of the formal derivative of a function $f : \{-1,1\}^n \to \mathbb{R}$.

$$D_i f(x) = \sum_{S \ni i} \widehat{f}_S x_{S - \{i\}}.$$

It is easy to see that,

$$D_i f(x) = \frac{1}{2} x_i [f(x) - f(x^{\oplus i})] = \frac{1}{2}\left(\frac{f(x) - f(x^{\oplus i})}{x_i}\right) \tag{6}$$

where "$x^{\oplus i}$" means "x with the i-th bit flipped."

For a Boolean function f, we have $D_i f(x) = \pm 1$ iff flipping the ith bit flips f; otherwise $D_i f(x) = 0$. So we have

$$\mathrm{Inf}_i(f) = \mathbf{E}[|D_i f(x)|].$$

Lemma 5.1. *Fix $i \ne j \in [n]$. Let $f, g : \{-1,1\}^n \to \mathbb{R}$ be functions such that f is independent of the i^{th} bit x_i and g is independent of the j^{th} bit x_j. Then*

$$\mathbf{E}_x[x_i x_j f(x) g(x)] \le \frac{\mathrm{Inf}_i(g) + \mathrm{Inf}_j(f)}{2}.$$

[1] after a very slight rescaling so the non-constant Fourier coefficients of p' have sum of squares equal to 1; this does not affect the bound we get because of the big-O.

PROOF. First, note that the influence of i^{th} coordinate on a function f can be written as $\mathrm{Inf}_i(f) =$

$$\begin{aligned}\mathbf{E}_{x_{-i}}[\mathrm{Var}_{x_i}[f(x)]] &= \mathbf{E}_x\left[\left(\frac{|f(x^{\oplus i})-f(x)|}{2}\right)^2\right]\\ &= \mathbf{E}_{x_{-i}}\left[|\mathbf{E}_{x_i}[x_i f(x)]|^2\right] \qquad (7)\end{aligned}$$

As f is independent of x_i and g is independent of x_j, we can write,

$$\begin{aligned}\mathbf{E}_x[x_i x_j f(x)g(x)] &= \mathbf{E}_{x_{-\{i,j\}}}\,\mathbf{E}_{x_i,x_j}\left[x_i x_j f(x)g(x)\right]\\ &= \mathbf{E}_{x_{-\{i,j\}}}\left[\mathbf{E}_{x_i}[x_i g(x)]\,\mathbf{E}_{x_j}[x_j f(x)]\right]\\ &\le \mathbf{E}_{x_{-\{i,j\}}}\left[\frac{1}{2}|\,\mathbf{E}_{x_i}[x_i g(x)]|^2 + \frac{1}{2}|\,\mathbf{E}_{x_j}[x_j f(x)]|^2\right]\\ &\le \frac{\mathrm{Inf}_j(f)+\mathrm{Inf}_i(g)}{2}.\end{aligned}$$

where the first inequality uses $ab \le \frac{1}{2}(a^2+b^2)$ and the second uses Equation 7. □

Theorem 1.2 is shown using an inductive argument over the degree d. Central to this inductive argument is the following lemma relating the influences of a degree-d PTF $\mathrm{sign}(p(x))$ to the degree-$(d-1)$ PTFs obtained by taking formal derivatives of p.

LEMMA 5.2. *For a PTF $f = \mathrm{sign}(p(x))$ on n variables and $i \in [n]$, $\mathrm{Inf}_i(f) = \mathbf{E}[f(x)x_i\mathrm{sign}(D_i p(x))]$.*

The following simple claim will be useful in the proof of the above lemma.

CLAIM 5.3. *For two real numbers a, b, if $\mathrm{sign}(a) \neq \mathrm{sign}(b)$ then $\mathrm{sign}(\mathrm{sign}(a)-\mathrm{sign}(b)) = \mathrm{sign}(a-b)$.*

PROOF OF LEMMA 5.2. The influence of the i^{th} coordinate is given by,

$$\begin{aligned}\mathrm{Inf}_i(f) &= \mathbf{E}\left[\frac{1}{2}|f(x)-f(x^{\oplus i})|\right]\\ &= \mathbf{E}\left[\frac{1}{2}\left(f(x)-f(x^{\oplus i})\right)\mathrm{sign}\left(f(x)-f(x^{\oplus i})\right)\right]\end{aligned} \tag{8}$$

Consider an x for which $f(x) \neq f(x^{\oplus i})$. In this case, we can use Claim 5.3 to conclude:

$$\begin{aligned}\mathrm{sign}\left(f(x)-f(x^{\oplus i})\right) &= \mathrm{sign}\left(p(x)-p(x^{\oplus i})\right)\\ &= \mathrm{sign}(2x_i D_i p(x)) = x_i\mathrm{sign}(D_i p(x)),\end{aligned}$$

using (6). Hence for an x with $f(x) \neq f(x^{\oplus i})$,

$$\begin{aligned}&\left(f(x)-f(x^{\oplus i})\right)\mathrm{sign}\left(f(x)-f(x^{\oplus i})\right)\\ =\;&\left(f(x)-f(x^{\oplus i})\right)x_i\mathrm{sign}(D_i p(x))\,.\end{aligned}$$

On the other hand, if $f(x) = f(x^{\oplus i})$ then the above equation continues holds since both the sides evaluate to 0. Substituting this equality into Equation 8 yields,

$$\begin{aligned}\mathrm{Inf}_i(f) &= \frac{1}{2}\mathbf{E}\left[f(x)x_i\mathrm{sign}(D_i p(x))\right]\\ &- \frac{1}{2}\mathbf{E}\left[f(x^{\oplus i})x_i\mathrm{sign}(D_i p(x))\right].\end{aligned}$$

Notice that the i^{th} coordinate $(x^{\oplus i})_i$ of $x^{\oplus i}$ is given by $-x_i$. Since $D_i p$ is independent of the i^{th} coordinate x_i, we have $D_i p(x) = D_i p(x^{\oplus i})$. Rewriting the above equation, we get

$$\begin{aligned}\mathrm{Inf}_i(f) &= \frac{1}{2}\mathbf{E}\left[f(x)x_i\mathrm{sign}(D_i p(x))\right]\\ &+ \frac{1}{2}\mathbf{E}\left[f(x^{\oplus i})(x^{\oplus i})_i\mathrm{sign}(D_i p(x^{\oplus i}))\right]\\ &= \mathbf{E}\left[f(x)x_i\mathrm{sign}(D_i p(x))\right]\end{aligned}$$

□

THEOREM 5.4. *Let $\mathrm{AS}(n,d)$ denote the max possible average sensitivity of any degree-d PTF on n variables. Then we have*

$$\mathrm{AS}(n,d) \le \sqrt{n + n\cdot\mathrm{AS}(n,d-1)}.$$

PROOF.

$$\begin{aligned}\mathrm{Inf}(f) &= \sum_i \mathrm{Inf}_i(f)\\ &= \sum_i \mathbf{E}[f(x)x_i\mathrm{sign}(D_i p(x))] \qquad \text{(by Lemma 5.2)}\\ &= \mathbf{E}[f(x)\sum_i x_i\mathrm{sign}(D_i p(x))]\\ &\le \sqrt{\mathbf{E}[f(x)^2]}\cdot\sqrt{\mathbf{E}[(\sum_i x_i\mathrm{sign}(D_i p(x)))^2]} \qquad (9)\\ &= 1\cdot\sqrt{\mathbf{E}[\sum_{i,j} x_i x_j\mathrm{sign}(D_i p(x))\mathrm{sign}(D_j p(x))]} \qquad (10)\\ &\le \sqrt{\mathbf{E}[\sum_i x_i^2\mathrm{sign}(D_i p(x))^2] + \sum_{i\neq j}\mathrm{Inf}_i(\mathrm{sign}(D_j p(x)))}\\ &= \sqrt{n + \sum_{i\neq j}\mathrm{Inf}_i(\mathrm{sign}(D_j p(x)))}.\end{aligned}$$

Here (9) is the Cauchy-Schwarz inequality, (10) is expanding the square. The last inequality uses Lemma 5.1 which we may apply since $D_i p(x)$ does not depend on x_i.

Observe that for any fixed j', we have $D_{j'}p(x)$ is a degree-$(d-1)$ polynomial and $\mathrm{sign}(D_{j'}p(x))$ is a degree-$(d-1)$ PTF. Hence, by definition we have,

$$\sum_{i\neq j'}\mathrm{Inf}(\mathrm{sign}(D_{j'}p(x))) \le \mathrm{AS}(n,d-1)\,,$$

for all $j' \in [n]$. Therefore the quantity $\sum_{i\neq j}\mathrm{Inf}(\mathrm{sign}(D_j p(x))) \le n\cdot\mathrm{AS}(n,d-1)$, finishing the proof. □

The bound on average sensitivity (Theorem 1.2) follows immediately from the above recursive relation.

PROOF OF THEOREM 1.2. Clearly, we have $\mathrm{AS}(n,0) = 0$. For $d = 1$, Theorem 5.4 yields $\mathrm{AS}(n,1) \le \sqrt{n}$. Now suppose $\mathrm{AS}(n,d) = 2n^{1-1/2^d}$ for $d \ge 1$, then by Theorem 5.4,

$$\mathrm{AS}(n,d{+}1) \le \sqrt{n + n\cdot\mathrm{AS}(n,d)} \le \sqrt{4n^{2-1/2^d}} = 2n^{1-1/2^{d+1}}\,,$$

finishing the proof. □

6. BOOLEAN AVERAGE SENSITIVITY VS NOISE SENSITIVITY

Our results on Boolean noise sensitivity are obtained via the following simple reduction which translates any upper bound on average sensitivity for degree-d PTFs over Boolean variables into a corresponding upper bound on noise sensitivity. This theorem is inspired by the proof of noise sensitivity of halfspaces by Peres [35].

THEOREM 6.1. *Let* $\mathrm{NS}(\epsilon, d)$ *denote the maximum noise sensitivity of a degree d-PTF at a noise rate of* ϵ. *For all* $0 \leq \epsilon \leq 1$ *if* $m = \lfloor \frac{1}{\epsilon} \rfloor$ *then* $\mathrm{NS}(\epsilon, d) \leq \frac{1}{m}\mathrm{AS}(m, d)$.

Theorem 1.3 follows immediately from this reduction along with our bounds on Boolean average sensitivity (Theorems 1.1 and 1.2), so it remains for us to prove Theorem 6.1.

6.1 Proof of Theorem 6.1

Let $f(x) = \mathrm{sign}(p(x))$ be a degee d-PTF. Let us denote $\delta = \frac{1}{m}$. As $\delta \geq \epsilon$, by the monotonicity of noise sensitivity we have $\mathrm{NS}_\epsilon(f) \leq \mathrm{NS}_\delta(f)$. In the following, we will show that $\mathrm{NS}_\delta(f) \leq \frac{1}{m}\mathrm{AS}(m, d)$ which implies the intended result. Recall that $\mathrm{NS}_\delta(f)$ is defined as

$$\mathrm{NS}_\delta(f) = \mathbf{Pr}_{x \sim_\delta y}\left[f(x) \neq f(y)\right] ,$$

where $x \sim_\delta y$ denotes that y is generated by flipping each bit of x independently with probability δ. An alternate way to generate y from x is as follows:

- Sample $r \in \{1, \ldots, m\}$ uniformly at random.
- Partition the bits of x into $m = \frac{1}{\delta}$ sets $S_1, S_2, \ldots, S_m$ by independently assigning each bit to a uniformly random set. Formally, a partition α is specified by a function $\alpha : \{1, \ldots, n\} \to \{1, \ldots, m\}$ mapping bit locations to their partition numbers, i.e., $i \in S_{\alpha(i)}$. A uniformly random partition is picked by sampling $\alpha(i)$ for each $i \in \{1, \ldots, n\}$ uniformly at random from $\{1, \ldots, m\}$.
- Flip the bits of x contained in the set S_r to obtain y.

Each bit of x belongs to the set S_r independently with probability $\frac{1}{m} = \delta$. Therefore, the vector y generated by the above procedure can equivalently be generated by flipping each bit of x with probability δ.

Inspired by the above procedure, we now define an alternate equivalent procedure to generate the pair $x \sim_\delta y$.

- Sample $a \in \{-1, 1\}^n$ uniformly at random.
- Sample a uniformly random partition $\alpha : \{1, \ldots, n\} \to \{1, \ldots, m\}$ of the bits of a.
- Sample $z \in \{-1, 1\}^m$ uniformly at random.
- Sample $r \in \{1, \ldots, m\}$ uniformly at random. Let $\tilde{z} = z^{\oplus r}$ and let $x_i = a_i z_{\alpha(i)}$, $y_i = a_i \tilde{z}_{\alpha(i)}$

Notice that x is uniformly distributed in $\{-1, 1\}^n$, since both a and z are uniformly distributed in $\{-1, 1\}^n$ and $\{-1, 1\}^m$ respectively. Furthermore, $\tilde{z}_i = z_i$ for all $i \neq r$ and $\tilde{z}_r = -z_r$. Therefore, y is obtained by flipping the bits of x in the coordinates belonging to the r^{th} partition. As the partition α is generated uniformly at random, this amounts to flipping each bit of x with probability exactly $\frac{1}{m} = \delta$.

The noise sensitivity of f can be rewritten as,

$$\mathrm{NS}_\delta(f) = \mathbf{Pr}_{a,\alpha,z,r}\left[f(x) \neq f(y)\right]$$

For a fixed choice of a and α, $f(x)$ is a function of z. In this light, let us define the function $f_{a,\alpha} : \{-1, 1\}^m \to \{-1, 1\}$ for each a, α as $f_{a,\alpha}(z) = f(x)$. Returning to the expression for noise sensitivity we get:

$$\begin{aligned}
\mathrm{NS}_\delta(f) &= \mathbf{Pr}_{a,\alpha,z,r}\left[f_{a,\alpha}(z) \neq f_{a,\alpha}(\tilde{z})\right] \\
&= \mathbf{E}_{a,\alpha,z,r}\left[\mathbf{1}[f_{a,\alpha}(z) \neq f_{a,\alpha}(z^{\oplus r})]\right] \\
&= \mathbf{E}_{a,\alpha,z}\left[\frac{1}{m}\sum_{r=1}^{m} \mathbf{1}\left[f_{a,\alpha}(z) \neq f_{a,\alpha}(z^{\oplus r})\right]\right] \\
&= \mathbf{E}_{a,\alpha}\left[\frac{1}{m}\sum_{r=1}^{m} \mathbf{E}_z\left[\mathbf{1}\left[f_{a,\alpha}(z) \neq f_{a,\alpha}(z^{\oplus r})\right]\right]\right] .
\end{aligned}$$

In the above calculation, the notation $\mathbf{1}[E]$ refers to the indicator function of the event E. Recall that, by definition of influences,

$$\mathrm{Inf}_r(f_{a,\alpha}) = \mathbf{E}_z\left[\mathbf{1}\left[f_{a,\alpha}(z) \neq f_{a,\alpha}(z^{\oplus r})\right]\right] ,$$

for all r. Thus, we can rewrite the noise sensitivity of f as

$$\mathrm{NS}_\delta(f) = \mathbf{E}_{a,\alpha}\left[\frac{1}{m}\sum_{r=1}^{m}\mathrm{Inf}_r(f_{a,\alpha})\right] = \frac{1}{m}\,\mathbf{E}_{a,\alpha}\left[\mathrm{Inf}(f_{a,\alpha})\right] . \tag{11}$$

We claim that $f_{a,\alpha}$ is a degree d-PTF in m variables. To see this observe that

$$f_{a,\alpha}(z) = \mathrm{sign}(p(x_1, \ldots, x_n)) = \mathrm{sign}\left(p(a_1 z_{\alpha(1)}, \ldots, a_n z_{\alpha(n)})\right) ,$$

which for a fixed choice of a, α is a degree d-PTF in z. Consequently, by definition of $\mathrm{AS}(m, d)$ we have $\mathrm{Inf}(f_{a,\alpha}) \leq \mathrm{AS}(m, d)$ for all a and α. Using this in (11), the result follows.

7. ACKNOWLEDGMENT

Ilias Diakonikolas is supported by NSF grant CCF-0728736, and by an Alexander S. Onassis Foundation Fellowship. Part of this work was done while visiting IBM Almaden. Prahladh Harsha is supported by THECB ARP grant 003658-0161-2007. Research done while the author was at the University of Texas at Austin. Adam Klivans is supported by NSF grants CCF-0643829, CCF-0728536 and THECB ARP grant 003658-0161-2007. Raghu Meka is partially supported by NSF grants CCF-0634811 and CCF-0916160 and THECB ARP grant 003658-0113-2007. Prasad Raghavendra is supported by NSF CCF-0343672. Part of this research was done while at the University of Washington and visiting Carnegie Mellon University. Rocco A. Servedio is supported by NSF grants CCF-0347282, CCF-0523664 and CNS-0716245, and by DARPA award HR0011-08-1-0069. Li-Yang Tan is supported by DARPA award no. HR0011-08-1-0069 and NSF Cybertrust grant no. CNS-0716245.

8. REFERENCES

[1] P. Austrin and J. Håstad. Randomly supported independence and resistance. In *Proc. 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 483–492. ACM, 2009.

[2] I. Benjamini, G. Kalai, and O. Schramm. Noise sensitivity of Boolean functions and applications to percolation. *Inst. Hautes Études Sci. Publ. Math.*, 90:5–43, 1999.

[3] E. Blais, R. O'Donnell, and K. Wimmer. Polynomial regression under arbitrary product distributions. In *Proc. 21st Annual Conference on Learning Theory (COLT)*, pages 193–204, 2008.

[4] V. Bogachev. *Gaussian measures.* Mathematical surveys and monographs, vol. 62, 1998.

[5] J. Bourgain and G. Kalai. Influences of variables and threshold intervals under group symmetries. *GAFA*, 7:438–461, 1997.

[6] N. Bshouty and C. Tamon. On the Fourier spectrum of monotone functions. *Journal of the ACM*, 43(4):747–770, 1996.

[7] A. Carbery and J. Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in R^n. *Mathematical Research Letters*, 8(3):233–248, 2001.

[8] I. Diakonikolas, P. Raghavendra, R. Servedio, and L.-Y. Tan. Average sensitivity and noise sensitivity of polynomial threshold functions, 2009. Available at http://arxiv.org/abs/0909.5011.

[9] I. Diakonikolas and R. Servedio. Improved approximation of linear threshold functions. In *Proc. 24th Annual IEEE Conference on Computational Complexity (CCC)*, pages 161–172, 2009.

[10] I. Diakonikolas, R. Servedio, L.-Y. Tan, and A. Wan. A regularity lemma, and low-weight approximators, for low-degree polynomial threshold functions. manuscript, 2009.

[11] I. Diakoniokolas, P. Gopalan, R. Jaiswal, R. Servedio, and E. Viola. Bounded independence fools halfspaces. In *Proc. 50th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 171–180, 2009.

[12] I. Dinur, E. Friedgut, G. Kindler, and R. O'Donnell. On the Fourier tails of bounded functions over the discrete cube. In *Proc. 38th ACM Symp. on Theory of Computing*, pages 437–446, 2006.

[13] W. Feller. *An introduction to probability theory and its applications.* John Wiley & Sons, 1968.

[14] E. Friedgut. Boolean functions with low average sensitivity depend on few coordinates. *Combinatorica*, 18(1):474–483, 1998.

[15] P. Gopalan, A. Kalai, and A. Klivans. Agnostically learning decision trees. In *Proc. 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 527–536, 2008.

[16] P. Gopalan and R. Servedio. Learning threshold-of-AC^0 circuits. Manuscript, 2009.

[17] C. Gotsman and N. Linial. Spectral properties of threshold functions. *Combinatorica*, 14(1):35–50, 1994.

[18] P. Harsha, A. Klivans, and R. Meka. Bounding the sensitivity of polynomial threshold functions. Available at http://arxiv.org/abs/0909.5175, 2009.

[19] J. Jackson, A. Klivans, and R. Servedio. Learnability beyond AC^0. In *Proc. 34th Annual ACM Symposium on Theory of Computing (STOC)*, pages 776–784, 2002.

[20] S. Janson. *Gaussian Hilbert Spaces.* Cambridge University Press, Cambridge, UK, 1997.

[21] J. Kahn, G. Kalai, and N. Linial. The influence of variables on boolean functions. In *Proc. 29th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 68–80, 1988.

[22] A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[23] A. Kalai, Y. Mansour, and E. Verbin. On agnostic boosting and parity learning. In *Proc. 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 629–638, 2008.

[24] D. Kane. The Gaussian surface area and noise sensitivity of degree-d polynomial threshold functions. CCC 2010, to appear, 2010.

[25] A. Klivans, R. O'Donnell, and R. Servedio. Learning intersections and thresholds of halfspaces. *Journal of Computer & System Sciences*, 68(4):808–840, 2004.

[26] A. Klivans, R. O'Donnell, and R. Servedio. Learning geometric concepts via Gaussian surface area. In *Proc. 49th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 541–550, 2008.

[27] N. Linial, Y. Mansour, and N. Nisan. Constant depth circuits, Fourier transform and learnability. *Journal of the ACM*, 40(3):607–620, 1993.

[28] E. Mossel. Lecture 4. Available at http://www.stat.berkeley.edu/~mossel/teach/206af05/scribes/sep8.pdf, 2005.

[29] E. Mossel and R. O'Donnell. On the noise sensitivity of monotone functions. *Random Structures and Algorithms*, 23(3):333–350, 2003.

[30] E. Mossel, R. O'Donnell, and K. Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. In *Proc. 46th Symposium on Foundations of Computer Science (FOCS)*, pages 21–30, 2005.

[31] R. O'Donnell. Lecture 16: The hypercontractivity theorem. Available at http://www.cs.cmu.edu/~odonnell/boolean-analysis/lecture16.pdf, 2007.

[32] R. O'Donnell, M. Saks, O. Schramm, and R. Servedio. Every decision tree has an influential variable. In *Proc. 46th Symposium on Foundations of Computer Science (FOCS)*, pages 31–39, 2005.

[33] R. O'Donnell and R. Servedio. Learning monotone decision trees in polynomial time. *SIAM J. Comput.*, 37(3):827–844, 2007.

[34] R. O'Donnell and R. Servedio. The Chow Parameters Problem. In *Proc. 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 517–526, 2008.

[35] Y. Peres. Noise stability of weighted majority, 2004.

[36] O. Schramm and J. Steif. Quantitative noise sensitivity and exceptional times for percolation. Ann. Math., to appear.

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

[38] Y. Shi. Lower bounds of quantum black-box complexity and degree of approximating polynomials by influence of boolean variables. *Inform. Process. Lett.*, 75(1-2):79–83, 2000.

[39] S. S. Shwartz, O. Shamir, and K. Sridharan. Agnostically learning halfspaces with margin errors. TTI Technical Report, 2009.