Bounded Independence Fools Degree-2 Threshold Functions

Ilias Diakonikolas[†] ilias@cs.columbia.edu

Daniel M. Kane[‡] dankane@math.harvard.edu

Jelani Nelson[§] minilek@mit.edu

Abstract

Let x be a random vector coming from any k-wise independent distribution over $\{-1,1\}^n$. For an n-variate degree-2 polynomial p, we prove that $\mathbf{E}[\mathrm{sgn}(p(x))]$ is determined up to an additive ε for $k = \mathrm{poly}(1/\varepsilon)$. This gives a large class of explicit pseudo-random generators against such functions and answers an open question of Diakonikolas et al. (FOCS 2009).

In the process, we develop a novel analytic technique we dub *multivariate FT-mollification*. This provides a generic tool to approximate bounded (multivariate) functions by *low-degree* polynomials (with respect to several different notions of approximation). A univariate version of the method was introduced by Kane et al. (SODA 2010) in the context of streaming algorithms. In this work, we refine it and generalize it to the multivariate setting. We believe that our technique is of independent interest. To illustrate its generality, we note that it implies a multidimensional generalization of Jackson's classical result in approximation theory due to (Newman and Shapiro, 1963).

To obtain our main result, we combine the *FT-mollification* technique with several linear algebraic and probabilistic tools. These include the invariance principle of of Mossell, O'Donnell and Oleszkiewicz, anti-concentration bounds for low-degree polynomials, an appropriate decomposition of degree-2 polynomials, and the Hanson-Wright tail bound for quadratic forms which takes the operator norm of the associated matrix into account. Our analysis is quite modular; it readily adapts to show that intersections of halfspaces and degree-2 threshold functions are fooled by bounded independence. From this it follows that $\Omega(1/\varepsilon^2)$-wise independence derandomizes the Goemans-Williamson hyperplane rounding scheme.

Our techniques unify, simplify, and in some cases improve several recent results in the literature concerning threshold functions. For the case of "regular" halfspaces we give a simple proof of an optimal independence bound of $\Theta(1/\varepsilon^2)$, improving upon Diakonikolas et al. (FOCS 2009) by polylogarithmic factors. This yields the first optimal derandomization of the Berry-Esséen theorem and – combined with the results of Kalai et al. (FOCS 2005) – implies a faster algorithm for the problem of agnostically learning halfspaces.

[1] Department of Computer Science, Columbia University. Research supported by NSF grant CCF-0728736, and by an Alexander S. Onassis Foundation Fellowship. Part of this work was done while interning at IBM Almaden.

[2] Harvard University, Department of Mathematics. Supported by a National Defense Science and Engineering Graduate (NDSEG) Fellowship.

[3] MIT Computer Science and Artificial Intelligence Laboratory. Supported by a National Defense Science and Engineering Graduate (NDSEG) Fellowship, and in part by the Center for Massive Data Algorithmics (MADALGO) - a center of the Danish National Research Foundation. Part of this work was done while interning at IBM Almaden.

1 Introduction

This paper is concerned with the power of limited independence to fool low-degree polynomial threshold functions. A degree-d *polynomial threshold function* (henceforth PTF), is a boolean function $f : \{-1,1\}^n \to \{-1,1\}$ expressible as $f(x) = \mathrm{sgn}(p(x))$, where p is an n-variate degree-d polynomial with real coefficients, and sgn is -1 for negative arguments and 1 otherwise. PTFs have played an important role in computer science since the early perceptron work of Minsky and Papert [37], and have since been extensively investigated in circuit complexity and communication complexity [3, 7, 12, 13, 22, 24, 34, 41, 42, 45, 46] and learning theory [32, 33, 47].

A distribution $\mathcal{D}$ on $\{-1,1\}^n$ is said to ε-fool a function $f : \{-1,1\}^n \to \{-1,1\}$ if

$$|\mathbf{E}_{x\sim\mathcal{D}}[f(x)] - \mathbf{E}_{x\sim\mathcal{U}}[f(x)]| \le \varepsilon$$

where $\mathcal{U}$ is the uniform distribution on $\{-1,1\}^n$. A distribution $\mathcal{D}$ on $\{-1,1\}^n$ is k-wise independent if every restriction of $\mathcal{D}$ to k coordinates is uniform on $\{-1,1\}^k$. Despite their simplicity, k-wise independent distributions have been a surprisingly powerful and versatile derandomization tool, fooling complex functions such as AC^0 circuits [5, 11, 44] and half-spaces [16]. As a result, this class of distributions has played a fundamental role in many areas of theoretical computer science.

Our Results and Techniques. The problem we study is the following: How large must $k = k(n,d,\varepsilon)$ be in order for *every* k-wise independent distribution on $\{-1,1\}^n$ to ε-fool the class of degree-d PTF's? The $d = 1$ case of this problem was recently considered in [16], where it was shown that $k(n,1,\varepsilon) = \widetilde{\Theta}(1/\varepsilon^2)$, independent of n. The main open problem in [16] was to identify $k = k(n,d,\varepsilon)$ for $d \ge 2$. In this work, we make progress on this question by proving the following:

Theorem 1.1 (Main Result). Any $\tilde{\Omega}(\varepsilon^{-9})$-wise independent distribution on $\{-1,1\}^n$ ε-fools all degree-2 PTFs.

Prior to this work, no nontrivial result was known for $d > 1$; it was not even known whether $o(n)$-wise independence suffices for constant ε. Standard explicit constructions of k-wise independent distributions over $\{\pm 1\}^n$ have seed length $O(k \cdot \log n)$ [1, 15] which is optimal up to constant factors. As a consequence, Theorem 1.1 gives a large class of explicit pseudo-random generators (PRGs) for degree-2 PTFs with seed length $\log(n) \cdot \widetilde{O}(\varepsilon^{-9})$.

Another consequence of Theorem 1.1 is that bounded independence suffices for the invariance principle of [38] in the case of degree-2 polynomials. Roughly, this says that for a “low influence” degree-2 polynomial p the distribution of $p(x)$ is essentially invariant if x is drawn from a k-wise distribution over n uniform random signs versus a k-wise distribution over n standard Gaussians. Under this interpretation, we believe that our result and its proof represent an advance of some substance in probability theory.

The techniques we employ to obtain our main result are quite robust. Our approach yields for example that Theorem 1.1 holds not only over the hypercube, but also over the n-variate Gaussian distribution. The proof also readily extends to show that the intersection of m halfspaces, or even m degree-2 threshold functions, is ε-fooled by $\mathrm{poly}(1/\varepsilon)$-wise independence for any constant m (over both the hypercube and the multivariate Gaussian); see Theorem 6.2. As a special case of the latter result we prove that the Goemans-Williamson hyperplane rounding scheme [21] (henceforth “GW rounding”) can be derandomized using $\Omega(1/\varepsilon^2)$-wise independence.[1]

[1] Concurrent independent work of [23] also implies $\Omega(\mathrm{polylog}(1/\varepsilon)/\varepsilon^2)$-independence suffices. Other derandomizations of GW rounding are known with better ε-dependence, though not solely via k-wise independence [31, 35, 48].

The method that we build may be viewed as a generic tool in approximation theory, and in particular for the approximation of functions by low-degree polynomials.[2] In Section 6 we use our method to obtain a simple proof of a classical result in polynomial approximation. We also obtain various quantitative improvements for problems related to streaming and agnostic learning.

Motivation and Related Work. The literature is rich with explicit generators for various natural classes of functions. In recent years, there has been much interest in not only constructing PRGs for natural complexity classes, but also in doing so with as broad and natural a family of PRGs as possible. One example is the recent work of Bazzi [5] on fooling depth-2 circuits (simplified by Razborov [44]), and of Braverman [11] on fooling AC^0, with bounded independence[3].

During the past year there has been a flurry of results on pseudo-random generators against threshold functions [8, 16, 23, 26, 31, 36, 43]. Most directly related to the results in this paper is the work of Meka and Zuckerman [36]. Simultaneously and independently from our work, they constructed PRGs against degree-d PTFs with seed length $\log n \cdot 2^{O(d)} \cdot (1/\varepsilon)^{8d+3}$ [36]. That is, their seed length for $d = 2$ is similar to ours (though our generator is better by a $\mathrm{poly}(1/\varepsilon)$ factor). Their PRG is not based on k-wise independence alone.

We note that, by a straightforward probabilistic argument, there exist generators with seed-length $O(d \log n + \log(1/\varepsilon))$ for degree-d PTFs. Hence, there is still a substantial gap between probabilistic and explicit constructions. We believe that the development of structural results such as the ones given in the current work may be useful for later developments of generators with better seed-length. In particular, we feel that understanding the degree of independence required to fool degree-d PTFs is an important step towards obtaining better explicit generators for these functions. For example, it is conceivable that $\mathrm{poly}(d/\varepsilon)$-independence suffices, which would yield generators with seed-length $\log n \cdot \mathrm{poly}(d/\varepsilon)$.

Organization. In Section 2 we give basic notation. Section 3 contains a high-level explanation of FT-mollification and explains how it is used to obtain our main result. Section 4 contains our FT-mollification theorem and a brief sketch of its proof. As a warmup for our main result, in Section 5 we show how our approach implies that $\Omega(1/\varepsilon^2)$-wise independence ε-fools "regular" halfspaces; the structure of this proof serves as a template which all later proofs follow. In Section 6, we show that our techniques also have connections to or yield improvements for various problems related to approximation theory, streaming, agnostic learning and fooling intersections of threshold functions. In Section 7 we present the proof of our main result. Due to space limitations, some full proofs are postponed to the appendix.

2 Notation

Let $p : \{-1,1\}^n \to \mathbb{R}$ be a polynomial and $p(x) = \sum_{S\subseteq[n]} \widehat{p}_S \chi_S$ be its Fourier-Walsh expansion, where $\chi_S(x) \stackrel{\mathrm{def}}{=} \prod_{i\in S} x_i$. The *influence* of variable i on p is $\mathrm{Inf}_i(p) \stackrel{\mathrm{def}}{=} \sum_{S\ni i} \widehat{p}_S^2$, and the *total influence* of p is $\mathrm{Inf}(p) = \sum_{i=1}^n \mathrm{Inf}_i(p)$. If $\mathrm{Inf}_i(p) \le \tau \cdot \mathrm{Inf}(p)$ for all i, we say that the polynomial p is *τ-regular*. If $f(x) = \mathrm{sgn}(p(x))$, where p is τ-regular, we say that f is a *τ-regular* PTF.

For $R \subseteq \mathbb{R}^d$ denote by $I_R : \mathbb{R}^d \to \{0,1\}$ its characteristic function. It will be convenient in some of the proofs to phrase our results in terms of ε-fooling $\mathbf{E}[I_{[0,\infty)}(p(x))]$ as opposed to $\mathbf{E}[\mathrm{sgn}(p(x))]$. It is straightforward that these are equivalent up to changing ε by a factor of 2.

[2] In fact our method is more general, and also provides good approximations by smooth functions with good derivative bounds. In some cases this move from polynomials to smooth functions is necessary, e.g. in [30].

[3] Note that a PRG for AC^0 with qualitatively similar – in fact slightly better – seed length had being already given by Nisan [40].

We frequently use $A \approx_\varepsilon B$ to denote that $|A - B| = O(\varepsilon)$, and we let the function $d_2(x, R)$ denote the L_2 distance from some $x \in \mathbb{R}^d$ to a region $R \subseteq \mathbb{R}^d$.

Finally, we familiarize the reader with some multi-index notation. A d-dimensional multi-index is a vector $\beta \in \mathbb{N}^d$ (here $\mathbb{N}$ is the nonnegative integers). For $\alpha, \beta \in \mathbb{N}^d$, we say $\alpha \leq \beta$ if the inequality holds coordinate-wise, and for such α, β we define $|\beta| = \sum_i \beta_i$, $\binom{\beta}{\alpha} = \prod_{i=1}^d \binom{\beta_i}{\alpha_i}$, and $\beta! = \prod_{i=1}^d \beta_i!$. For $x \in \mathbb{R}^d$ we use x^β to denote $\prod_{i=1}^d x_i^{\beta_i}$, and for $f : \mathbb{R}^d \to \mathbb{R}$ we use $\partial^\beta f$ to denote $\frac{\partial^{|\beta|}}{\partial x_1^{\beta_1} \cdots \partial x_d^{\beta_d}} f$.

3 Overview of our proof of Theorem 1.1

The program of our proof follows the outline of the proof in [16]: we first prove that bounded independence fools the class of *regular* degree-2 PTF's (Step 1), then reduce the general case to the regular case (Step 2) to show that bounded independence fools all degree-2 PTF's. The bulk of our proof is to establish Step 1; this is the most challenging part of this work and where our main technical contribution lies. Step 2 is achieved by adapting the recent results of [17]. We stress that proving Step 1 for the degree-2 case poses significant technical challenges. It turns out that the proof requires a conceptual departure from the approach used in the degree-1 case. We elaborate on this below.

Let $f : \{-1, 1\}^n \to \{-1, 1\}$ be a boolean function. To show that f is fooled by k-wise independence, it suffices – and is in fact necessary – to prove the existence of two degree-k "sandwiching" polynomials $q_u, q_l : \{-1, 1\}^n \to \{-1, 1\}$ that approximate f in L_1-norm (see e.g. [5, 9]). (Let us remark here that, because of the additional "sandwiching" condition this notion of approximation is at least as hard as L_1.) Even though this is an n-dimensional approximation problem, it may be possible to exploit the additional structure of the function under consideration to reduce it to a low-dimensional problem. This is exactly what is done in both [16] and [30] for the case of regular halfspaces.

We now briefly explain the approaches of [16] and [30]. Let $f(x) = \mathrm{sgn}(\langle w, x\rangle)$ be an ε^2-regular halfspace, i.e. $\|w\|_2 = 1$ and $\max_i |w_i| \leq \varepsilon$. The works of [16, 30] use the Berry-Esséen theorem, which states that the random variable $\langle w, x\rangle$ behaves approximately like a standard Gaussian and hence can be treated as if it was one-dimensional. Thus, both [16] and [30] construct (implicitly in the latter) a (different in each case) univariate polynomial $P : \mathbb{R} \to \mathbb{R}$ that is a good "upper sandwich" L_1-approximation to the sign function under the normal distribution in $\mathbb{R}$. The desired n-variate sandwiching polynomials are then obtained (roughly) by setting $q_u(x) = P(\langle w, x\rangle)$ and $q_l(x) = -P(-\langle w, x\rangle)$. It turns out that this approach suffices for the case of halfspaces. In [16] the polynomial P is constructed using classical approximation theory tools. In [30] it is obtained by taking a truncated Taylor expansion of a certain smooth approximation to the sign function, constructed via a method dubbed "Fourier Transform mollification" (henceforth FT-mollification).

Let $f(x) = \mathrm{sgn}(p(x))$ be a regular degree-2 PTF. A first natural attempt to handle this case would be to again use some *univariate* polynomial approximation Q to the sign function – potentially allowing its degree to increase – and then take $q_u(x) = Q(p(x))$, as before. Such an approach is easily seen to fail for both constructions outlined above – for essentially the same reason. This is not a coincidence; it is conjectured [19] that *no* univariate L_1 ε-approximating polynomial for the sign function (i.e., without even requiring the sandwiching condition) can have $2^{o(1/\varepsilon^2)}$-degree. We elaborate on this issue in Section E.

We now describe FT-mollification and our departure from the univariate approach.

3.1 FT-mollification FT-mollification is a general procedure to obtain a smooth function with bounded derivatives that approximates some bounded function f. The univariate version of the method in the context of derandomization was introduced in [30]. In this paper we refine the technique and generalize it to the multivariate setting, and later use it to prove our main theorem. We remark here that the FT-mollification construction given in the current work is not only a generalization of that in [30], but is redone from scratch and is simpler, while also yielding improved bounds even in univariate applications (see Section A.1 for details).

For the univariate case, where $f : \mathbb{R} \to \mathbb{R}$, [30] defined $\tilde{f}^c(x) = (c \cdot \hat{b}(c \cdot t) * f(t))(x)$ for a parameter c, where $\hat{b}$ has unit integral and is the Fourier transform of a smooth function b of compact support (a so-called *bump function*). Here "$*$" denotes convolution. The idea of smoothing functions via convolution with a smooth approximation of the Dirac delta function is old, dating back to "Friedrichs mollifiers" [18] in 1944. Indeed, the only difference between Friedrichs mollification and FT-mollification is that in the former, one convolves f with the scaled bump function, and not its Fourier transform. The switch to the Fourier transform is made to have better control on the high-order derivatives of the resulting smooth function, which is crucial for our application.

The method can be illustrated as follows. Let $X = \sum_i a_i X_i$ for independent X_i. Suppose we would like to argue that $\mathbf{E}[f(X)] \approx_\varepsilon \mathbf{E}[f(Y)]$, where $Y = \sum_i a_i Y_i$ for k-wise independent Y_i's that are individually distributed as the X_i. Let $\tilde{f}^c$ be the FT-mollified version of f. If the parameter $c = c(\varepsilon)$ is appropriately selected, we can guarantee that $|f(x) - \tilde{f}^c(x)| < \varepsilon$ "almost everywhere", and furthermore have "good" upper bounds on the high-order derivatives of $\tilde{f}^c$. We could then hope to show the following chain of inequalities: $\mathbf{E}[f(X)] \approx_\varepsilon \mathbf{E}[\tilde{f}^c(X)] \approx_\varepsilon \mathbf{E}[\tilde{f}^c(Y)] \approx_\varepsilon \mathbf{E}[f(Y)]$. To justify the first inequality, f and $\tilde{f}^c$ are close almost everywhere, and so it suffices to argue that X is sufficiently anti-concentrated in the small region where they are not close. The second inequality would use Taylor's theorem, bounding the error via upper bounds on moment expectations of X and the high-order derivatives of $\tilde{f}^c$. Showing the final inequality would be similar to the first, except that one needs to justify that even under k-wise independence the distribution of Y is sufficiently anti-concentrated. The argument outlined above was used in [30] to provide an alternative proof that bounded independence fools regular halfspaces, and to optimally derandomize Indyk's moment estimation algorithm in data streams [27].

We now describe our switch to multivariate FT-mollification. Let $f : \{-1,1\}^n \to \{-1,1\}$ be arbitrary, and let $S = f^{-1}(1) \subseteq \mathbb{R}^n$. Then, fooling $\mathbf{E}[f(x)]$ and fooling $\mathbf{E}[I_S(x)]$ are equivalent. A natural attempt to this end would be to generalize FT-mollification to n dimensions, then FT-mollify I_S and argue as above using the multivariate Taylor's theorem. Such an approach is perfectly valid, but as one might expect, there is a penalty for working over high dimensions. Both our quantitative bounds on the error introduced by FT-mollifying, and the error coming from the multivariate Taylor's theorem, increase with the dimension. Our approach is then to find a *low-dimensional representation* of such a region S which allows us to obtain the desired bounds. We elaborate below on how this can be accomplished in our setting.

3.2 Our Approach Let $f = \mathrm{sgn}(p)$ be a regular multilinear degree-2 PTF with $\|p\|_2 = 1$ (wlog). Let us assume for simplicity that p is a quadratic form; handling the additive linear form and constant is easier. Our approach is now as follows. We decompose p as $p_1 - p_2 + p_3$, where p_1, p_2 are positive semidefinite quadratic forms with no small non-zero eigenvalues and p_3 is indefinite with all eigenvalues small in magnitude; such a decomposition follows from elementary linear algebra.

Then, as suggested by the aforementioned, we would like to identify a low-dimensional region

$R \subseteq \mathbb{R}^d$ such that $I_{\{z:p(z)\geq 0\}}(x)$ can be written as $I_R(F(x))$ for some $F : \{-1,1\}^n \to \mathbb{R}^d$ that depends on the p_i, then FT-mollify I_R. The region R is selected as follows: note we can write $p_3(x) = x^T A_{p_3} x$, where A_{p_3} is a real symmetric matrix with trace Υ. We consider the region $R = \{z : z_1^2 - z_2^2 + z_3 + \Upsilon \geq 0\} \subseteq \mathbb{R}^3$ and define $F(x) = (\sqrt{p_1(x)}, \sqrt{p_2(x)}, p_3(x) - \Upsilon)$, then observe that $I_R(F(x)) = 1$ iff $p(x) \geq 0$. (Recall that p_1, p_2 are positive-semidefinite, hence the first two coordinates are always real.) We then prove via FT-mollification that $\mathbf{E}[I_R(F(x))]$ is preserved to within ε by bounded independence. Due to our choice of F, when applying Taylor's theorem our error grows only like $2^{O(k)} \cdot c^k \cdot (\mathbf{E}[\sqrt{p_1(x)}^k] + \mathbf{E}[\sqrt{p_2(x)}^k] + \mathbf{E}[(p_3(x) - \Upsilon)^k])/k^k$ for some (non-constant) c in our proof, and we want this error to be ε. Essentially, these square roots save us since kth moments of quadratic forms can grow like k^k, which would nullify the k^k in the denominator of Taylor's theorem; by having square roots, we only have to deal with $(k/2)$th moments. To handle p_3, we use a moment bound for quadratic forms with small eigenvalues. The fact that we need p_1, p_2 to not only be positive semidefinite, but to also have no small eigenvalues, is needed because quadratic forms with no small non-zero eigenvalues satisfy good tail bounds. This is revelant because $\tilde{I}_R^c(F(x))$ and $I_R(F(x))$ are *not* close for $F(x)$ near the boundary of R, and we can show that the probability of this event is small when p_1, p_2 satisfy good tail bounds.

4 Multivariate FT-mollification

We now state and sketch the proof of our FT-mollification theorem, which yields generic smoothing guarantees for arbitrary bounded functions mapping $\mathbb{R}^d$ to $\mathbb{R}$. The full proof is in Section A. In the proof of our main theorem (Theorem 1.1), we are concerned with $d = 4$. In some of the other applications of our technique mentioned in Section 6, d can be a growing parameter, e.g. it is the number of halfspaces when fooling intersections of halfspaces. In what follows, we refer to $\tilde{F}^c$ as the *FT-mollification* of F ("FT" for "Fourier Transform", for reasons that become clear in Section A).

Theorem 4.1. Let $F : \mathbb{R}^d \to \mathbb{R}$ be bounded, $c > 0$ arbitrary. There exists $\tilde{F}^c : \mathbb{R}^d \to \mathbb{R}$ satisfying

i. $\|\partial^\beta \tilde{F}^c\|_\infty \leq \|F\|_\infty \cdot (2c)^{|\beta|}$ for all $\beta \in \mathbb{N}^d$.

ii. Fix some $x \in \mathbb{R}^d$. Then if $|F(x) - F(y)| \leq \varepsilon$ whenever $\|x - y\|_2 \leq \delta$ for some $\varepsilon, \delta \geq 0$, then $|\tilde{F}^c(x) - F(x)| \leq \varepsilon + \|F\|_\infty \cdot O(d^2/(c^2\delta^2))$.

iii. $\tilde{F}^c$ is nonnegative if F is nonnegative.

Proof (Sketch). In Section A we show the existence of a probability density B on $\mathbb{R}^d$ satisfying $\mathbf{E}_{x\sim B}[\|x\|_2^2] = O(d^2)$, and $\|\partial^\beta B\|_1 \leq 2^{|\beta|}$ for all $\beta \in \mathbb{N}^d$. This density B is obtained by taking a "smooth enough" function $b : \mathbb{R}^d \to \mathbb{R}$ of compact support with $\int_{\mathbb{R}^d} b^2(y)dy = 1$, then letting B be the square of its Fourier transform. We then define $B_c(x) = c^d \cdot B(cx)$, and $\tilde{F}^c(x) = (B_c * F)(x) = \int_{\mathbb{R}^d} B_c(y)F(x-y)dy$.

For (i), using basic properties of convolution we show $\|\tilde{F}^c\|_\infty \leq \|F\|_\infty \cdot c^{|\beta|} \cdot \left\|\partial^\beta B\right\|_1$, at which point we use our bounds on $\|\partial^\beta B\|_1$. For (ii), since B is a probability density we have $\int_{\mathbb{R}^d} B_c(y)dy = 1$ for all c. Thus, $\int_{\mathbb{R}^d} B_c(x-y)F(y)dy = F(x) + \int_{\mathbb{R}^d}(F(y) - F(x))B_c(x-y)dy$. We then split the domain of integration into the regions $\|x - y\|_2 < \delta$ and $\|x - y\|_2 \geq \delta$. The integral over the first region is bounded by ε, and over the second region by the product of $\|F\|_\infty$ and a tail bound for B, which we can obtain by the second moment method since B has bounded variance. Item (iii) follows since for F nonnegative, $\tilde{F}^c$ is the convolution of two nonnegative functions. ∎

The following theorem is a corollary of Theorem 4.1 in the case F is the indicator function of a subset $R \subseteq \mathbb{R}^d$. In Theorem 4.2, and in later invocations of the theorem, we use the following notation: for $R \subset \mathbb{R}^d$, we let ∂R denote the boundary of R (specifically, ∂R denotes the set of points $x \in \mathbb{R}^d$ such that for every $\varepsilon > 0$, the ball about x of radius ε intersects both R and $\mathbb{R}^d \backslash R$).

Theorem 4.2. For any $R \subseteq \mathbb{R}^d$ and $x \in \mathbb{R}^d$, $|I_R(x) - \tilde{I}_R^c(x)| \le \min\{1, O((\frac{d}{c \cdot d_2(x, \partial R)})^2)\}$.

5 Warmup: fooling regular halfspaces

In this section, as a warmup to our main result we show how to use Theorem 4.2 to provide a simple proof that $\Omega(1/\varepsilon^2)$-wise independence fools the class of ε^2-regular halfspaces, i.e. halfspaces $\{x : \langle w, x\rangle \ge \theta\} \subseteq \{-1,1\}^n$ where $|w_i| \le \varepsilon$ for all i and $\|w\|_2 = 1$. This result is new and in fact it is optimal up to constant factors (see e.g. [16] for a straightforward $\Omega(1/\varepsilon^2)$ lower bound). This improves upon the bounds of [16, 30] by polylog$(1/\varepsilon)$ factors.

Theorem 5.1. Let $H_{w,\theta} = \{x : \langle w, x\rangle \ge \theta\}$ be a subset of $\{-1,1\}^n$ such that $|w_i| \le \varepsilon$ for all $i \in [n]$ with $\|w\|_2 = 1$, i.e. $H_{w,\theta}$ is ε^2-regular. Suppose $x_1, \ldots, x_n$ are independent Bernoulli, and $y_1, \ldots, y_n$ are k-wise independent Bernoulli for $k \ge C/\varepsilon^2$ for a sufficiently large even constant C. Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$. Then $|\mathbf{Pr}[x \in H_{w,\theta}] - \mathbf{Pr}[y \in H_{w,\theta}]| = O(\varepsilon)$.

Proof. Let $X = \langle w, x\rangle, Y = \langle w, y\rangle$. It is equivalent to show $|\mathbf{E}[I_{[\theta,\infty)}(X)] - \mathbf{E}[I_{[\theta,\infty)}(Y)]| = O(\varepsilon)$. We show the following chain of inequalities for $c = 1/\varepsilon$:

$$\mathbf{E}[I_{[\theta,\infty)}(X)] \approx_\varepsilon \mathbf{E}[\tilde{I}^c_{[\theta,\infty)}(X)] \approx_\varepsilon \mathbf{E}[\tilde{I}^c_{[\theta,\infty)}(Y)] \approx_\varepsilon \mathbf{E}[I_{\theta,\infty)}(Y)]$$

Here $\tilde{I}^c_{[\theta,\infty)}$ is as in Theorem 4.2, where $R = [\theta, \infty)$ and $d = 1$. Note then $d_2(z, \partial R)$ is just $|z - \theta|$.

(**A**) $\mathbf{E}[\mathbf{I}_{[\theta,\infty)}(\mathbf{X})] \approx_\varepsilon \mathbf{E}[\tilde{\mathbf{I}}^{\mathbf{c}}_{[\theta,\infty)}(\mathbf{X})]$: We have

$$\begin{aligned}|\mathbf{E}[I_{[\theta,\infty)}(X)] - \mathbf{E}[\tilde{I}^c_{[\theta,\infty)}(X)]| &\le \mathbf{E}[|I_{[\theta,\infty)}(X) - \tilde{I}^c_{[\theta,\infty)}(X)|] \\ &\le \mathbf{Pr}[|X-\theta| < \varepsilon] + \sum_{s=0}^{\infty} \mathbf{Pr}[2^s\varepsilon \le |X - \theta| < 2^{s+1}\varepsilon] \cdot O(c^{-2}2^{-2s}\varepsilon^{-2}) \\ &\le O(\varepsilon) + O(c^{-2}\varepsilon^{-2}) \cdot \sum_{s=0}^{\infty} 2^{-2s} \cdot \mathbf{Pr}[|X - \theta| < 2^{s+1}\varepsilon] \\ &= O(\varepsilon) + O(c^{-2}\varepsilon^{-2}) \cdot O(\varepsilon)\end{aligned}$$

since $\mathbf{Pr}[|X - \theta| \le t] = O(t + \varepsilon)$ for any $t > 0$, by ε^2-regularity and the Berry-Esséen Theorem. The above is $O(\varepsilon)$ by choice of c.

(**B**) $\mathbf{E}[\tilde{\mathbf{I}}^{\mathbf{c}}_{[\theta,\infty)}(\mathbf{X})] \approx_\varepsilon \mathbf{E}[\tilde{\mathbf{I}}^{\mathbf{c}}_{[\theta,\infty)}(\mathbf{Y})]$: By Taylor's theorem, $\tilde{I}^c_{[\theta,\infty)}(z) = P_{k-1}(z) \pm \|(\tilde{I}^c_{[\theta,\infty)})^{(k)}\|_\infty \cdot |z|^k/k!$ for $z \in \mathbb{R}$ and $f^{(k)}$ being the kth derivative of f, where P_{k-1} is a degree-$(k-1)$ polynomial. By k-wise independence, $\mathbf{E}[P_{k-1}(X)] = \mathbf{E}[P_{k-1}(Y)]$ and $\mathbf{E}[X^k] = \mathbf{E}[Y^k]$. For k even, $|z|^k = z^k$. Hence,

$$|\mathbf{E}[\tilde{I}^c_{[\theta,\infty)}(X)] - \mathbf{E}[\tilde{I}^c_{[\theta,\infty)}(Y)]| \le 2 \cdot \frac{\|(\tilde{I}^c_{[\theta,\infty)})^{(k)}\|_\infty \cdot \mathbf{E}[X^k]}{k!} \le 2^{O(k)} \cdot \frac{c^k \cdot k^{k/2}}{k^k},$$

which is $O(\varepsilon)$ since $k = \Omega(c^2)$. The last inequality used Theorem 4.2 to bound $\|(\tilde{I}^c_{[\theta,\infty)})^{(k)}\|_\infty$, and Khintchine's inequality to bound $\mathbf{E}[X^k]$ (also, a simple proof bounding $\mathbf{E}[X^k]$ is in Lemma B.1).

(C) $\mathbf{E}[\mathbf{I}_{[\theta,\infty)}(\mathbf{Y})] \approx_\varepsilon \mathbf{E}[\tilde{\mathbf{I}}^{\mathbf{c}}_{[\theta,\infty)}(\mathbf{Y})]$: This is argued identically as in the first inequality, but we now must show that even under $\Omega(1/\varepsilon^2)$-wise independence we still have $\mathbf{Pr}[|Y-\theta| \le \varepsilon] = O(\varepsilon)$. Suppose we had a function $f : \mathbb{R} \to \mathbb{R}$ such that (1) $f \ge I_{[\theta-\varepsilon,\theta+\varepsilon]}$ on $\mathbb{R}$ (implying for example $\mathbf{E}[f(Y)] \ge \mathbf{E}[I_{[\theta-\varepsilon,\theta+\varepsilon]}(Y)]$), (2) $\mathbf{E}[f(X)] = O(\varepsilon)$, and (3) $\|f^{(\ell)}\|_\infty \le O(1/\varepsilon)^\ell$ for all $\ell \ge 0$. Given (2) and (3), we can apply Taylor's theorem just as above to show $|\mathbf{E}[f(X)] - \mathbf{E}[f(Y)]| = O(\varepsilon)$, i.e. $\mathbf{E}[f(Y)] = O(\varepsilon)$. Using (1) then gives our desired upper bound on $\mathbf{E}[I_{[\theta-\varepsilon,\theta+\varepsilon]}(Y)] = \mathbf{Pr}[|Y-\theta| \le \varepsilon]$.

It simply remains to exhibit such a function f: we take $f = 2 \cdot \tilde{I}^{c'}_{[\theta-2\varepsilon,\theta+2\varepsilon]}$ for c' a sufficiently large constant times $1/\varepsilon$. To show (1), for $x \notin [\theta-\varepsilon,\theta+\varepsilon]$ we have $I_{[\theta-\varepsilon,\theta+\varepsilon]} = 0$, whereas f is nonnegative. For $x \in [\theta-\varepsilon,\theta+\varepsilon]$, we have $\min\{x-(\theta-2\varepsilon), x-(\theta+2\varepsilon)\} \ge \varepsilon$, implying $f(x) \ge 1$ by Theorem 4.2 and choice of c', and the fact that $I_{[\theta-2\varepsilon,\theta+2\varepsilon]}(x) = 1$ for such x. Item (2) follows by applying step (A) above. Item (3) follows from (i) of Theorem 4.1 and Taylor's theorem (as in (B) above). ■

The proof structure we gave above is similar to that in [30]. In particular, both use the same chain of three inequalities. However, due to differences in the FT-mollification guarantees of [30], the proof there gave a worse bound on k by a polylog$(1/\varepsilon)$ factor. Furthermore, the proof we give here of (C) is arguably more intuitive than that in [30], which relied on some complex analysis. One consequence of Theorem 5.1 is that the Berry-Esséen theorem is derandomized by $\Omega(1/\varepsilon^2)$-independence, which is optimal given an example in [16]. Specifically, Theorem 5.1 implies, after also carrying out the same argument under the Gaussian instead of Bernoulli measure, that $\sup_{t\in\mathbb{R}} |\mathbf{Pr}[\langle w,x\rangle \le t] - \mathbf{Pr}[\langle w,g\rangle \le t]| \le \varepsilon$ as long as the x_i and g_i are each $\Omega(1/\varepsilon^2)$-wise independent and $\|w\|_\infty \le \varepsilon$, where the x_i are Bernoulli and the g_i are Gaussian. The original Berry-Esséen theorem required independent x_i and g_i, and [16, 30] required polylog$(1/\varepsilon)/\varepsilon^2$-wise independence.

The theorem we proved above is a worthwhile warmup for the following reason: *all* our FT-mollification proofs that bounded independence fools various functions follow the same generic template as above. Specifically, when showing bounded independence fools some $f : \mathbb{R}^n \to \{0,1\}$, our proofs always begin by identifying a low-dimensional region $R \subset \mathbb{R}^d$, and a map $F : \mathbb{R}^n \to \mathbb{R}^d$, such that $f(x) = I_R(F(x))$. Then, fooling $\mathbf{E}[f(x)]$ and $\mathbf{E}[I_R(F(x))]$ are equivalent. In the proof above, $d = 1$, $R = [\theta,\infty)$, $F(x) = \langle w,x\rangle$. Our proofs then always proceed via the chain $\mathbf{E}[I_R(F(x))] \approx_\varepsilon \mathbf{E}[\tilde{I}^c_R(F(x))] \approx_\varepsilon \mathbf{E}[\tilde{I}^c_R(F(y))] \approx_\varepsilon \mathbf{E}[I_R(F(y))]$. The first requires anticoncentration near the boundary of R. The second uses the multivariate Taylor's theorem, employing moment bounds on the $F(x)_i$. The third requires showing anticoncentration around the boundary of R even under bounded independence; similarly as to above, to show anticoncentration in some set S under bounded independence, we argue through the function $\tilde{I}^c_{\{x:d_2(x,S)\le\varepsilon\}}$ for some sufficiently large c.

6 Other connections and implications of our techniques

To illustrate the generality of our methods, we now give a list of applications and connections to several problems.

6.1 A multidimensional Jackson's theorem

In 1963, Newman and Shapiro gave the following multidimensional Jackson's theorem for the unit L_2-ball:

Theorem 6.1 (Newman and Shapiro [39]). For $F : \mathbb{R}^m \to \mathbb{R}$ define

$$\omega(F,\delta) = \sup_{\substack{\|x\|_2,\|y\|_2\le 1\\ \|x-y\|_2\le\delta}} |F(x)-F(y)|.$$

For any $k \geq 1$ there exists a polynomial p_k of degree k with $\sup_{\|x\|_2 \leq 1} |F(x) - p_k(x)| = O(\omega(F, m/k))$.

In Section J we show how Theorem 6.1 can be recovered by FT-mollification followed by Taylor expansion to degree k.

6.2 Fooling intersections of halfspaces Our approach also implies that the intersection of halfspaces (or even degree-2 threshold functions) is fooled by bounded independence. While our main theorem implies that poly$(1/\varepsilon)$-wise independence fools GW rounding, we can obtain much better polynomial dependence by noting that to fool GW rounding it suffices to fool the intersection of two halfspaces under the Gaussian measure.

This is because in the GW rounding scheme for MaxCut, each vertex u is first mapped to a vector x_u of unit norm, and the side of a bipartition u is placed in is decided by $\mathrm{sgn}(\langle x_u, r\rangle)$ for a random Gaussian vector r. For a vertex u, let H_u^+ be the halfspace $\langle x_u, r\rangle > 0$, and let H_u^- be the halfspace $\langle -x_u, r\rangle > 0$. Then note that the edge (u, v) is cut if and only if $r \in (H_u^+ \cap H_v^-) \cup (H_u^- \cap H_v^+)$, i.e. $I_{(H_u^+ \cap H_v^-) \cup (H_u^- \cap H_v^+)}(r) = 1$. Since $H_u^+ \cap H_v^-$ and $H_u^- \cap H_v^+$ are disjoint, we have $I_{(H_u^+ \cap H_v^-) \cup (H_u^- \cap H_v^+)}(r) = I_{H_u^+ \cap H_v^-}(r) + I_{H_u^- \cap H_v^+}(r)$. Since the sum of expectations of this quantity over all edges (u, v) gives us the expected number of edges that are cut, the following theorem implies that to achieve a maximum cut within a factor $.878... - \varepsilon$ of optimal in expectation, it suffices that the entries of the random normal vector r have entries that are $\Omega(1/\varepsilon^2)$-wise independent. The proof of the theorem is in Section H.

Theorem 6.2. Let $H_i = \{x : \langle a_i, x\rangle > \theta_i\}$ for $i \in [m]$ be m halfspaces, with $\|a_i\|_2 = 1$ for all i. Let X be a vector of n i.i.d. Gaussians, and Y be a vector of k-wise independent Gaussians. Then for $k \geq \mathrm{poly}(m)/\varepsilon^2$, $|\mathbf{Pr}[X \in \cap_{i=1}^m H_i] - \mathbf{Pr}[Y \in \cap_{i=1}^m H_i]| < \varepsilon$.

The proof is essentially the same as in Section 5 and can be summarized succinctly: FT-mollify the indicator function of $R = \{z : \forall i \in [m]\ z_i \geq \theta_i\} \subset \mathbb{R}^m$, then consider the FT-mollification of I_R evaluated at $(\langle a_1, X\rangle, \ldots, \langle a_m, X\rangle)$. We also in Section H discuss how our proof of Theorem 6.2 generalizes to handle intersections of degree-2 PTF's, as well as generalizations to case that X, Y are Bernoulli vectors as opposed to Gaussian. Our dependence on m in all cases is polynomial.

We note that recent and independent work in [23] also shows that bounded independence fools intersections of halfspaces. They consider more distributions and also have a slightly lower polynomial dependence on m, though their proof is considerably more involved and also gives dependence on ε which is worse by poly-logarithmic factors.

6.3 Agnostic Learning Let $\mathcal{C}$ be a *concept class* (a set of functions $f : X \to Y$). For a distribution $\mathcal{D}$ over pairs $(x, y) \in X \times Y$ define $\mathsf{err}(f) = \mathbf{Pr}_{\mathcal{D}}[f(x) \neq y]$ and $\mathsf{opt} = \min_{f \in \mathcal{C}} \mathsf{err}(f)$. In agnostic learning, given an $\varepsilon > 0$ and independent samples from $\mathcal{D}$, the task is to efficiently compute a hypothesis h satisfying $\mathsf{err}(h) \leq \mathsf{opt} + \varepsilon$. The following theorem of Kalai et al translates low-degree polynomial approximators in L_1 to efficient agnostic learning algorithms (where the range Y is assumed to be $\{\pm 1\}$).[4]

Theorem 6.3 ([29, Theorem 5]). Suppose $\min_{\deg(p) \leq d} \mathbf{E}_{\mathcal{D}_X}[|p(x) - c(x)|] \leq \varepsilon$ for some degree $d = d(\varepsilon)$, some distribution $\mathcal{D}$ over $X \times \{-1, 1\}$ with marginal $\mathcal{D}_X$, and any $c \in \mathcal{C}$. Then there is an algorithm using $m = \mathrm{poly}(n^d/\varepsilon)$ examples $Z = \{(x_i, y_i)\}_{i=1}^m$ from $\mathcal{D}$ which runs in time poly(m) and outputs an $h \in \mathcal{C}$ satisfying $\mathbf{E}_Z[\mathsf{err}(h)] \leq \mathsf{opt} + \varepsilon$.

[4] The statement of the theorem is slightly different in [29]; they require an upper bound of ε^2 on the L_2^2 error of $p(x) - c(x)$ instead of ε on the L_1 error, but this is only used in their proof to obtain a bound on the L_1 error.

Fix $\mathcal{C}$ to be the class of halfspaces, $X = \mathbb{R}$ and $\mathcal{D}_X$ being the standard normal distribution. Note that by symmetry the n-dimensional problem can be reduced to an L_1 polynomial approximation of the sign function under the Gaussian distribution (on the real line). [29] further reduce this problem to the corresponding L_2 approximation problem (which they analyze using the Hermite polynomials) and thus get $d(\varepsilon) = O(1/\varepsilon^4)$. This bound can be further improved by directly constructing L_1 approximators. The current work improves the bound to $O(1/\varepsilon^2)$ (which is optimal up to constant factors [20]). This is achieved by FT-mollifying and Taylor-expanding to degree $O(1/\varepsilon^2)$, as done in Section 5. Our proof does not require specific properties of the distribution $\mathcal{D}_X$ beyond moment bounds (to bound the error from Taylor's theorem), and anticoncentration bounds (to bound $\mathbf{Pr}_{x \sim \mathcal{D}_X}[|x - \theta| \le \varepsilon]$ for $\theta \in \mathbb{R}$). Thus, in addition to improving the degree bound, our proof also generalizes readily to other distributions.

6.4 Streaming algorithms The introduction of the FT-mollification technique in [30] was made to obtain an optimal algorithm for moment estimation in data streams, a problem first studied by Alon, Matias, and Szegedy [2]. In this problem, a vector $x \in \mathbb{R}^n$ receives m updates $(i_1, v_1), \ldots, (i_m, v_m)$, with update (i, v) causing the change $x_i \leftarrow x_i + v$. At the end of the stream, we must output a $(1 \pm \varepsilon)$ approximation to the value $F_p = \sum_i |x_i|^p$ with 2/3 probability. One goal is to use as little space as possible to process the stream. The work of [30] showed that Indyk's algorithm [27] is fooled by bounded independence, using FT-mollification, yielding a space-optimal algorithm for F_p-estimation for all real $0 < p < 2$. Specifically, [30] showed $\mathrm{polylog}(1/\varepsilon)/\varepsilon^p$-wise independence suffices. Plugging our new FT-mollification theorem into the argument of [30] shows that in fact $\Omega(1/\varepsilon^p)$-wise independence suffices (the $\mathrm{polylog}(1/\varepsilon)$ factors are removed).

7 Proof of Theorem 1.1

We now give our proof of Theorem 1.1. In Section 7.1 we state a central moment bound we use for quadratic forms with small eigenvalues. Section 7.2 analyzes the regular case of our main theorem, and Section 7.3 reduces the general case to the regular case.

7.1 A spectral moment bound for quadratic forms For a quadratic form $p(x) = \sum_{i \le j} a_{i,j} x_i x_j$, we can associate a real symmetric matrix A_p which has the $a_{i,i}$ on the diagonals and $a_{\min\{i,j\},\max\{i,j\}}/2$ on the offdiagonals, so that $p(x) = x^T A_p x$. Our proof of Theorem 1.1 makes use of a moment bound for quadratic forms which takes into account the maximum eigenvalue of A_p. We give a proof of this moment bound which builds upon ideas of Whittle [50], who showed the hypercontractive inequality for degree-2 polynomials when comparing q-norms to 2-norms (see Theorem D.1). We learned recently that another proof of this moment bound can be derived from a tail inequality for quadratic forms given in [25]. We provide our own proof to be self-contained.

Recall the *Frobenius norm* of $A \in \mathbb{R}^{n \times n}$ is $\|A\|_2 = \sqrt{\sum_{i,j=1}^{n,n} A_{i,j}^2} = \sqrt{\sum_i \lambda_i^2} = \sqrt{\mathrm{tr}(A^2)}$, where tr denotes trace and A has eigenvalues $\lambda_1, \ldots, \lambda_n$. Define $\|A\|_\infty$ to be the largest magnitude of an eigenvalue of A. We can now state the main theorem of this section, which plays a crucial role in our analysis of the regular case of Theorem 1.1. Our proof can be found in Section B.

Theorem 7.1. Let $A \in \mathbb{R}^{n \times n}$ be symmetric and $x \in \{-1, 1\}^n$ be random. Then for all $k \ge 2$, $\mathbf{E}[|(x^T A x) - \mathrm{tr}(A)|^k] \le C^k \cdot \max\{\sqrt{k}\|A\|_2, k\|A\|_\infty\}^k$, where C is an absolute constant.

Note if $\sum_{i \le j} a_{i,j}^2 \le 1$ then $\|A_p\|_\infty \le 1$, in which case Theorem 7.1 recovers a similar moment bound as the one obtained via hypercontractivity. Thus, in the special case of bounding kth moments of degree-2 polynomials against their 2nd moment, Theorem 7.1 can be viewed as a generalization of the hypercontractive inequality (and of Whittle's inequality).

7.2 Fooling regular degree-2 threshold functions In this section we show the following.

Theorem 7.2. Let $0 < \varepsilon < 1$ be given. Let $X_1, \ldots, X_n$ be independent Bernoulli and $Y_1, \ldots, Y_n$ be $2k$-wise independent Bernoulli for k a sufficiently large multiple of $1/\varepsilon^8$. If p is multilinear and of degree 2 with $\sum_{|S|>0} \widehat{p}_S^2 = 1$, and $\mathrm{Inf}_i(p) \le \tau$ for all i, then

$$\mathbf{E}[\mathrm{sgn}(p(X))] - \mathbf{E}[\mathrm{sgn}(p(Y))] = O(\varepsilon + \tau^{1/9}).$$

Throughout this section, p always refers to the polynomial of Theorem 7.2, and τ refers to the maximum influence of any variable in p. Observe p (over the hypercube) can be written as $q+p_4+C$, where q is a multilinear quadratic form, p_4 is a linear form, and C is a constant. Furthermore, $\|A_q\|_2 \le 1/2$ and $\sum_S \widehat{p_4}_S^2 \le 1$. Using the spectral theorem for real symmetric matrices, we write $p = p_1 - p_2 + p_3 + p_4 + C$ where p_1, p_2, p_3 are quadratic forms satisfying $\lambda_{\min}(A_{p_1}), \lambda_{\min}(A_{p_2}) \ge \delta$, $\|A_{p_3}\|_\infty < \delta$, and $\|A_{p_i}\|_2 \le 1/2$ for $1 \le i \le 3$, and also with p_1, p_2 positive semidefinite (see Lemma D.7 for details on how this is accomplished). Here $\lambda_{\min}(A)$ denotes the smallest magnitude of a non-zero eigenvalue of A. Throughout this section we let $p_1, \ldots, p_4$, C, δ be as discussed here. We use Υ to denote $\mathrm{tr}(A_{p_3})$. The value δ will be set later in the proof of Theorem 7.2.

It will be convenient to define the map $M_p : \mathbb{R}^n \to \mathbb{R}^4$ for $M_p(x) = (\sqrt{p_1(x)}, \sqrt{p_2(x)}, p_3(x) - \Upsilon, p_4(x))$. Note the the first two coordinates of $M_p(x)$ are indeed real since p_1, p_2 are positive semidefinite. To show Theorem 7.2, we follow the template of Section 5, by showing that $\mathbf{E}[I_R(M_p(X))]$ is determined by k-wise independence for $R = \{z : z_1^2 - z_2^2 + z_3 + z_4 + C + \Upsilon \ge 0\} \subset \mathbb{R}^4$ (note $I_R(M_p(x))$ iff $p(x) \ge 1$).

Before giving the proof of Theorem 7.2, we first state Lemma 7.3, which states that for $F : \mathbb{R}^4 \to \mathbb{R}$, $F(M_p(x))$ is fooled by bounded independence as long as F is even in x_1, x_2 and certain technical conditions are satisfied. The proof of Lemma 7.3 essentially follows from Taylor's theorem, using moment bounds on linear and quadratic forms to bound the error term. We provide a proof sketch here, and the full proof can be found in Section F.

Lemma 7.3. Let $\varepsilon > 0$ be arbitrary. Let $F : \mathbb{R}^4 \to \mathbb{R}$ be even in each of its first two arguments such that $\|\partial^\beta \tilde{F}^c\|_\infty = O(\alpha^{|\beta|})$ for all multi-indices $\beta \in \mathbb{N}^4$ and some $\alpha > 1$. Suppose $1/\delta \ge B\alpha$ for a sufficiently large constant B. Let $X_1, \ldots, X_n$ be independent Bernoulli, and $Y_1, \ldots, Y_n$ be k'-independent Bernoulli for $k' = 2k$ with $k \ge \max\{\log(1/\varepsilon), B\alpha/\sqrt{\delta}, B\alpha^2\}$ an even integer. Write $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$. Then $|\mathbf{E}[F(M_p(X))] - \mathbf{E}[F(M_p(Y))]| < \varepsilon$.

Proof (Sketch). Taylor-expand F to obtain a polynomial P_{k-1} containing all monomials up to degree $k-1$. Since $F(x)$ is even in x_1, x_2, we can assume P_{k-1} is a polynomial in x_1^2, x_2^2, x_3, x_4. Let $x \in \mathbb{R}^4$ be arbitrary. Via Taylor's theorem one can show $|\mathbf{E}[F(M_p(X))] - \mathbf{E}[F(M_p(Y))]| \le \alpha^k 2^{O(k)} \cdot (\mathbf{E}[(p_1(X))^{k/2}] + \mathbf{E}[(p_2(X))^{k/2}] + \mathbf{E}[(p_3(X) - \Upsilon)^k] + \mathbf{E}[(p_4(X))^k])/k^k$, since $\mathbf{E}[P_{k-1}(M_p(X))] = \mathbf{E}[P_{k-1}(M_p(Y))]$ via sufficient independence. The lemma follows by applying standard moment bounds to p_1, p_2, p_4, and the bound of Theorem 7.1 to $p_3 - \Upsilon$. ■

In proving Theorem 7.2, we will need a lemma which states that p is anticoncentrated even when evaluated on Bernoulli random variables which are k-wise independent. We state the necessary lemma and provide a proof sketch here; the full proof is in Section F.

Lemma 7.4. For $\varepsilon' > 0$, let $k \ge D/(\varepsilon')^4$ for a universal constant $D > 0$. Let $Y_1, \ldots, Y_n$ be k-wise independent Bernoulli, and let $t \in \mathbb{R}$ be arbitrary. Then $\mathbf{Pr}[|p(Y) - t| < \varepsilon'] \le O(\sqrt{\varepsilon'} + \tau^{1/9})$.

Proof (Sketch). The proof in spirit works similarly to the proof in Section 5 of anticoncentration under bounded independence of linear forms (specifically, step (C) in the proof of Theorem 5.1). We define the region $T_{t,\varepsilon'} = \{z : |z_1^2 - z_2^2 + z_3 + z_4 + C + \Upsilon - t| < \varepsilon'\} \subset \mathbb{R}^4$ and note $\mathbf{Pr}[|p(Y) - t| < \varepsilon'] = \mathbf{E}[I_{T_{t,\varepsilon'}}(M_p(Y))]$. Then, just as when proving (C) of Theorem 5.1, we would like a smooth function f which upper bounds $I_{T_{t,\varepsilon'}}$ and has small expectation under full independence, so that we may apply Taylor's theorem (specifically, Lemma 7.3) to show that its expectation is also small under bounded independence. To accomplish this, we define the region $S_{\rho,t,\varepsilon'} = \{z : d_2(z, T_{t,\varepsilon'}) \le \rho\}$ then take f to be $2 \cdot \tilde{I}^c_{S_{\rho,t,\varepsilon'}}$ for appropriately selected $\rho, c > 0$. ■

The following lemma follows from the Invariance Principle, the hypercontractive inequality, and Lemma 7.4. The proof is in Section F.

Lemma 7.5. Let $\eta, \eta' \ge 0$ be given, and let $Y_1, \ldots, Y_n$ be k-independent Bernoulli for k as in Lemma 7.4 with $\varepsilon' = \min\{\eta/\sqrt{\delta}, \eta'\}$. Also assume $k \ge \lceil 2/\delta \rceil$. Then

$$\mathbf{Pr}[|p(X) - t| \le \eta \cdot (\sqrt{p_1(X)} + \sqrt{p_2(X)} + 1) + \eta'] = O(\sqrt{\eta'} + (\eta^2/\delta)^{1/4} + \tau^{1/9} + \exp(-\Omega(1/\delta))).$$

We are now ready to prove the main theorem of this section.

Proof (of Theorem 7.2). Consider the region $R \subset \mathbb{R}^4$ defined by $R = \{z : z_1^2 - z_2^2 + z_3 + z_4 + C + \Upsilon \ge 0\}$. Then note that $I_{[0,\infty)}(p(x)) = 1$ if and only if $I_R(M_p(x)) = 1$. It thus suffices to show that I_R is fooled in expectation by bounded independence.

We set $\rho = \varepsilon^4$, $c = 1/\rho$, and $1/\delta = 2Bc$ for B the constant in the statement of Lemma 7.3. We now show a chain of inequalities to give our theorem:

$$\mathbf{E}[I_R(M_p(X))] \approx_{\varepsilon+\tau^{1/9}} \mathbf{E}[\tilde{I}^c_R(M_p(X))] \approx_{\varepsilon} \mathbf{E}[\tilde{I}^c_R(M_p(Y))] \approx_{\varepsilon+\tau^{1/9}} \mathbf{E}[I_R(M_p(Y))]$$

$\mathbf{E[I_R(M_p(X))] \approx_{\varepsilon+\tau^{1/9}} E[\tilde{I}^c_R(M_p(X))]}$: First, note

$$d_2(z, \partial R) \ge \frac{1}{2} \cdot \min\left\{\frac{|z_1^2 - z_2^2 + z_3 + z_4 + C + \Upsilon|}{2(|z_1| + |z_2| + 1)}, \sqrt{|z_1^2 - z_2^2 + z_3 + z_4 + C + \Upsilon|}\right\}.$$

This is because by adding a vector v to z, we can change each individual coordinate of z by at most $\|v\|_2$, and can thus change the value of $|z_1^2 - z_2^2 + z_3 + z_4 + C + \Upsilon - t| - \varepsilon'$ by at most $2\|v\|_2 \cdot (|z_1| + |z_2| + 1) + \|v\|_2^2$. Applying Lemma 7.5,

$$\begin{aligned}\mathbf{Pr}[d_2(M_p(X), \partial R) \le w] &\le \mathbf{Pr}[|p(X)| \le 4w \cdot (\sqrt{p_1(X)} + \sqrt{p_2(X)} + 1)] + \mathbf{Pr}[|p(X)| \le 4w^2] \\ &= O(w + \sqrt{w} + (w^2/\delta)^{1/4} + \tau^{1/9} + \exp(-\Omega(1/\delta)))\end{aligned}$$

Now, noting $|\mathbf{E}[I_R(M_p(X))] - \mathbf{E}[\tilde{I}^c_R(M_p(X))]| \le \mathbf{E}[|I_R(M_p(X))] - \tilde{I}^c_R(M_p(X))|]$,

$$\begin{aligned}&|\mathbf{E}[I_R(M_p(X))] - \mathbf{E}[\tilde{I}^c_R(M_p(X))]| \\ &\quad\le \mathbf{Pr}[d_2(M_p(X), \partial R) \le 2\rho] + O\left(\sum_{s=1}^{\infty} 2^{-2s} \cdot \mathbf{Pr}[2^s\rho < d_2(M_p(X), \partial R) \le 2^{s+1}\rho]\right) \\ &\quad\le O(\sqrt{\rho} + (\rho^2/\delta)^{1/4} + \tau^{1/9} + \exp(-\Omega(1/\delta)) \\ &\qquad + O\left(\sum_{s=1}^{\infty} 2^{-2s} \cdot (\sqrt{2^{s+1}\rho} + (2^{2s+2}\rho^2/\delta)^{1/4} + \tau^{1/9} + \exp(-\Omega(1/\delta)))\right) = O(\varepsilon + \tau^{1/9})\end{aligned}$$

by choice of ρ, δ and applications of Theorem 4.2 and Lemma 7.5.

$\mathbf{E}[\tilde{\mathbf{I}}^{\mathbf{c}}_{\mathbf{R}}(\mathbf{M}_{\mathbf{p}}(\mathbf{X}))] \approx_{\varepsilon} \mathbf{E}[\tilde{\mathbf{I}}^{\mathbf{c}}_{\mathbf{R}}(\mathbf{M}_{\mathbf{p}}(\mathbf{Y}))]$: As in Eq. (F.5), we can assume $\tilde{I}^c_R$ is even in x_1, x_2. We apply Lemma 7.3 with $\alpha = 2c$, noting that $1/\delta = B\alpha$ and that our setting of k is sufficiently large.

$\mathbf{E}[\tilde{\mathbf{I}}^{\mathbf{c}}_{\mathbf{R}}(\mathbf{M}_{\mathbf{p}}(\mathbf{Y}))] \approx_{\varepsilon+\tau^{1/9}} \mathbf{E}[\mathbf{I}_{\mathbf{R}}(\mathbf{M}_{\mathbf{p}}(\mathbf{Y}))]$: The argument is identical as with the first inequality. We remark that we do have sufficient independence to apply Lemma 7.5 since, mimicking our analysis of the first inequality, we have

$$\begin{aligned} &\mathbf{Pr}[|p(Y)| \leq 4\rho \cdot (\sqrt{p_1(Y)} + \sqrt{p_2(Y)} + 1)] + \mathbf{Pr}[|p(Y)| \leq 4\rho^2] \\ &\quad \leq \mathbf{Pr}[|p(Y)| \leq 4\rho \cdot (\sqrt{p_1(Y)} + \sqrt{p_2(Y)} + 1)] + \mathbf{Pr}[|p(Y)| \leq \varepsilon^2] \end{aligned} \tag{7.1}$$

since $\rho^2 = o(\varepsilon^2)$ (we only changed the second summand). To apply Lemma 7.5 to Eq. (7.1), we need $k \geq \lceil 2/\delta \rceil$, which is true, and $k = \Omega(1/(\varepsilon'')^4)$, for $\varepsilon'' = \min\{\rho/\sqrt{\delta}, \varepsilon^2\} = \varepsilon^2$, which is also true. Lemma 7.5 then tells us Eq. (7.1) is $O(\varepsilon + \tau^{1/9})$.

■

Our main theorem of this Section (Theorem 7.2) also holds under the case that the X_i, Y_i are standard normal, and without any error term depending on τ. We give a proof in Section F.2, by reducing back to the Bernoulli case.

7.3 Reduction to the regular case

In this section, we complete the proof of Theorem 1.1. We accomplish this by providing a reduction from the general case to the regular case. In fact, such a reduction can be shown to hold for any degree $d \geq 1$ and establishes the following:

Theorem 7.6. Suppose K_d-wise independence ε-fools the class of τ-regular degree-d PTF's, for some parameter $0 < \tau \leq \varepsilon$. Then $(K_d + L_d)$-wise independence ε-fools all degree-d PTFs, where $L_d = (1/\tau) \cdot \big(d \log(1/\tau)\big)^{O(d)}$.

Noting that τ-regularity implies that the maximum influence of any particular variable is at most $d \cdot \tau$, Theorem 7.2 implies that degree-2 PTF's that are τ-regular, for $\tau = O(\varepsilon^9)$, are ε-fooled by K_2-wise independence for $K_2 = O(\varepsilon^{-8}) = \mathrm{poly}(1/\varepsilon)$. By plugging in $\tau = O(\varepsilon^9)$ in the above theorem we obtain Theorem 1.1. The proof of Theorem 7.6 is obtained by a simple adaptation of the regularity lemma in [17][5]. Here we give a sketch, with details in Section G.

Proof (Sketch). (of Theorem 7.6). Any boolean function f on $\{-1,1\}^n$ can be expressed as a binary decision tree where each internal node is labeled by a variable, every root-to-leaf path corresponds to a restriction ρ that fixes the variables as they are set on the path, and every leaf is labeled with the restricted subfunction f_ρ. The main claim is that, if f is a degree-d PTF, then it has such a decision-tree representation with certain strong properties. In particular, given an arbitrary degree-d PTF $f = \mathrm{sgn}(p)$, by [17] there exists a decision tree $\mathcal{T}$ of depth $(1/\tau) \cdot \big(d \log(1/\tau)\big)^{O(d)}$, so that with probability $1 - \tau$ over the choice of a random root-to-leaf path[6] ρ, the restricted subfunction (leaf) $f_\rho = \mathrm{sgn}(p_\rho)$ is either a τ-regular degree-d PTF or τ-close to a constant function.

Our proof of Theorem 7.6 is based on the above structural lemma. Under the uniform distribution, there is some particular distribution on the leaves (the tree is not of uniform height); then conditioned on the restricted variables the variables still undetermined at the leaf are still uniform.

[5] We note that [36] prove a very similar regularity lemma to obtain their PRGs for degree-d PTF's. One could alternatively use this instead of [17]. For $d = 2$ this would give a worse bound of $\widetilde{\Omega}(\varepsilon^{-18})$.

[6] A "random root-to-leaf path" corresponds to the standard uniform random walk on the tree.

With $(K_d + L_d)$-wise independence, a random walk down the tree arrives at each leaf with the same probability as in the uniform case (since the depth of the tree is at most L_d). Hence, the probability mass of the "bad" leaves is at most $\tau \leq \varepsilon$ even under bounded independence. Furthermore, the induced distribution on each leaf (over the unrestricted variables) is K_d-wise independent. Consider a good leaf. Either the leaf is τ-regular, in which case we can apply Theorem 7.2, or it is τ-close to a constant function. At this point though we arrive at a technical issue. The statement and proof in [17] concerning "close-to-constant" leaves holds only under the uniform distribution. For our result, we need a stronger statement that holds under any distribution (on the variables that do not appear in the path) that has sufficiently large independence. By simple modifications of the proof in [17], we show that the statement holds even under $O(d \cdot \log(1/\tau))$-wise independence. ■

Acknowledgments

We thank Piotr Indyk and Rocco Servedio for comments that improved the presentation of this work, Ryan O'Donnell for bringing our attention to the problem of the intersection of threshold functions, Assaf Naor for bringing our attention to the reference [25], and Michael Ganzburg for useful email correspondence.

References

[1] Noga Alon, László Babai, and Alon Itai. A fast and simple randomized parallel algorithm for the maximal independent set problem. *J. Algorithms*, 7(4):567–583, 1986.

[2] Noga Alon, Yossi Matias, and Mario Szegedy. The Space Complexity of Approximating the Frequency Moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

[3] James Aspnes, Richard Beigel, Merrick L. Furst, and Steven Rudich. The expressive power of voting polynomials. *Combinatorica*, 14(2):1–14, 1994.

[4] Per Austrin and Johan Håstad. Randomly supported independence and resistance. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 483–492, 2009.

[5] Louay Bazzi. Polylogarithmic independence can fool DNF formulas. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 63–73, 2007.

[6] William Beckner. Inequalities in Fourier analysis. *Annals of Mathematics*, 102(1):159–182, 1975.

[7] Richard Beigel. Perceptrons, PP, and the Polynomial Hierarchy. *Computational Complexity*, 4:339–349, 1994.

[8] Ido Ben-Eliezer, Shachar Lovett, and Ariel Yadin. Polynomial threshold functions: Structure, approximation and pseudorandomness. *CoRR*, abs/0911.3473, 2009.

[9] Itai Benjamini, Ori Gurel-Gurevich, and Ron Peled. On k-wise independent distributions and boolean functions. Available at `http://www.wisdom.weizmann.ac.il/~origurel/`, 2007.

[10] Aline Bonami. Étude des coefficients de Fourier des fonctions de $L^p(G)$. *Ann. Inst. Fourier*, 20:335–402, 1970.

[11] Mark Braverman. Poly-logarithmic independence fools AC^0 circuits. In *Proceedings of the 24th Annual IEEE Conference on Computational Complexity (CCC)*, pages 3–8, 2009.

[12] Jehoshua Bruck. Harmonic analysis of polynomial threshold functions. *SIAM J. Discrete Math.*, 3(2):168–177, 1990.

[13] Jehoshua Bruck and Roman Smolensky. Polynomial threshold functions, AC^0 functions and spectral norms. *SIAM J. Comput.*, 21(1):33–42, 1992.

[14] Anthony Carbery and James Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical Research Letters*, 8(3):233–248, 2001.

[15] Benny Chor and Oded Goldreich. On the power of two-point based sampling. *Journal of Complexity*, 5(1):96–106, March 1989.

[16] Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco A. Servedio, and Emanuele Viola. Bounded independence fools halfspaces. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 171–180, 2009.

[17] Ilias Diakonikolas, Rocco A. Servedio, Li-Yang Tan, and Andrew Wan. A regularity lemma, and low-weight approximators, for low-degree polynomial threshold functions. In *Proceedings of the 25th Annual IEEE Conference on Computational Complexity (CCC), to appear*, 2010. CoRR abs/0909.4727.

[18] Kurt Otto Friedrichs. The identity of weak and strong extensions of differential operators. *Transactions of the American Mathematical Society*, 55(1):132–151, 1944.

[19] Michael I. Ganzburg. personal communication.

[20] Michael I. Ganzburg. *Limit theorems of polynomial approximation with exponential weights.* Memoirs of the American Mathematical Society, 2008.

[21] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42:1115–1145, 1995.

[22] Mikael Goldmann, Johan Håstad, and Alexander A. Razborov. Majority gates vs. general weighted threshold gates. *Computational Complexity*, 2:277–300, 1992.

[23] Parikshit Gopalan, Ryan O'Donnell, Yi Wu, and David Zuckerman. Fooling functions of halfspaces under product distributions. In *Proceedings of the 25th Annual IEEE Conference on Computational Complexity (CCC), to appear*, 2010.

[24] András Hajnal, Wolfgang Maass, Pavel Pudlák, Mario Szegedy, and György Turán. Threshold circuits of bounded depth. *J. Comput. Syst. Sci.*, 46:129–154, 1993.

[25] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Statist.*, 42(3):1079–1083, 1971.

[26] Prahladh Harsha, Adam Klivans, and Raghu Meka. An invariance principle for polytopes. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing (STOC), to appear*, 2010.

[27] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM*, 53(3):307–323, 2006.

[28] Steven G. Johnson. Saddle-point integration of C_∞ "bump" functions. Manuscript. Available at `http://math.mit.edu/~stevenj/bump-saddle.pdf`.

[29] Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. *SIAM J. Comput.*, 37(6):1777–1805, 2008.

[30] Daniel M. Kane, Jelani Nelson, and David P. Woodruff. On the exact space complexity of sketching and streaming small norms. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1161–1178, 2010.

[31] Zohar Karnin, Yuval Rabani, and Amir Shpilka. Explicit dimension reduction and its applications. *Electronic Colloquium on Computational Complexity (ECCC)*, (121), 2009.

[32] Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning intersections and thresholds of halfspaces. *J. Comput. Syst. Sci.*, 68(4):808–840, 2004.

[33] Adam R. Klivans and Rocco A. Servedio. Learning DNF in time $2^{\tilde{O}(n^{1/3})}$. *J. Comput. Syst. Sci.*, 68(2):303–318, 2004.

[34] Matthias Krause and Pavel Pudlák. Computing boolean functions by polynomials and threshold circuits. *Computational Complexity*, 7(4):346–370, 1998.

[35] Sanjeev Mahajan and Ramesh Hariharan. Derandomizing semidefinite programming based approximation algorithms. In *Proceedings of the 36th Symposium on Foundations of Computer Science (FOCS)*, pages 162–169, 1995.

[36] Raghu Meka and David Zuckerman. Pseudorandom generators for polynomial threshold functions. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing (STOC), to appear (see also CoRR abs/0910.4122)*, 2010.

[37] Marvin A. Minsky and Seymour L. Papert. *Perceptrons.* MIT Press, Cambridge, MA, 1969 (expanded edition 1988).

[38] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. *Annals of Mathematics*, 171(1):295–341, 2010.

[39] D.J. Newman and H.S. Shapiro. Jackson's theorem in higher dimensions. *On approximation theory (Proc. Conf. Oberwolfach 1963)*, MR 32(310):208–219, 1964.

[40] Noam Nisan. Pseudorandom bits for constant depth circuits. *Combinatorica*, 11(1):63–70, 1991.

[41] Noam Nisan. The communication complexity of threshold gates. In *Proceedings of Combinatorics, Paul Erdös is Eighty*, pages 301–315, 1994.

[42] Ryan O'Donnell and Rocco A. Servedio. Extremal properties of polynomial threshold functions. *J. Comput. Syst. Sci.*, 74(3):298–312, 2008.

[43] Yuval Rabani and Amir Shpilka. Explicit construction of a small epsilon-net for linear threshold functions. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 649–658, 2009.

[44] Alexander A. Razborov. A simple proof of Bazzi's theorem. *ACM Transactions on Computation Theory*, 1(1), 2009.

[45] Alexander A. Razborov and Alexander A. Sherstov. The sign-rank of AC^0. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 57–66, 2008.

[46] Michael E. Saks. *Slicing the hypercube*, pages 211–257. London Mathematical Society Lecture Note Series 187, 1993.

[47] Alexander A. Sherstov. The intersection of two halfspaces has high threshold degree. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2009.

[48] D. Sivakumar. Algorithmic derandomization via complexity theory. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, pages 619–626, 2002.

[49] Gilbert Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 4th edition, 2009.

[50] Peter Whittle. Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.*, 5(3):302–305, 1960.

Appendix

A FT-mollification proofs

In the proof sketch of Theorem 4.1 in Section 4, we required the existence of a probability density B on $\mathbb{R}^d$ satisfying $\mathbf{E}_{x\sim B}[\|x\|_2^2] = O(d^2)$, and $\|\partial^\beta B\|_1 \le 2^{|\beta|}$ for all $\beta \in \mathbb{N}^d$. In this section, we show that such a B exists.

Definition A.1. In *hyperspherical coordinates* in $\mathbb{R}^d$, we represent a point $x = (x_1, \ldots, x_d)$ by $x_i = r\cos(\phi_i)\prod_{j=1}^{i-1}\sin(\phi_j)$ for $i < d$, and $x_d = r\prod_{j=1}^{d-1}\sin(\phi_j)$. Here $r = \|x\|_2$ and the ϕ_i satisfy $0 \le \phi_i \le \pi$ for $i < d-1$, and $0 \le \phi_{d-1} < 2\pi$.

Fact A.2. Let J be the Jacobian matrix corresponding to the change of variables from Cartesian to hyperspherical coordinates. Then

$$\det(J) = r^{d-1}\prod_{i=1}^{d-2}\sin^{d-1-i}(\phi_i).$$

We define the function $b : \mathbb{R}^d \to \mathbb{R}$ by

$$b(x) = \sqrt{C_d}\cdot\begin{cases} 1 - \|x\|_2^2 & \text{for } \|x\|_2 < 1 \\ 0 & \text{otherwise}\end{cases}.$$

The value C_d is chosen so that $\|b\|_2 = 1$. Note b is not smooth (its mixed partials do not exist on the unit sphere), but we will only ever need that $\frac{\partial}{\partial x_i}b$ is square-integrable for all $i \in [d]$.

Henceforth, we make the setting

$$A_d = C_d \cdot \int_0^{2\pi}\int_{[0,\pi]^{d-2}}\left(\prod_{i=1}^{d-2}\sin^{d-1-i}(\phi_i)\right)d\phi_1 d\phi_2\cdots d\phi_{d-1}.$$

We let $\hat{b} : \mathbb{R}^d \to \mathbb{R}$ denote the Fourier transform of b, i.e.

$$\hat{b}(t) = \frac{1}{(\sqrt{2\pi})^d}\int_{\mathbb{R}^d} b(x)e^{-i\langle x,t\rangle}dx.$$

We note $\hat{b}$ does in fact take real values, since $\widehat{(b^*)}(t) = (\hat{b})^*(-t)$ and $\hat{b}$ is symmetric about the origin (here b^* denotes complex conjugation of b). Finally, we define $B = \hat{b}^2$.

Lemma A.3. B is a probability density on $\mathbb{R}^d$.

Proof. B is nonnegative since it is the square of a real function. Also, $\int_{\mathbb{R}^d} B(x)dx = \|\hat{b}\|_2^2$, which equals $\|b\|_2^2 = 1$ by Plancherel's theorem. ■

We now bound the L_1 norm of mixed partials of B.

Lemma A.4. For any $\beta \in \mathbb{N}^d$, $\|\partial^\beta B\|_1 \le 2^{|\beta|}$.

Proof. We have

$$\partial^\beta B = \sum_{\alpha\le\beta}\binom{\beta}{\alpha}\left(\partial^\alpha\hat{b}\right)\cdot\left(\partial^{\beta-\alpha}\hat{b}\right)$$

Thus,

$$\begin{aligned}
\left\|\partial^\beta B\right\|_1 &= \left\|\sum_{\alpha\le\beta}\binom{\beta}{\alpha}\left(\partial^\alpha\hat{b}\right)\cdot\left(\partial^{\beta-\alpha}\hat{b}\right)\right\|_1 \\
&\le \sum_{\alpha\le\beta}\binom{\beta}{\alpha}\left\|\partial^\alpha\hat{b}\right\|_2\cdot\left\|\partial^{\beta-\alpha}\hat{b}\right\|_2 && \text{(A.1)}\\
&= \sum_{\alpha\le\beta}\binom{\beta}{\alpha}\left\|x^\alpha\cdot b\right\|_2\cdot\left\|x^{\beta-\alpha}\cdot b\right\|_2 && \text{(A.2)}\\
&\le \sum_{\alpha\le\beta}\binom{\beta}{\alpha} && \text{(A.3)}\\
&= 2^{|\beta|} && \text{(A.4)}
\end{aligned}$$

Eq. (A.1) follows by Cauchy-Schwarz. Eq. (A.2) follows from Plancherel's theorem, since the Fourier transform of $\partial^\alpha\hat{b}$ is $x^\alpha\cdot b$, up to factors of i. Eq. (A.3) follows since $\|x^\alpha\cdot b\|_2\le\|b\|_2=1$. Eq. (A.4) is seen combinatorially. Suppose we have $2d$ buckets A_i^j for $(i,j)\in[d]\times[2]$. We also have $|\beta|$ balls, with each having one of d types with β_i balls of type i. Then the number of ways to place balls into buckets such that balls of type i only go into some A_i^j is $2^{|\beta|}$ (each ball has 2 choices). However, it is also $\sum_{\alpha\le\beta}\binom{\beta}{\alpha}$, since for every placement of balls we must place some number α_i balls of type i in A_i^1 and $\beta_i-\alpha_i$ balls in A_i^2. ■

Finally, we show the desired variance bound.

Lemma A.5. $\mathbf{E}_{x\sim B}[\|x\|_2^2]=d(d+4)/2$.

Proof. Write

$$S=\mathbf{E}_{x\sim B}[\|x\|_2^2]=\int_{\mathbb{R}^d}\|x\|_2^2\cdot B(x)dx=\sum_{i=1}^d\left(\int_{\mathbb{R}^d}x_i^2\cdot B(x)dx\right).$$

Recalling that $B=\hat{b}^2$, the Fourier transform of B is $(2\pi)^{-d/2}(b*b)$. The above integral is $(2\pi)^{d/2}$ times the Fourier transform of $x_i^2\cdot B$, evaluated at 0. Since multiplying a function by $i\cdot x_j$ corresponds to partial differentiation by x_j in the Fourier domain,

$$S=\sum_{i=1}^d\left(\frac{\partial^2}{\partial x_i^2}(b*b)\right)(0)=\sum_{i=1}^d\left(\left(\frac{\partial}{\partial x_i}b\right)*\left(\frac{\partial}{\partial x_i}b\right)\right)(0)=\sum_{i=1}^d\left\|\frac{\partial}{\partial x_i}b\right\|_2^2$$

with the last equality using that $\frac{\partial}{\partial x_i}b$ is odd.

We have, for x in the unit ball,

$$\left(\frac{\partial}{\partial x_i}b\right)(x)=-2x_i$$

so that, after switching to hyperspherical coordinates,

$$\sum_{i=1}^d\left\|\frac{\partial}{\partial x_i}b\right\|_2^2=A_d\cdot\int_0^1 4r^{d+1}dr. \tag{A.5}$$

Now, by definition of b,

$$\begin{aligned}\|b\|_2^2 &= A_d \cdot \int_0^1 r^{d-1} + r^{d+3} - 2r^{d+1} dr \\ &= A_d \cdot \frac{8}{d(d+2)(d+4)}\end{aligned}$$

We also have by Eq. (A.5) that

$$\sum_{i=1}^{d} \left\| \frac{\partial}{\partial x_i} b \right\|_2^2 = A_d \cdot \frac{4}{d+2}.$$

The claim follows since $\|b\|_2^2 = 1$. ■

We now give a full proof of Theorem 4.1.

Theorem 4.1 (restatement). *Let $F : \mathbb{R}^d \to \mathbb{R}$ be bounded and $c > 0$ be arbitrary. Then there exists $\tilde{F}^c : \mathbb{R}^d \to \mathbb{R}$ satisfying*

i. $\|\partial^\beta \tilde{F}^c\|_\infty \le \|F\|_\infty \cdot (2c)^{|\beta|}$ *for all* $\beta \in \mathbb{N}^d$.

ii. Fix some $x \in \mathbb{R}^d$. *Then if* $|F(x) - F(y)| \le \varepsilon$ *whenever* $\|x - y\|_2 \le \delta$ *for some* $\varepsilon, \delta \ge 0$, *then* $|\tilde{F}^c(x) - F(x)| \le \varepsilon + \|F\|_\infty \cdot O(d^2/(c^2\delta^2))$.

iii. $\tilde{F}^c$ *is nonnegative if* F *is nonnegative.*

Proof. Item (iii) follows since for F nonnegative, $\tilde{F}^c$ is the convolution of two nonnegative functions.

For (i),

$$\begin{aligned}\left|\left(\partial^\beta \tilde{F}^c\right)(x)\right| &= \left|\left(\partial^\beta (B_c * F)\right)(x)\right| \\ &= \left|\left(\left(\partial^\beta B_c\right) * F\right)(x)\right| \\ &= \left|\int_{\mathbb{R}^d} \left(\partial^\beta B_c\right)(y) F(x-y) dy\right| \\ &\le \|F\|_\infty \cdot \left\|\partial^\beta B_c\right\|_1 \\ &= \|F\|_\infty \cdot c^{|\beta|} \cdot \left\|\partial^\beta B\right\|_1 \qquad \text{(A.6)} \\ &\le \|F\|_\infty \cdot (2c)^{|\beta|}\end{aligned}$$

For (ii),

$$\begin{aligned}\tilde{F}^c(x) &= (B_c * F)(x) \\ &= \int_{\mathbb{R}^d} B_c(x-y) F(y) dy \\ &= F(x) + \int_{\mathbb{R}^d} (F(y) - F(x)) B_c(x-y) dy \qquad \text{(A.7)} \\ &= F(x) + \int_{\|x-y\|_2 < \delta} (F(y) - F(x)) B_c(x-y) + \int_{\|x-y\|_2 \ge \delta} (F(y) - F(x)) B_c(x-y)\end{aligned}$$

where Eq. (A.7) uses that $\int_{\mathbb{R}^d} B_c(x-y)dy = \int_{\mathbb{R}^d} B(y)dy = 1$ (recall B is a probability density). The first integral above is at most $\varepsilon \cdot \int_{\mathbb{R}^d} B_c(x-y)dy = \varepsilon$ in magnitude. The second integral is at most $\|F\|_\infty \cdot \int_{\|u\|_2 \geq \delta} B_c(u)du = \|F\|_\infty \cdot \int_{\|z\|_2 \geq c\delta} B(z)dz$ in magnitude. Now to bound this integral, we use Markov's inequality, which gives $\mathbf{Pr}_{v\sim B}[\|v\|_2 \geq t\sqrt{\mathbf{E}[\|v\|_2^2]}] \leq 1/t^2$. Recalling that $\mathbf{E}_{v\sim B}[\|v\|_2^2] = O(d^2)$, we have $\int_{\|v\|_2 \geq td} B(v)dv = O(1/t^2)$. Thus $\int_{\|z\|_2 \geq c\delta} B(z)dz = O(d^2/(c^2\delta^2))$. ■

We also prove the following lemma, which can be useful when FT-mollifying over high dimensions.

Lemma A.6. Let $c > 0$ be given, and $F : \mathbb{R}^d \to \mathbb{R}$ be bounded. Let p_k be the Taylor expansion of $\tilde{F}^c$ up to degree k about the origin. Then, for any $x \in \mathbb{R}^d$,

$$|\tilde{F}^c(x) - p_k(x)| \leq \|F\|_\infty \cdot \frac{(2c \cdot \|x\|_2)^{k+1}}{(k+1)!}$$

Proof. By the univariate Taylor's theorem, restricting F to the line passing through the origin and x,

$$\begin{aligned} |\tilde{F}^c(x) - p_k(x)| &\leq \frac{1}{(k+1)!} \cdot |(D_x^{k+1}\tilde{F}^c)(y)| \\ &= \frac{\|x\|_2^{k+1}}{(k+1)!} \cdot |(D_{x/\|x\|_2}^{k+1}\tilde{F}^c)(y)|, \end{aligned}$$

where D_x is the directional derivative in direction x, and y is some point on the line from the origin to x. Set $u = x/\|x\|_2$. Now, as in the proof of Theorem 4.1, we can obtain

$$|(D_u^k\tilde{F}^c)(y)| \leq \|F\|_\infty \cdot \left\|D_u^k B_c\right\|_1.$$

Since B is spherically symmetric, $D_u^{k+1}B_c = D_{e_1}^k B_c$, and thus $\left\|D_u^{k+1}B_c\right\|_1 \leq (2c)^{k+1}$. ■

A.1 Differences in FT-mollification constructions In this subsection we sketch the main differences between the univariate FT-mollification construction in [30], and the FT-mollification construction in this work. We note that [30] only considered FT-mollifying indicator functions of intervals of the real line, and thus we compare the construction there with that of Theorem 4.2.

In [30], for $[a,b] \subset \mathbb{R}$ one takes a bump function b (i.e. a smooth function of compact support), and defines $\tilde{I}_{[a,b]}^c = \hat{b}_c * I_{[a,b]}$ for $\hat{b}$ the Fourier transform of b. For a particularly chosen b in [30], the conclusion is that for $c \geq \text{polylog}(1/\varepsilon)/\varepsilon$, one has that $|\tilde{I}_{[a,b]}^c(x) - I_{[a,b]}(x)| \leq \varepsilon$ as long as $x \notin [a-\varepsilon, a+\varepsilon] \cup [b-\varepsilon, b+\varepsilon]$. For x inside these two intervals, we are only guaranteed that $|\tilde{I}_{[a,b]}^c(x)| = O(1)$. Further, we have $\|(\tilde{I}_{[a,b]}^c)^\ell\|_\infty = O(c^\ell)$. The "polylog$(1/\varepsilon)$" term in the requirement for c arises as a consequence of the function b chosen. Specifically, the proof given in [30] relied on fast decay of $\hat{b}$; the polylog$(1/\varepsilon)$ term above corresponds to $\hat{b}^{-1}(\varepsilon)$. The choice of b was then quite important, since the specifics of the decay rate of $\hat{b}$ played into the conclusion of the theorem. Proving good bounds on the decay rate of $\hat{b}$ then required some tedious calculation, and proving *sharp* bounds resorted to advanced methods such as saddle-point integration [28].

In the FT-mollification construction in this current work, when FT-mollifying $I_{[a,b]}$ we do not give an "all-or-nothing" guarantee (i.e., that the FT-mollification is within ε for x far from $\{a,b\}$, with no guarantee when near $\{a,b\}$). Rather, the quality of the approximation of $I_{[a,b]}$ by its FT-mollification continuously improves as x moves farther away from the boundary. Furthermore, in all our applications it is enough that the rate at which this quality of approximation improves be quadratic. To accomplish this, we convolve not with $\hat{b}$, but with $\hat{b}^2$, which is nonnegative. Since $\hat{b}^2$ is nonnegative and has integral 1, we can view it as a probability density then resort to probabilistic arguments to show that our convolution produces a good conclusion (specifically, we use that $\hat{b}^2$ has bounded variance then apply the second moment method). Almost any "smooth enough" b of compact supports yields a $\hat{b}^2$ with sufficiently small variance while maintaining good bounds on $\|(\hat{b}^2)^{(\ell)}\|_1$, and thus our choice of b is largely irrelevant. Furthermore, our proof of bounded variance is far simpler and less calculation-intensive than the decay bounds needed in [30]. Lastly, straightforward generalization of the approach in [30] to dimension d would result in error bounds which blow up at least exponentially in d (which for example would yield bounds of the form "$\exp(m) \cdot \mathrm{poly}(1/\varepsilon)$-wise independence fools interesections of m halfspaces"), while our current approach has only $\mathrm{poly}(d)$ blowup. As a bonus, bounds even in our univariate applications improve by $\mathrm{polylog}(1/\varepsilon)$ factors when using the newer FT-mollification.

B Proof of Theorem 7.1

We first give two lemmas, the second of which is a discrete analog of one of Whittle's lemmas.

Lemma B.1. For $a \in \mathbb{R}^n$, $x \in \{-1,1\}^n$ random, and $k \geq 2$ an even integer, $\mathbf{E}[(a^T x)^k] \leq \|a\|_2^k \cdot k^{k/2}$.

Proof. We may replace the coefficients of a by their absolute values. Note that when expanded $(a^T x)^k$ is a sum of monomials in the coefficients of x. Notice that if we replace x by a vector of independent normal distributions, y, the expectation of each monomial is only made larger. Hence $\mathbf{E}[(a^T x)^k] \leq \mathbf{E}[(a^T y)^k]$. On the other hand $a^T y$ is distributed as a normal with standard deviation $\|a\|_2$ and hence $\mathbf{E}[(a^T y)^k] = \|a\|_2^k k!/(2^k \cdot (k/2)!)$. ■

Lemma B.2. If X, Y are independent with $\mathbf{E}[Y] = 0$ and if $k \geq 2$, then $\mathbf{E}[|X|^k] \leq \mathbf{E}[|X - Y|^k]$.

Proof. Consider the function $f(y) = |X - y|^k$. Since $f^{(2)}$, the second derivative of f, is nonnegative on $\mathbb{R}$, the claim follows by Taylor's theorem since $|X - Y|^k \geq |X|^k - kY(\mathrm{sgn}(X) \cdot X)^{k-1}$. ■

We are now prepared to prove our Theorem 7.1.

Theorem 7.1 (restatement). *Let $A \in \mathbb{R}^{n \times n}$ be symmetric and $x \in \{-1,1\}^n$ be random. Then for all $k \geq 2$, $\mathbf{E}[|(x^T A x) - tr(A)|^k] \leq C^k \cdot \max\{\sqrt{k}\|A\|_2, k\|A\|_\infty\}^k$, where C is an absolute constant.*

Proof (of Theorem 7.1). Without loss of generality we can assume $\mathrm{tr}(A) = 0$. This is because if one considers $A' = A - (\mathrm{tr}(A)/n) \cdot I$, then $x^T A x - \mathrm{tr}(A) = x^T A' x$, and we have $\|A'\|_2 \leq \|A\|_2$ and $\|A'\|_\infty \leq 2\|A\|_\infty$. We now start by proving our theorem for k a power of 2 by induction on k. For $k = 2$, $\mathbf{E}[(x^T A x)^2] = 4\sum_{i<j} A_{i,j}^2$ and $\|A\|_2^2 = \sum_i A_{i,i}^2 + 2\sum_{i<j} A_{i,j}^2$. Thus $\mathbf{E}[(x^T A x)^2] \leq 2\|A\|_2^2$. Next we assume the statement of our Theorem for $k/2$ and attempt to prove it for k.

We note that by Lemma B.2,

$$\mathbf{E}[|x^T A x|^k] \leq \mathbf{E}[|x^T A x - y^T A y|^k] = \mathbf{E}[|(x+y)^T A (x-y)|^k],$$

where $y \in \{-1,1\}^n$ is random and independent of x. Notice that if we swap x_i with y_i then $x + y$ remains constant as does $|x_j - y_j|$ and that $x_i - y_i$ is replaced by its negation. Consider averaging over all such swaps. Let $\xi_i = ((x+y)^T A)_i$ and $\eta_i = x_i - y_i$. Let z_i be 1 if we did not swap and -1 if we did. Then $(x+y)^T A(x-y) = \sum_i \xi_i \eta_i z_i$. Averaging over all swaps,

$$\mathbf{E}_z[|(x+y)^T A(x-y)|^k] \leq \left(\sum_i \xi_i^2 \eta_i^2\right)^{k/2} \cdot k^{k/2} \leq 2^k k^{k/2} \cdot \left(\sum_i \xi_i^2\right)^{k/2}.$$

The first inequality is by Lemma B.1, and the second uses that $|\eta_i| \leq 2$. Note that

$$\sum_i \xi_i^2 = \|A(x+y)\|_2^2 \leq 2\|Ax\|_2^2 + 2\|Ay\|_2^2,$$

and hence

$$\mathbf{E}[|x^T Ax|^k] \leq 2^k \sqrt{k}^k \mathbf{E}[(2\|Ax\|_2^2 + 2\|Ay\|_2^2)^{k/2}] \leq 4^k \sqrt{k}^k \mathbf{E}[(\|Ax\|_2^2)^{k/2}],$$

with the final inequality using Minkowski's inequality (namely that $|\mathbf{E}[|X+Y|^p]|^{1/p} \leq |\mathbf{E}[|X|^p]|^{1/p} + |\mathbf{E}[|Y|^p]|^{1/p}$ for any random variables X, Y and any $1 \leq p < \infty$).

Next note $\|Ax\|_2^2 = \langle Ax, Ax\rangle = x^T A^2 x$. Let $B = A^2 - \frac{\mathrm{tr}(A^2)}{n} I$. Then $\mathrm{tr}(B) = 0$. Also, $\|B\|_2 \leq \|A\|_2 \|A\|_\infty$ and $\|B\|_\infty \leq \|A\|_\infty^2$. The former holds since

$$\|B\|_2^2 = \left(\sum_i \lambda_i^4\right) - \left(\sum_i \lambda_i^2\right)^2 \Big/ n \leq \sum_i \lambda_i^4 \leq \|A\|_2^2 \|A\|_\infty^2.$$

The latter holds since the eigenvalues of B are $\lambda_i^2 - (\sum_{j=1}^n \lambda_j^2)/n$ for each $i \in [n]$. The largest eigenvalue of B is thus at most that of A^2, and since $\lambda_i^2 \geq 0$, the smallest eigenvalue of B cannot be smaller than $-\|A\|_\infty^2$.

We then have

$$\mathbf{E}[(\|Ax\|_2^2)^{k/2}] = \mathbf{E}\left[\left|\|A\|_2^2 + x^T Bx\right|^{k/2}\right] \leq 2^k \max\{\|A\|_2^k, \mathbf{E}[|x^T Bx|^{k/2}]\}.$$

Hence employing the inductive hypothesis on B we have that

$$\begin{aligned}
\mathbf{E}[|x^T Ax|^k] &\leq 8^k \max\{\sqrt{k}\|A\|_2, C^{k/2} k^{3/4} \|B\|_2, C^{k/2} k \sqrt{\|B\|_\infty}\}^k \\
&\leq 8^k C^{k/2} \max\{\sqrt{k}\|A\|_2, k^{3/4}\sqrt{\|A\|_2\|A\|_\infty}, k\|A\|_\infty\}^k \\
&= 8^k C^{k/2} \max\{\sqrt{k}\|A\|_2, k\|A\|_\infty\}^k,
\end{aligned}$$

with the final equality holding since the middle term above is the geometric mean of the other two, and thus is dominated by at least one of them. This proves our hypothesis as long as $C \geq 64$.

To prove our statement for general k, set $k' = 2^{\lceil \log_2 k \rceil}$. Then by the power mean inequality and our results for k' a power of 2, $\mathbf{E}[|x^T Ax|^k] \leq (\mathbf{E}[|x^T Ax|^{k'}])^{k/k'} \leq 128^k \max\{\sqrt{k}\|A\|_2, k\|A\|_\infty\}^k$. ■

C Basic linear algebra facts

In this subsection we record some basic linear algebraic facts used in our proofs.

We start with two elementary facts.

Fact C.1. If $A, P \in \mathbb{R}^{n \times n}$ with P invertible, then the eigenvalues of A and $P^{-1}AP$ are identical.

Fact C.2. For $A \in \mathbb{R}^{n \times n}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$, and for integer $k > 0$, $\mathrm{tr}(A^k) = \sum_i \lambda_i^k$.

Note Fact C.1 and Fact C.2 imply the following.

Fact C.3. For a real matrix $A \in \mathbb{R}^{n \times n}$ and invertible matrix $P \in \mathbb{R}^{n \times n}$,

$$\|P^{-1}AP\|_2 = \|A\|_2.$$

The following standard result will be useful:

Theorem C.4 (Spectral Theorem [49, Section 6.4]). If $A \in \mathbb{R}^{n\times n}$ is symmetric, there exists an orthogonal $Q \in \mathbb{R}^{n\times n}$ with $\Lambda = Q^T AQ$ diagonal. In particular, all eigenvalues of A are real.

Definition C.5. For a real symmetric matrix A, we define $\lambda_{\min}(A)$ to be the smallest magnitude of a non-zero eigenvalue of A (in the case that all eigenvalues are 0, we set $\lambda_{\min}(A) = 0$). We define $\|A\|_\infty$ to be the largest magnitude of an eigenvalue of A.

We now give a simple lemma that gives an upper bound on the magnitude of the trace of a symmetric matrix with positive eigenvalues.

Lemma C.6. Let $A \in \mathbb{R}^{n\times n}$ be symmetric with $\lambda_{\min}(A) > 0$. Then $|\mathrm{tr}(A)| \leq \|A\|_2^2/\lambda_{\min}(A)$.

Proof. We have

$$\begin{aligned} |\mathrm{tr}(A)| &= \left|\sum_{i=1}^n \lambda_i\right| \\ &\leq \frac{\|A\|_2}{\lambda_{\min}(A)} \cdot \sqrt{\sum_{i=1}^n \lambda_i^2} \\ &= \frac{\|A\|_2^2}{\lambda_{\min}(A)} \end{aligned}$$

We note $\sum_{i=1}^n \lambda_i^2 = \|A\|_2^2$, implying the final equality. Also, there are at most $\|A\|_2^2/(\lambda_{\min}(A))^2$ non-zero λ_i. The sole inequality then follows by Cauchy-Schwarz. ■

D Useful facts about polynomials

D.1 Facts about low-degree polynomials. We view $\{-1, 1\}^n$ as a probability space endowed with the uniform probability measure. For a function $f : \{-1, 1\}^n \to \mathbb{R}$ and $r \geq 1$, we let $\|f\|_r$ denote $(\mathbf{E}_x[|f(x)|^r])^{1/r}$.

Our first fact is a consequence of the well-known hypercontractivity theorem.

Theorem D.1 (Hypercontractivity [6, 10]). If f is a degree-d polynomial and $1 \leq r < q \leq \infty$,

$$\|f\|_q \leq \sqrt{\frac{q-1}{r-1}}^{\,d} \|f\|_r.$$

Our second fact is an anticoncentration theorem for low-degree polynomials over independent standard Gaussian random variables.

Theorem D.2 (Gaussian Anticoncentration [14]). For f a non-zero, n-variate, degree-d polynomial,

$$\mathbf{Pr}[|f(G_1, \ldots, G_n) - t| \leq \varepsilon \cdot \mathbf{Var}[f]] = O(d\varepsilon^{1/d})$$

for all $\varepsilon \in (0, 1)$ and $t \in \mathbb{R}$. Here $G_1, \ldots, G_n \sim \mathcal{N}(0, 1)$ are independent. (Here, and henceforth, $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2.)

The following is a statement of the Invariance Principle of Mossell, O'Donnell, and Oleszkiewicz [38], in the special case when the random variables X_i are Bernoulli.

Theorem D.3 (Invariance Principle [38]). Let $X_1, \ldots, X_n$ be independent ± 1 Bernoulli, and let p be a degree-d multilinear polynomial with $\sum_{|S|>0} \widehat{p}_S^2 = 1$ and $\max_i \mathrm{Inf}_i(p) \leq \tau$. Then

$$\sup_t |\mathbf{Pr}[p(X_1, \ldots, X_n) \leq t] - \mathbf{Pr}[p(G_1, \ldots, G_n) \leq t]| = O(d\tau^{1/(4d+1)})$$

where the $G_i \sim \mathcal{N}(0,1)$ are independent.

The following tail bound argument is standard (see for example [4]). We repeat the argument here just to point out that only bounded independence is required.

Theorem D.4 (Tail bound). If f is a degree-d polynomial, $t > 8^{d/2}$, and X is drawn at random from a $(dt^{2/d})$-wise independent distribution over $\{-1,1\}^n$, then

$$\mathbf{Pr}[|f(X)| \geq t\|f\|_2] = \exp(-\Omega(dt^{2/d})).$$

Proof. Suppose $k > 2$. By Theorem D.1,

$$\mathbf{E}[|f(X)|^k] \leq k^{dk/2} \cdot \|f\|_2^k,$$

implying

$$\mathbf{Pr}[|f(X)| \geq t\|f\|_2] \leq (k^{d/2}/t)^k \tag{D.1}$$

by Markov's inequality. Set $k = 2 \cdot \lfloor t^{2/d}/4 \rfloor$ and note $k > 2$ as long as $t > 8^{d/2}$. Now the right hand side of Eq. (D.1) is at most $2^{-dk/2}$, as desired. Finally, note independence was only used to bound $\mathbf{E}[|f(X)|^k]$, which for k even equals $\mathbf{E}[f(X)^k]$ and is thus determined by dk-independence. ■

D.2 Facts about quadratic forms. The following facts are concerned with quadratic forms, i.e. polynomials $p(x) = \sum_{i \leq j} a_{i,j} x_i x_j$. We often represent a quadratic form p by its associated symmetric matrix A_p, where

$$(A_p)_{i,j} = \begin{cases} a_{i,j}/2, & i < j \\ a_{j,i}/2, & i > j \\ a_{i,j}, & i = j \end{cases}$$

so that $p(x) = x^T A_p x$.

The following is a bound on moments for quadratic forms.

Lemma D.5. Let $f(x)$ be a degree-2 polynomial. Then, for $X = (X_1, \ldots, X_n)$ a vector of independent Bernoullis,

$$\mathbf{E}[|f(X)|^k] \leq 2^{O(k)}(\|A_f\|_2 k^k + |\mathrm{tr}(A_f)|^k).$$

Proof. Over the hypercube we can write $f = q + \mathrm{tr}(A_f)$ where q is multilinear. Note $\|A_q\|_2 \leq \|A_f\|_2$. Then by Theorem D.1,

$$\begin{aligned} \mathbf{E}[|f(x)|^k] &= \mathbf{E}[|q(x) + \mathrm{tr}(A_f)|^k] \\ &\leq \sum_{i=0}^{k} (\|A_f\|_2 \cdot i)^i |\mathrm{tr}(A_f)|^{k-i} \\ &\leq \sum_{i=0}^{k} (\|A_f\|_2 \cdot k)^i |\mathrm{tr}(A_f)|^{k-i} \\ &= 2^{O(k)} \max\{\|A_f\|_2 \cdot k, |\mathrm{tr}(A_f)|\}^k \end{aligned}$$

■

The following corollary now follows from Theorem D.4 and Lemma C.6.

Corollary D.6. Let f be a quadratic form with A_f positive semidefinite, $\|A_f\|_2 \leq 1$, and $\lambda_{\min}(A_f) \geq \delta$ for some $\delta \in (0, 1]$. Then, for x chosen at random from a $\lceil 2/\delta \rceil$-independent family over $\{-1, 1\}^n$,

$$\mathbf{Pr}[f(x) > 2/\delta] = \exp(-\Omega(1/\delta)).$$

Proof. Write $f = g + C$ via Lemma C.6 with $0 \leq C \leq 1/\delta$ and g multilinear, $\|A_g\|_2 \leq \|A_f\|_2 \leq 1$. Apply Theorem D.4 to g with $t = 1/\delta$. ■

The following lemma gives a decomposition of any multi-linear quadratic form as a sum of quadratic forms with special properties for the associated matrices. It is used in the proof of Theorem 7.2.

Lemma D.7. Let $\delta > 0$ be given. Let f be a multilinear quadratic form. Then f can be written as $f_1 - f_2 + f_3$ for quadratic forms f_1, f_2, f_3 where:

1. A_{f_1}, A_{f_2} are positive semidefinite with $\lambda_{\min}(A_{f_1}), \lambda_{\min}(A_{f_2}) \geq \delta$.
2. $\|A_{f_3}\|_\infty < \delta$.
3. $\|A_{f_1}\|_2, \|A_{f_2}\|_2, \|A_{f_3}\|_2 \leq \|A_f\|_2$.

Proof. Since A_f is real and symmetric, we can find an orthogonal matrix Q such that $\Lambda = Q^T A_f Q$ is diagonal. Each diagonal entry of Λ is either at least δ, at most $-\delta$, or in between. We create a matrix P containing all entries of Λ which are at least δ, with the others zeroed out. We similarly create N to have all entries at most $-\delta$. We place the remaining entries in R. We then set $A_{f_1} = QPQ^T, A_{f_2} = QNQ^T, A_{f_3} = QRQ^T$. Note $\|\Lambda\|_2^2 = \|A_f\|_2^2$ by Fact C.3, so since we remove terms from Λ form each A_{f_i}, their Frobenius norms can only shrink. The eigenvalue bounds hold by construction and Fact C.1. ■

E Why the previous approaches failed

Here we attempt to provide an explanation as to why the approaches of [16] and [30] fail to fool degree-2 PTFs. Furthermore, Ganzburg has shared with us a conjecture [19] that for any distribution with tail bound $e^{-O(|x|)}$, any polynomial which L_1-approximates the sign function to within ε with respect to that distribution must have degree $2^{\Omega(1/\varepsilon^2)}$. Since the degree of sandwiching polynomials with small L_1-error *characterizes* the independence required to fool a boolean function, this conjecture would imply that any attempt to find a univariate polynomial q_ε such that $\mathbf{E}[|q_\varepsilon(p(x)) - \mathrm{sgn}(p(x))|] \leq \varepsilon$ would require $\deg(q_\varepsilon) = 2^{\Omega(1/\varepsilon^2)}$ if $\deg(p) = 2$, given known (and tight) tail bounds for degree-2 polynomials, i.e. it would be impossible to prove a $\mathrm{poly}(1/\varepsilon)$ bound on the amount of independence required via a univariate approach.

E.1 Why the approximation theory approach failed The analysis in [16] crucially exploits the strong concentration and anti-concentration properties of the gaussian distribution. (Recall that in the linear regular case, the random variable $\langle w, x\rangle$ is approximately Gaussian.) Now consider a regular degree-2 polynomial p and the corresponding PTF $f = \mathrm{sgn}(p)$. Since p is regular, it has still has "good" concentration and anti-concentration properties – though quantitatively inferior than those of the Gaussian. Hence, one would hope to argue as follows: use the univariate polynomial P (constructed using approximation theory), allowing its degree to increase if necessary, and carry out the analysis of the error as in the linear case.

The reason this fails is because the (tight) concentration properties of p – as implied by hypercontractivity – are not sufficient for the analysis to bound the error of the approximation, even if we let the degree of the polynomial P tend to infinity. (Paradoxically, the error coming from the worst-case analysis becomes worse as the degree of P increases.)

Without going into further details, we mention that an additional problem for univariate approximations to work is this: the (tight) anti-concentration properties of p – obtained via the Invariance Principle and the anti-concentration bounds of [14] – are quantitatively weaker than what is required to bound the error, even in the region where P has small point-wise error (from the sgn function).

E.2 Why the analysis for univariate FT-mollification failed We discuss why the argument in [30] failed to generalize to higher degree. Recall that the argument was via the following chain of inequalities:

$$\mathbf{E}[I_{[0,\infty)}(p(X))] \approx_\varepsilon \mathbf{E}[\tilde{I}^c_{[0,\infty)}(p(X))] \approx_\varepsilon \mathbf{E}[\tilde{I}^c_{[0,\infty)}(p(Y))] \approx_\varepsilon \mathbf{E}[I_{[0,\infty)}(p(Y))] \qquad \text{(E.1)}$$

The step that fails for high-degree PTFs is the second inequality in Eq. (E.1), which was argued by Taylor's theorem. Our bounds on derivatives of $\tilde{I}^c_{[0,\infty)}$, the FT-mollification of $I_{[0,\infty)}$ for a certain parameter $c = c(\varepsilon)$ to make sure $|I_{[0,\infty)} - \tilde{I}^c_{[0,\infty)}| < \varepsilon$ "almost everywhere", are such that $||(\tilde{I}^c_{[0,\infty)})^{(k)}||_\infty \geq 1$ for all k. Thus, we have that the error term from Taylor's theorem is at least $\mathbf{E}[(p(X))^k]/k!$. The problem comes from the numerator. Since we can assume the sum of squared coefficients of p is 1 (note the sgn function is invariant to scaling of its argument), known (and tight) moment bounds (via hypercontractivity) only give us an upper bound on $\mathbf{E}[(p(x))^k]$ which is larger than $k^{dk/2}$, where degree$(p) = d$. Thus, the error from Taylor's theorem does not decrease to zero by increasing k for $d \geq 2$, since we only are able to divide by $k! \leq k^k$ (in fact, strangely, increasing the amount of independence k *worsens* this bound).

F Proofs omitted from Section 7.2

F.1 Boolean setting.

Lemma F.1. For a quadratic form f and random $x \in \{-1,1\}^n$,

$$\mathbf{E}[|f(x)|^k] \leq 2^{O(k)} \cdot (\|A_f\|_2 k^k + (\|A_f\|_2^2/\lambda_{\min}(A_f))^k).$$

Proof. Combine Lemma C.6 and Lemma D.5. ■

Lemma 7.3 (restatement). *Let $\varepsilon > 0$ be arbitrary. Let $F : \mathbb{R}^4 \to \mathbb{R}$ be even in each of its first two arguments such that $\|\partial^\beta \tilde{F}^c\|_\infty = O(\alpha^{|\beta|})$ for all multi-indices $\beta \in \mathbb{N}^4$ and some $\alpha > 1$. Suppose $1/\delta \geq B\alpha$ for a sufficiently large constant B. Let $X_1, \ldots, X_n$ be independent Bernoulli, and $Y_1, \ldots, Y_n$ be k'-independent Bernoulli for $k' = 2k$ with $k \geq \max\{\log(1/\varepsilon), B\alpha/\sqrt{\delta}, B\alpha^2\}$ an even integer. Write $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$. Then $|\mathbf{E}[F(M_p(X))] - \mathbf{E}[F(M_p(Y))]| < \varepsilon$.*

Proof. We Taylor-expand F to obtain a polynomial P_{k-1} containing all monomials up to degree $k-1$. Since $F(x)$ is even in x_1, x_2, we can assume P_{k-1} is a polynomial in x_1^2, x_2^2, x_3, x_4. Let $x \in \mathbb{R}^4$ be arbitrary. We apply Taylor's theorem to bound $R(x) = |F(x) - P_{k-1}(x)|$. Define

$x_* = \max_i\{|x_i|\}$. Then

$$\begin{aligned} R(x) &\leq \alpha^k \cdot \sum_{|\beta|=k} \frac{|x_1|^{\beta_1}\cdot|x_2|^{\beta_2}\cdot|x_3|^{\beta_3}\cdot|x_4|^{\beta_4}}{\beta_1!\cdot\beta_2!\cdot\beta_3!\cdot\beta_4!} \\ &\leq \alpha^k x_*^k \cdot \sum_{|\beta|=k} \frac{1}{\beta_1!\cdot\beta_2!\cdot\beta_3!\cdot\beta_4!} \\ &= \alpha^k x_*^k \cdot \frac{1}{k!}\cdot \sum_{|\beta|=k} \binom{k}{\beta_1,\ldots,\beta_4} \\ &\leq \alpha^k 4^k \cdot \frac{x_1^k + x_2^k + x_3^k + x_4^k}{k!}, \end{aligned} \tag{F.1}$$

with the absolute values unnecessary in the last inequality since k is even. We now observe

$$\begin{aligned} &|\mathbf{E}[F(M_p(X))] - \mathbf{E}[F(M_p(Y))]| \\ &\qquad \leq \alpha^k 2^{O(k)} \cdot \frac{\mathbf{E}[(p_1(X))^{k/2}] + \mathbf{E}[(p_2(X))^{k/2}] + \mathbf{E}[(p_3(X)-\Upsilon)^k] + \mathbf{E}[(p_4(X))^k]}{k^k} \end{aligned}$$

since (a) every term in $P_{k-1}(M_p(X))$ is a monomial of degree at most $2k-2$ in the X_i, by evenness of P_{k-1} in x_1, x_2, and is thus determined by $2k$-independence, (b) $\sqrt{p_1(X)}, \sqrt{p_2(X)}$ are real by positive semidefiniteness of p_1, p_2 (note that we are only given that the high order partial derivatives are bounded by $O(\alpha^k)$ on the reals; we have no guarantees for complex arguments), and (c) the moment expectations above are equal for X and Y since they are determined by $2k$-independence.

We now bound the error term above. We have

$$\mathbf{E}[(p_1(X))^{k/2}] = 2^{O(k)}(k^{k/2} + \delta^{-k/2})$$

by Lemma F.1, with the same bound holding for $\mathbf{E}[(p_2(X))^{k/2}]$. We also have

$$\mathbf{E}[(p_3(X) - \Upsilon)^k] \leq 2^{O(k)} \cdot \max\{\sqrt{k}, (\delta k)\}^k$$

by Theorem 7.1. We finally have

$$\mathbf{E}[(p_4(X))^k] \leq k^{k/2}$$

by Lemma B.1. Thus in total,

$$|\mathbf{E}[F(M_p(X))] - \mathbf{E}[F(M_p(Y))]| \leq 2^{O(k)} \cdot ((\alpha/\sqrt{k})^k + (\alpha/(k\sqrt{\delta}))^k + (\alpha\delta)^k),$$

which is at most ε for sufficiently large B by our lower bounds on k and $1/\delta$. ■

To prove Lemma 7.4, we make use of the following lemma, which follows from the Invariance Principle, the hypercontractive inequality, and the anticoncentration bound of [14]. Here p_1, p_2, δ are as in Section 7.2 (recall $p = p_1 - p_2 + p_3 + p_4 + C$ where p_1, p_2 are positive semidefinite with minimum non-zero eigenvalues at least δ).

Lemma F.2. Let $\eta, \eta' \geq 0, t \in \mathbb{R}$ be given, and let $X_1, \ldots, X_n$ be independent Bernoulli. Then

$$\mathbf{Pr}[|p(X) - t| \leq \eta \cdot (\sqrt{p_1(X)} + \sqrt{p_2(X)} + 1) + \eta'] = O(\sqrt{\eta'} + (\eta^2/\delta)^{1/4} + \tau^{1/9} + \exp(-\Omega(1/\delta))).$$

Proof. Applying Corollary D.6, we have

$$\mathbf{Pr}[\sqrt{p_1(X)} \geq \sqrt{2/\delta}] = \exp(-\Omega(1/\delta)),$$

and similarly for $\sqrt{p_2(X)}$. We can thus bound our desired probability by

$$\mathbf{Pr}[|p(X) - t| \leq 2\eta\sqrt{2/\delta} + \eta + \eta'] + \exp(-\Omega(1/\delta)).$$

By Theorem D.2, together with Theorem D.3, we can bound the probability in the lemma statement by

$$O(\sqrt{\eta'} + (\eta^2/\delta)^{1/4} + \tau^{1/9} + \exp(-\Omega(1/\delta))).$$

■

We now prove our anticoncentration lemma in the case of limited independence.

Lemma 7.4 (restatement). *Let ε' be given. Suppose $k \geq D/(\varepsilon')^4$ for a sufficiently large constant $D > 0$. Let $Y_1, \ldots, Y_n$ be k-wise independent Bernoulli, and let $t \in \mathbb{R}$ be arbitrary. Then*

$$\mathbf{Pr}[|p(Y) - t| < \varepsilon'] \leq O(\sqrt{\varepsilon'} + \tau^{1/9}).$$

Proof. Define the region $T_{t,\varepsilon'} = \{(x_1, x_2, x_3, x_4) : |x_1^2 - x_2^2 + x_3 + x_4 + C + \Upsilon - t| < \varepsilon'\}$, and also the region $S_{\rho,t,\varepsilon'} = \{x : d_2(x, T_{t,\varepsilon'}) \leq \rho\}$ for $\rho \geq 0$. Consider the FT-mollification $\tilde{I}^c_{S_{\rho,t,\varepsilon'}}$ of $I_{S_{\rho,t,\varepsilon'}}$ for $c = A/\rho$, with A a large constant to be determined later. We note a few properties of $\tilde{I}^c_{S_{\rho,t,\varepsilon'}}$:

i. $\|\partial^\beta \tilde{I}^c_{S_{\rho,t,\varepsilon'}}\|_\infty \leq (2c)^{|\beta|}$

ii. $\tilde{I}^c_{S_{\rho,t,\varepsilon'}}(x) \geq \frac{1}{2} \cdot I_{T_{t,\varepsilon'}}(x)$

iii. $\tilde{I}^c_{S_{\rho,t,\varepsilon'}}(x) = \max\left\{1, O\left((c \cdot d_2(x, T_{t,\varepsilon'}))^{-2}\right)\right\}$ for any x with $d_2(x, T_{t,\varepsilon'}) \geq 2\rho$

Item (i) is straightforward from Theorem 4.1. For item (ii), note that if $x \in T_{t,\varepsilon'}$, then $d_2(x, \partial S_{\rho,t,\varepsilon'}) \geq \rho$, implying

$$|\tilde{I}^c_{S_{\rho,t,\varepsilon'}}(x) - 1| = O\left(\frac{1}{c^2\rho^2}\right),$$

which is at most $1/2$ for A a sufficiently large constant. Furthermore, $\tilde{I}^c_{S_{\rho,t,\varepsilon'}}$ is nonnegative. Finally, for (iii), by Theorem 4.2 we have

$$\begin{aligned}
\tilde{I}^c_{S_{\rho,t,\varepsilon'}}(x) &= \max\left\{1, O\left((c \cdot d_2(x, \partial S_{\rho,t,\varepsilon'}))^{-2}\right)\right\} \\
&\leq \max\left\{1, O\left((c \cdot d_2(x, S_{\rho,t,\varepsilon'}))^{-2}\right)\right\} \\
&\leq \max\left\{1, O\left((c \cdot (d_2(x, T_{t,\varepsilon'}) - \rho))^{-2}\right)\right\} \\
&\leq \max\left\{1, O\left((c \cdot d_2(x, T_{t,\varepsilon'}))^{-2}\right)\right\}
\end{aligned}$$

with the last inequality using that $d_2(x, T_{t,\varepsilon'}) \geq 2\rho$.

Noting $\mathbf{Pr}[|p(Z) - t| < \varepsilon'] = \mathbf{E}[I_{T_{t,\varepsilon'}}(M_p(Z))]$ for any random variable $Z = (Z_1, \ldots, Z_n)$, item (ii) tells us that

$$\mathbf{Pr}[|p(Z) - t| \leq \varepsilon'] \leq 2 \cdot \mathbf{E}[\tilde{I}^c_{S_{\rho,t,\varepsilon'}}(M_p(Z))]. \tag{F.2}$$

We now proceed in two steps. We first show $\mathbf{E}[\tilde{I}^c_{S_{\rho,t,\varepsilon'}}(M_p(X))] = O(\sqrt{\varepsilon'} + \tau^{1/9})$ by applications of Lemma F.2. We then show $\mathbf{E}[\tilde{I}^c_{S_{\rho,t,\varepsilon'}}(M_p(Y))] = O(\sqrt{\varepsilon'} + \tau^{1/9})$ by applying Lemma 7.3, at which point we will have proven our lemma via Eq. (F.2) with $Z = Y$.

$\mathbf{E}[\tilde{\mathbf{I}}^{\mathbf{c}}_{\mathbf{S}_{\rho,\mathbf{t},\varepsilon'}}(\mathbf{M_p}(\mathbf{X}))] = \mathbf{O}(\sqrt{\varepsilon'} + \tau^{\mathbf{1/9}})$: We first observe that for $x \notin T_{t,\varepsilon'}$,

$$d_2(x, T_{t,\varepsilon'}) \geq \frac{1}{2} \cdot \min\left\{\frac{|x_1^2 - x_2^2 + x_3 + x_4 + C + \Upsilon - t| - \varepsilon'}{2(|x_1| + |x_2| + 1)}, \sqrt{|x_1^2 - x_2^2 + x_3 + x_4 + C + \Upsilon - t| - \varepsilon'}\right\}. \tag{F.3}$$

This is because by adding a vector v to x, we can change each individual coordinate of x by at most $\|v\|_2$, and can thus change the value of $|x_1^2 - x_2^2 + x_3 + x_4 + C + \Upsilon - t| - \varepsilon'$ by at most $2\|v\|_2 \cdot (|x_1| + |x_2| + 1) + \|v\|_2^2$.

Now let $X \in \{-1, 1\}^n$ be uniformly random. We thus have that, for any particular $w > 0$,

$$\begin{aligned}
\mathbf{Pr}[0 < d_2(M_p(X), T_{t,\varepsilon'}) \leq w] &\leq \mathbf{Pr}\left[\min\left\{\frac{|p(X) - t| - \varepsilon'}{2(\sqrt{p_1(X)} + \sqrt{p_2(X)} + 1)}, \sqrt{|p(X) - t| - \varepsilon'}\right\} \leq 2w\right] \\
&\leq \mathbf{Pr}[|p(X) - t| \leq 4w \cdot (\sqrt{p_1}(X) + \sqrt{p_2(X)} + 1) + \varepsilon'] \\
&\quad + \mathbf{Pr}[|p(X) - t| \leq 4w^2 + \varepsilon'] \\
&= O(\sqrt{\varepsilon'} + w + \sqrt{w} + (w^2/\delta)^{1/4} + \tau^{1/9} + \exp(-\Omega(1/\delta)))
\end{aligned}$$

with the last inequality holding by Lemma F.2.

Now, by item (iii),

$$\begin{aligned}
&\mathbf{E}[\tilde{I}^c{}_{S_{\rho,t,\varepsilon'}}(M_p(X))] \\
&\quad \leq \mathbf{Pr}[d_2(M_p(X), T_{t,\varepsilon'}) \leq 2\rho] + O\left(\sum_{s=1}^{\infty} 2^{-2s} \cdot \mathbf{Pr}[2^s\rho < d_2(M_p(X), T_{t,\varepsilon'}) \leq 2^{s+1}\rho]\right) \\
&\quad \leq O(\sqrt{\varepsilon'} + \sqrt{\rho} + (\rho^2/\delta)^{1/4} + \tau^{1/9} + \exp(-\Omega(1/\delta)) \\
&\qquad + O\left(\sum_{s=1}^{\infty} 2^{-2s} \cdot (\sqrt{\varepsilon'} + 2^{s+1}\rho + \sqrt{2^{s+1}\rho} + (2^{2s+2}\rho^2/\delta)^{1/4} + \tau^{1/9} + \exp(-\Omega(1/\delta)))\right) \\
&\quad = O(\sqrt{\varepsilon'} + \sqrt{\rho} + (\rho^2/\delta)^{1/4} + \tau^{1/9} + \exp(-\Omega(1/\delta))
\end{aligned} \tag{F.4}$$

We now make the settings

$$\rho = (\varepsilon')^2, \qquad \frac{1}{\delta} = 2Bc = \frac{2AB}{\rho}.$$

where $B > 1$ is the sufficiently large constant in Lemma 7.3. Thus Eq. (F.4) is now $O(\sqrt{\varepsilon'} + \tau^{1/9})$. (We remark that a different δ is used when proving Theorem 7.2.)

$\mathbf{E}[\tilde{\mathbf{I}}^{\mathbf{c}}_{\mathbf{S}_{\rho,\mathbf{t},\varepsilon'}}(\mathbf{M_p}(\mathbf{Y}))] = \mathbf{O}(\sqrt{\varepsilon'} + \tau^{\mathbf{1/9}})$: It suffices to show

$$\mathbf{E}[\tilde{I}^c_{S_{\rho,t,\varepsilon'}}(M_p(Y))] \approx_\varepsilon \mathbf{E}[\tilde{I}^c_{S_{\rho,t,\varepsilon'}}(M_p(X))].$$

We remark that $\tilde{I}^c_{S_{\rho,t,\varepsilon'}}$ can be assumed to be even in both x_1, x_2. If not, then consider the symmetrization

$$(\tilde{I}^c_{S_{\rho,t,\varepsilon'}}(x_1,x_2,x_3,x_4)+\tilde{I}^c_{S_{\rho,t,\varepsilon'}}(-x_1,x_2,x_3,x_4)+\tilde{I}^c_{S_{\rho,t,\varepsilon'}}(x_1,-x_2,x_3,x_4)+\tilde{I}^c_{S_{\rho,t,\varepsilon'}}(-x_1,-x_2,x_3,x_4))/4, \tag{F.5}$$

which does not affect any of our properties (i),(ii), (iii).

Now, by our choice of k, δ and item (i), we have by Lemma 7.3 (with $\alpha = 2c$) that

$$|\mathbf{E}[\tilde{I}^c_{S_{\rho,t,\varepsilon'}}(M_p(X))] - \mathbf{E}[\tilde{I}^c_{S_{\rho,t,\varepsilon'}}(M_p(Y))]| < \varepsilon'.$$

This completes our proof by applying Eq. (F.2) with $Z = Y$. ■

Lemma 7.5 (restatement). *Let $\eta, \eta' \geq 0$ be given, and let $Y_1, \ldots, Y_n$ be k-independent Bernoulli for k as in Lemma 7.4 with $\varepsilon' = \min\{\eta/\sqrt{\delta}, \eta'\}$. Also assume $k \geq \lceil 2/\delta \rceil$. Then*

$$\mathbf{Pr}[|p(X) - t| \leq \eta \cdot (\sqrt{p_1(X)} + \sqrt{p_2(X)} + 1) + \eta'] = O(\sqrt{\eta'} + (\eta^2/\delta)^{1/4} + \tau^{1/9} + \exp(-\Omega(1/\delta))).$$

Proof. There were two steps in the proof of Lemma F.2 which required using the independence of the X_i. The first was in the application of Corollary D.6, but that only required $\lceil 2/\delta \rceil$-wise independence, which is satisfied here. The next was in using the anticoncentration of $p(X)$ (the fact that $\mathbf{Pr}[|p(X) - t| < s] = O(\sqrt{s} + \tau^{1/9})$ for any $t \in \mathbb{R}$ and $s > 0$). However, given Lemma 7.4, anticoncentration still holds under k-independence. ■

F.2 Gaussian Setting In the following Theorem we show that the conclusion of Theorem 7.2 holds even under the Gaussian measure.

Theorem F.3. Let $0 < \varepsilon < 1$ be given. Let $G = (G_1, \ldots, G_n)$ be a vector of independent standard normal random variables, and $G' = (G'_1, \ldots, G'_n)$ be a vector of $2k$-wise independent standard normal random variables for k a sufficiently large multiple of $1/\varepsilon^8$. If $p(x) = \sum_{i \leq j} a_{i,j} x_i x_j$ has $\sum_{i \leq j} a^2_{i,j} = 1$,

$$\mathbf{E}[\mathrm{sgn}(p(G))] - \mathbf{E}[\mathrm{sgn}(p(G'))] = O(\varepsilon).$$

Proof. Our proof is by a reduction to the Bernoulli case, followed by an application of Theorem 7.2. We replace each G_i with $Z_i = \sum_{j=1}^N X_{i,j}/\sqrt{N}$ for a sufficiently large N to be determined later. We also replace each G'_i with $Z'_i = \sum_{j=1}^N Y_{i,j}/\sqrt{N}$. We determine these $X_{i,j}, Y_{i,j}$ as follows. Let $\Phi : \mathbb{R} \to [0,1]$ be the cumulative distribution function (CDF) of the standard normal. Define $T_{-1,N} = -\infty$, $T_{N,N} = \infty$, and $T_{k,N} = \Phi^{-1}(2^{-N} \sum_{j=0}^k \binom{N}{k})$ for $0 \leq k \leq N$. Now, after a G_i is chosen according to a standard normal distribution, we identify the unique k_i such that $T_{k_i-1,N} \leq G_i < T_{k_i,N}$. We then randomly select a subset of k_i of the $X_{i,j}$ to make 1, and we set the others to -1. The $Y_{i,j}$ are defined similarly. It should be noted that the $X_{i,j}, Y_{i,j}$ are Bernoulli random variables, with the $X_{i,j}$ being independent and the $Y_{i,j}$ being $2k$-wise independent. Furthermore, we define the nN-variate polynomial $p' : \{-1,1\}^{nN} \to \mathbb{R}$ to be the one obtained from this procedure, so that $p(G) = p'(X)$. We then define $p''(x) = \alpha \cdot p'(x)$ for $\alpha = (\sum_{i<j} a^2_{i,j} + (1 - 1/N) \sum_i a^2_{i,i})^{-1}$ so that the sum of squared coefficients in p'' (ignoring constant terms, some of which arise because the $x^2_{i,j}$ terms are 1 on the hypercube) is 1. It should be observed that $1 \leq \alpha \leq 1 + 1/(N-1)$.

Now, we make the setting $\epsilon = \log^{1/3}(N)/\sqrt{N}$. By the Chernoff bound,

$$\mathbf{Pr}[|k_i - N/2| \geq \epsilon N/2] = o(1) \text{ as } N \text{ grows.} \tag{F.6}$$

Claim F.4. If $(1-\epsilon)N/2 \le k_i \le (1+\epsilon)N/2$, then $|T_{k_i,N} - T_{k_i+1,N}| = o(1)$.

Before proving the claim, we show how now we can use it to prove our Theorem. We argue by the following chain of inequalities:

$$\mathbf{E}[\mathrm{sgn}(p(G))] \approx_\varepsilon \mathbf{E}[\mathrm{sgn}(p''(X))] \approx_\varepsilon \mathbf{E}[\mathrm{sgn}(p''(Y))] \approx_\varepsilon \mathbf{E}[\mathrm{sgn}(p(G'))].$$

$\mathbf{E}[\mathrm{sgn}(\mathbf{p}(\mathbf{G}))] \approx_\varepsilon \mathbf{E}[\mathrm{sgn}(\mathbf{p}''(\mathbf{X}))]$: First we condition on the event $\mathcal{E}$ that $|Z_i - G_i| \le \varepsilon^3/n^2$ for all $i \in [n]$; this happens with probability $1-o(1)$ as N grows by coupling Claim F.4 and Eq. (F.6), and applying a union bound over all $i \in [n]$. We also condition on the event $\mathcal{E}'$ that $|G_i| = O(\sqrt{\log(n/\varepsilon)})$ for all $i \in [n]$, which happens with probability $1-\varepsilon^2$ by a union bound over $i \in [n]$ since a standard normal random variable has probability $e^{-\Omega(x^2)}$ of being larger than x in absolute value. Now, conditioned on $\mathcal{E}, \mathcal{E}'$, we have

$$|p(G)-p''(X)| \le n^2(\varepsilon^3/n^2)^2 + (\varepsilon^3/n^2)\sum_i |G_i| \left(\sum_j |a_{i,j}|\right) \le \varepsilon^2 + (\varepsilon^3/n^2)\cdot O(\sqrt{\log(n/\varepsilon)})\cdot \sum_{i,j} |a_{i,j}|.$$

We note $\sum_{i,j} a_{i,j}^2 = 1$, and thus $\sum_{i,j} |a_{i,j}| \le n$ by Cauchy-Schwarz. We thus have that $|p'(X) - p(G)| \le \varepsilon^2$ with probability at least $1-\varepsilon^2$, and thus $|p''(X) - p(G)| \le \varepsilon^2 + |(\alpha-1)\cdot p(X)|$ with probability at least $1-\varepsilon^2$. We finally condition on the event $\mathcal{E}''$ that $|(\alpha-1)\cdot p'(X)| \le \varepsilon^2$. Since p' can be written as a multilinear quadratic form with sum of squared coefficients at most 1, plus its trace $\mathrm{tr}(A_{p'})$ (which is $\sum_i a_{i,i} \le \sqrt{n}$, by Cauchy-Schwarz), we have

$$\mathbf{Pr}[|(\alpha-1)\cdot p'(X)| \ge \varepsilon^2] \le \mathbf{Pr}[|p'(X)| \ge \varepsilon^2\cdot(N-1)] = o(1),$$

which for large enough N and the fact that $\|p'\|_2 = O(1+\mathrm{tr}(A_{p'}))$ irrespective of N, is at most

$$\mathbf{Pr}[|p'(X)| \ge c\cdot\log(1/\varepsilon)\|p'\|_2],$$

for a constant c we can make arbitrarily large by increasing N. We thus have $\mathbf{Pr}[\mathcal{E}''] \ge 1-\varepsilon^2$ by Theorem D.4. Now, conditioned on $\mathcal{E}\wedge\mathcal{E}'\wedge\mathcal{E}''$, $\mathrm{sgn}(p''(X)) \ne \mathrm{sgn}(p(G))$ can only occur if $|p''(X)| = O(\varepsilon^2)$. However, by anticoncentration (Theorem D.2) and the Invariance Principle (Theorem D.3), this occurs with probability $O(\varepsilon)$ for N sufficiently large (note the maximum influence of p'' goes to 0 as $N\to\infty$).

$\mathbf{E}[\mathrm{sgn}(\mathbf{p}''(\mathbf{X}))] \approx_\varepsilon \mathbf{E}[\mathrm{sgn}(\mathbf{p}''(\mathbf{Y}))]$: Since the maximum influence τ of any $x_{i,j}$ in p'' approaches 0 as $N\to\infty$, we can apply Theorem 7.2 for N sufficiently large (and thus τ sufficiently small).

$\mathbf{E}[\mathrm{sgn}(\mathbf{p}''(\mathbf{Y}))] \approx_\varepsilon \mathbf{E}[\mathrm{sgn}(\mathbf{p}(\mathbf{G}'))]$: This case is argued identically as in the first inequality, except that we use anticoncentration of $p''(Y)$, which follows from Lemma 7.4, and we should ensure that we have sufficient independence to apply Theorem D.4 with $t = O(\log(1/\varepsilon))$, which we do.

Proof (of Claim F.4). The claim is argued by showing that for k_i sufficiently close to its expectation (which is $N/2$), the density function of the Gaussian (i.e. the derivative of its CDF) is sufficiently large that the distance we must move from $T_{k_i,N}$ to $T_{k_i+1,N}$ to change the CDF by $\Theta(1/\sqrt{N}) \ge 2^{-N}\binom{N}{k_i+1}$ is small. We argue the case $(1-\epsilon)N/2 \le k_i \le N/2$ since the case $N/2 \le k_i \le (1+\epsilon)N/2$ is argued symmetrically. Also, we consider only the case $k_i = (1-\epsilon)N/2$ exactly, since the magnitude of the standard normal density function is smallest in this case.

Observe that each Z_i is a degree-1 polynomial in the $X_{i,j}$ with maximum influence $1/N$, and thus by the Berry-Esséen Theorem,

$$\sup_{t\in\mathbb{R}} |\mathbf{Pr}[Z_i \le t] - \mathbf{Pr}[G_i \le t]| \le \frac{1}{\sqrt{N}}.$$

Also note that

$$\mathbf{Pr}[G_i \le T_{k_i,N}] = \mathbf{Pr}\left[Z_i \le \frac{2k_i}{\sqrt{N}} - \sqrt{N}\right]$$

by construction. We thus have

$$\begin{aligned}\mathbf{Pr}[G_i \le T_{k_i,N}] &= \mathbf{Pr}\left[G_i \le \frac{2k_i}{\sqrt{N}} - \sqrt{N}\right] \pm \frac{1}{\sqrt{N}} \\ &= \mathbf{Pr}[G_i \le \log^{1/3}(N)] \pm \frac{1}{\sqrt{N}}\end{aligned}$$

By a similar argument we also have

$$\mathbf{Pr}[G_i \le T_{k_i+1,N}] = \mathbf{Pr}\left[G_i \le \log^{1/3}(N) + \frac{2}{\sqrt{N}}\right] \pm \frac{1}{\sqrt{N}}$$

Note though for $t = \Theta(\log^{1/3}(N))$, the density function f of the standard normal satisfies $f(t) = e^{-t^2/2} = N^{-o(1)}$. Thus, in this regime we can change the CDF by $\Theta(1/\sqrt{N})$ by moving only $N^{o(1)}/\sqrt{N} = o(1)$ along the real axis, implying $T_{k_i+1,N} - T_{k_i,N} = o(1)$. ■

■

G Proofs from Section 7.3

G.1 Proof of Theorem 7.6 We begin by stating the following structural lemma:

Theorem G.1. Let $f(x) = \mathrm{sgn}(p(x))$ be any degree-d PTF. Fix any $\tau > 0$. Then f is equivalent to a decision tree $\mathcal{T}$ of depth $\mathrm{depth}(d,\tau) \overset{\mathrm{def}}{=} (1/\tau)\cdot(d\log(1/\tau))^{O(d)}$ with variables at the internal nodes and a degree-d PTF $f_\rho = \mathrm{sgn}(p_\rho)$ at each leaf ρ, with the following property: with probability at least $1-\tau$, a random path from the root reaches a leaf ρ such that either: (i) f_ρ is τ-regular degree-d PTF, or (ii) For any $O(d\cdot\log(1/\tau))$-independent distribution $\mathcal{D}'$ over $\{-1,1\}^{n-|\rho|}$ there exists $b \in \{-1,1\}$ such that $\mathbf{Pr}_{x\sim\mathcal{D}'}[f_\rho(x) \ne b] \le \tau$.

We now prove Theorem 7.6 assuming Theorem G.1. We will need some notation. Consider a leaf of the tree $\mathcal{T}$. We will denote by ρ both the set of variables that appear on the corresponding root-to-leaf path and the corresponding partial assignment; the distinction will be clear from context. Let $|\rho|$ be the number of variables on the path. We identify a leaf ρ with the corresponding restricted subfunction $f_\rho = \mathrm{sgn}(p_\rho)$. We call a leaf "good" if it corresponds to either a τ-regular PTF or to a "close-to constant" function. We call a leaf "bad" otherwise. We denote by $L(\mathcal{T})$, $GL(\mathcal{T})$, $BL(\mathcal{T})$ the sets of leaves, good leaves and bad leaves of $\mathcal{T}$ respectively.

In the course of the proof we make repeated use of the following standard fact:

Fact G.2. Let $\mathcal{D}$ be a k-wise independent distribution over $\{-1,1\}^n$. Condition on any fixed values for any $t \le k$ bits of $\mathcal{D}$, and let $\mathcal{D}'$ be the projection of $\mathcal{D}$ on the other $n-t$ bits. Then $\mathcal{D}'$ is $(k-t)$-wise independent.

Throughout the proof, $\mathcal{D}$ denotes a $(K_d + L_d)$-wise independent distribution over $\{-1,1\}^n$. Consider a random walk on the tree $\mathcal{T}$. Let $LD(\mathcal{T},\mathcal{D})$ (resp. $LD(\mathcal{T},\mathcal{U})$) be the leaf that the random walk will reach when the inputs are drawn from the distribution $\mathcal{D}$ (resp. the uniform distribution). The following straightforward lemma quantifies the intuition that these distributions are the same. This holds because the tree has small depth and $\mathcal{D}$ has sufficient independence.

Lemma G.3. For any leaf $\rho \in L(\mathcal{T})$ we have $\mathbf{Pr}\big[LD(\mathcal{T},\mathcal{D}) = \rho\big] = \mathbf{Pr}\big[LD(\mathcal{T},\mathcal{U}) = \rho\big]$.

The following lemma says that, if ρ is a good leaf, the distribution induced by $\mathcal{D}$ on ρ $O(\varepsilon)$-fools the restricted subfunction f_ρ.

Lemma G.4. Let $\rho \in GL(\mathcal{T})$ be a good leaf and consider the projection $\mathcal{D}_{[n]\setminus\rho}$ of $\mathcal{D}$ on the variables not in ρ. Then we have $\big|\mathbf{Pr}_{x\sim\mathcal{D}_{[n]\setminus\rho}}[f_\rho(x)=1] - \mathbf{Pr}_{y\sim\mathcal{U}_{[n]\setminus\rho}}[f_\rho(y)=1]\big| \le 2\varepsilon$.

Proof. If f_ρ is τ-regular, by Fact G.2 and recalling that $|\rho| \le \mathrm{depth}(d,\tau) \le L_d$, the distribution $\mathcal{D}_{[n]\setminus\rho}$ is K_d-wise independent. Hence, the statement follows by assumption. Otherwise, f_ρ is ε-close to a constant, i.e. there exists $b \in \{-1,1\}$ so that for any $t = O(d\log(1/\tau))$-wise distribution $\mathcal{D}'$ over $\{-1,1\}^{n-|\rho|}$ we have $\mathbf{Pr}_{x\sim\mathcal{D}'}[f_\rho(x) \neq b] \le \tau$ $(*)$. Since $L_d >> t$, Fact G.2 implies that $(*)$ holds both under $\mathcal{D}_{[n]\setminus\rho}$ and $\mathcal{U}_{[n]\setminus\rho}$, hence the statement follows in this case also, recalling that $\tau \le \varepsilon$. ■

The proof of Theorem 7.6 now follows by a simple averaging argument. By the decision-tree decomposition of Theorem G.1, we can write

$$\mathbf{Pr}_{x\sim\mathcal{D}'_n}[f(x)=1] = \sum_{\rho\in L(T)} \mathbf{Pr}\big[LD(\mathcal{T},\mathcal{D}')=\rho\big]\cdot\mathbf{Pr}_{y\in\mathcal{D}'_{[n]\setminus\rho}}\big[f_\rho(y)=1\big]$$

where $\mathcal{D}'$ is either $\mathcal{D}$ or the uniform distribution $\mathcal{U}$. By Theorem G.1 and Lemma G.3 it follows that the probability mass of the bad leaves is at most ε under both distributions. Therefore, by Lemma G.3 and Lemma G.4 we get

$$\Big|\mathbf{Pr}_{x\sim\mathcal{D}}[f(x)=1] - \mathbf{Pr}_{x\sim\mathcal{U}}[f(x)=1]\Big| \le \varepsilon +$$
$$\sum_{\rho\in GL(T)} \mathbf{Pr}\big[LD(\mathcal{T},\mathcal{U})=\rho\big]\cdot\big|\mathbf{Pr}_{y\in\mathcal{U}_{[n]\setminus\rho}}\big[f_\rho(y)=1\big] - \mathbf{Pr}_{y\in\mathcal{D}_{[n]\setminus\rho}}\big[f_\rho(y)=1\big]\big| \le 3\varepsilon.$$

This completes the proof of Theorem 7.6.

G.2 Proof of Theorem G.1 In this section we provide the proof of Theorem G.1. For the sake of completeness, we give below the relevant machinery from [17]. We note that over the hypercube every polynomial can be assumed to be multilinear, and so whenever we discuss a polynomial in this section it should be assumed to be multilinear. We start by defining the notion of the critical index of a polynomial:

Definition G.5 (critical index). Let $p : \{-1,1\}^n \to \mathbb{R}$ and $\tau > 0$. Assume the variables are ordered such that $\mathrm{Inf}_i(p) \ge \mathrm{Inf}_{i+1}(p)$ for all $i \in [n-1]$. The *τ-critical index* of p is the least i such that:

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

If Eq. (G.1) 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.*

We will be concerned with polynomials p of degree-d. The work in [17] establishes useful random restriction lemmas for low-degree polynomials. Roughly, they are as follows: Let p be a degree-d polynomial. If the τ-critical index of p is zero, then $f = \mathrm{sgn}(p)$ is τ-regular and there is nothing to prove.

- If the τ-critical index of p is "very large", then a random restriction of "few" variables causes $f = \mathrm{sgn}(p)$ to become a "close-to-constant" function with probability $1/2^{O(d)}$. We stress that the distance between functions is measured in [17] with respect to the uniform distribution on inputs. As previously mentioned, we extend this statement to hold for any distribution with sufficiently large independence.

- If the τ-critical index of p is positive but not "very large", then a random restriction of a "small" number of variables – the variables with largest influence in p – causes p to become "sufficiently" regular with probability $1/2^{O(d)}$.

Formally, we require the following lemma which is a strengthening of Lemma 10 in [17]:

Lemma G.6. Let $p : \{-1,1\}^n \to \mathbb{R}$ be a degree-d polynomial and assume that its variables are in order of non-increasing influence. Let $0 < \tau', \beta < 1/2$ be parameters. Fix $\alpha = \Theta(d \log\log(1/\beta) + d \log d)$ and $\tau'' = \tau' \cdot (C' d \ln d \ln(1/\tau'))^d$, where C' is a universal constant. One of the following statements holds true:

1. The function $f = \mathrm{sgn}(p)$ is τ'-regular.
2. With probability at least $1/2^{O(d)}$ over a random restriction ρ fixing the first $L' = \alpha/\tau'$ variables of p, the function $f_\rho = \mathrm{sgn}(p_\rho)$ is β-close to a constant function. In particular, under any $O(d \log(1/\beta))$-wise independent distribution $\mathcal{D}'$ there exists $b \in \{-1,1\}$ such that $\mathbf{Pr}_{x \sim \mathcal{D}'}[f_\rho(x) \neq b] \leq \tau'$.
3. There exists a value $k \leq \alpha/\tau'$, such that with probability at least $1/2^{O(d)}$ over a random restriction ρ fixing the first k variables of p, the polynomial p_ρ is τ''-regular.

By applying the above lemma in a recursive manner we obtain Theorem G.1. This is done exactly as in the proof of Theorem 1 in [17]. We remark that in every recursive application of the lemma, the value of the parameter β is set to τ. This explains why $O(d \log(1/\tau))$-independence suffices in the second statement of Theorem G.1. Hence, to complete the proof of Theorem G.1, it suffices to establish Lemma G.6.

Proof (of Lemma G.6). We now sketch the proof of the lemma. The first statement of the lemma corresponds to the case that the value ℓ of τ'-critical index is 0, the second to the case that it is $\ell > L'$ and the third to $1 \leq \ell \leq L'$.

The proof of the second statement proceeds in two steps. Let H denote the first L' most influential variables of p and $T = [n] \setminus H$. Let $p'(x_H) = \sum_{S \subseteq H} \widehat{p}(S) x_S$. We first argue that with probability at least $2^{-\Omega(d)}$ over a random restriction ρ to H, the restricted polynomial $p_\rho(x_T)$ will have a "large" constant term $\widehat{p}_\rho(\emptyset) = p'(\rho)$, in particular at least $\theta = 2^{-\Omega(d)}$. The proof is based on the fact that, since the critical index is large, almost all of the Fourier weight of the polynomial p lies in p', and it makes use of a certain anti-concentration property over the hypercube. Since the randomness is over H and the projection of $\mathcal{D}$ on those variables is still uniform, the argument holds unchanged under $\mathcal{D}$.

In the second step, by an application of a concentration bound, we show that for at least half of these restrictions to H the surviving (non-constant) coefficients of p_ρ, i.e. the Fourier coefficients of the polynomial $p_\rho(x_T) - p'(\rho)$, have small ℓ_2 norm; in particular, we get that $\|p_\rho - p'_\rho\|_2 \leq \log(1/\beta)^{-d}$. We call such restrictions good. Since the projection of $\mathcal{D}$ on these "head" variables is uniform, the concentration bound applies as is.

Finally, we need to show that, for the good restrictions, the event the "tail" variables x_T change the value of the function f_ρ, i.e. $\operatorname{sgn}(p_\rho(x_T) + p'(\rho)) \neq \operatorname{sgn}(p'(\rho))$ has probability at most β. This event has probability at most

$$\mathbf{Pr}_{x_T}[|p_\rho(x_T) - p'(\rho)| \geq \theta].$$

This is done in [17] using a concentration bound on the "tail", assuming full independence. Thus, in this case, we need to modify the argument since the projection of $\mathcal{D}$ on the "tail" variables is not uniform. However, a careful inspection of the parameters reveals that the concentration bound needed above actually holds even under an assumption of $O(d\log(1/\beta))$-independence for the "tail" x_T. In particular, given the upper bound on $\|p_\rho - p'_\rho\|_2$ and the lower bound on θ, it suffices to apply Theorem D.4 for $t = \log(1/\beta)^{d/2}$, which only requires $(dt^{2/d})$-wise independence. Hence, we are done in this case too.

The proof of the third statement remains essentially unchanged for the following reason: One proceeds by considering a random restriction of the variables of p up to the τ-critical index – which in this case is small. Hence, the distribution induced by $\mathcal{D}$ on this space is still uniform. Since the randomness is over these "head" variables, all the arguments remain intact and the claim follows. ■

H Appendix to Section 6.2

We here give a proof of Theorem 6.2.

Theorem 6.2 (restatement). *Let $m > 1$ be an integer. Let $H_i = \{x : \langle a_i, x\rangle > \theta_i\}$ for $i \in [m]$, with $\|a_i\|_2 = 1$ for all i. Let X be a vector of n i.i.d. Gaussians, and Y be a vector of k-wise independent Gaussians. Then for $k = \Omega(m^6/\varepsilon^2)$,*

$$|\mathbf{Pr}[X \in \cap_{i=1}^m H_i] - \mathbf{Pr}[Y \in \cap_{i=1}^m H_i]| < \varepsilon$$

Proof. Define $F : \mathbb{R}^n \to \mathbb{R}^m$ by $F(x) = (\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle)$, and let R be the region $\{x : \forall i\ x_i > \theta_i\}$. Similarly as in Section 5, we show a chain of inequalities after setting $\rho = \varepsilon/m$ and $c = m/\rho$:

$$\mathbf{E}[I_R(F(X))] \approx_\varepsilon \mathbf{E}[\tilde{I}_R^c(F(X))] \approx_\varepsilon \mathbf{E}[\tilde{I}_R^c(F(Y))] \approx_\varepsilon \mathbf{E}[I_R(F(Y))]. \tag{H.1}$$

For the first inequality, observe $d_2(x, \partial R) \geq \min_i\{|x_i - \theta_i|\}$. Then by a union bound,

$$\mathbf{Pr}[d_2(F(X), \partial R) \leq w] \leq \mathbf{Pr}[\min_i\{|\langle a_i, X\rangle - \theta_i|\} \leq w] \leq \sum_{i=1}^m \mathbf{Pr}[|\langle a_i, X\rangle - \theta_i| \leq w] = O(mw).$$

Now,

$$\begin{aligned}
|\mathbf{E}[I_R(F(X))] - \mathbf{E}[\tilde{I}_R^c(F(X))]| &\leq \mathbf{E}[|I_R(F(X))] - \tilde{I}_R^c(F(X))|] \\
&\leq \mathbf{Pr}[d_2(F(X), \partial R) \leq 2\rho] \\
&\quad + O\left(\sum_{s=1}^\infty \left(\frac{m^2}{c^2 2^{2s}\rho^2}\right) \cdot \mathbf{Pr}[d_2(F(X), \partial R) \leq 2^{s+1}\rho]\right)
\end{aligned} \tag{H.2}$$

$$
\begin{aligned}
&= \mathbf{Pr}[d_2(F(X), \partial R) \le 2\rho] + O\left(\sum_{s=1}^{\infty} 2^{-2s} \cdot \mathbf{Pr}[d_2(F(X), \partial R) \le 2^{s+1}\rho]\right) \\
&= O(m\rho) \\
&= O(\varepsilon)
\end{aligned}
$$

where Eq. (H.2) follows from Theorem 4.2.

The last inequality in Eq. (H.1) is argued identically, except that we need to have anticoncentration of the $|\langle a_i, Y\rangle|$ in intervals of size no smaller than $\rho = \varepsilon/m$; this was already shown to hold under $\Omega(1/\rho^2)$-wise independence in the proof of Theorem 5.1.

For the middle inequality we use Taylor's theorem, as was done in Section 5. If we truncate the Taylor polynomial at degree-$(k-1)$ for k even, then by our derivative bounds on mixed partials of $\tilde{I}_R^c$ from Theorem 4.1, the error term is bounded by

$$(2c)^k \cdot m^k \cdot \frac{\sum_{i=1}^m \mathbf{E}[\langle a_i, X\rangle^k]}{k!} \le (cm)^k \cdot 2^{O(k)}/k^{k/2},$$

with the inequality holding by Lemma B.1, and the m^k arising as the analogue of the 4^k term that arose in Eq. (F.1). This is at most ε for k a sufficiently large constant times $(cm)^2$, and thus overall $k = \Omega(m^6/\varepsilon^2)$-wise independence suffices. ■

Remark H.1. A couple improvements are possible to reduce the dependence on m in Theorem 6.2. We presented the simplest proof we are aware of which obtains a polynomial dependence on m, for clarity of exposition. See Section I for an improvement on the dependence on m to quartic.

Our approach can also show that bounded independence fools the intersection of any constant number m of degree-2 threshold functions. Suppose the degree-2 polynomials are $p_1, \ldots, p_m$. Exactly as in Section 7.2 we decompose each p_i into $p_{i,1} - p_{i,2} + p_{i,3} + p_{i,4} + C_i$. We then define a region $R \subset \mathbb{R}^{4m}$ by $\{x : \forall i \in [m]\ x_{4i-3}^2 - x_{4i-2}^2 + x_{4i-1} + x_{4i} + C_i + \mathrm{tr}(A_{p_{i,3}}) > 0\}$, and the map $F : \mathbb{R}^n \to \mathbb{R}^{4m}$ by

$$F(x) = (M_{p_1}(X), \ldots, M_{p_n}(X))$$

for the map $M_p : \mathbb{R}^n \to \mathbb{R}^4$ defined in Section 7.2. The goal is then to show $\mathbf{E}[I_R(F(X))] \approx_\varepsilon \mathbf{E}[I_R(F(Y))]$, which is done identically as in the proof of Theorem 7.2. We simply state the theorem here:

Theorem H.2. Let $m > 1$ be an integer. Let $H_i = \{x : p_i(x) \ge 0\}$ for $i \in [m]$, for some degree-2 polynomials $p_i : \mathbb{R}^n \to \mathbb{R}$. Let X be a vector of n i.i.d. Gaussians, and Y be a vector of k-wise independent Gaussians with $k = \Omega(\mathrm{poly}(m)/\varepsilon^8)$. Then,

$$|\mathbf{Pr}[X \in \cap_{i=1}^m H_i] - \mathbf{Pr}[Y \in \cap_{i=1}^m H_i]| < \varepsilon$$

Identical conclusions also hold for X, Y being drawn from $\{-1,1\}^n$, since we can apply the decision tree argument from Theorem G.1 to each of the m polynomial threshold functions separately so that, by a union bound, with probability at least $1 - m\tau'$ each of the m PTF restrictions is either τ'-close to a constant function, or is τ'-regular. Thus for whatever setting of τ sufficed for the case $m = 1$ ($\tau = \varepsilon^2$ for halfspaces [16] and $\tau = \varepsilon^9$ for degree-2 threshold functions (Theorem 7.2)), we set $\tau' = \tau/m$ then argue identically as before.

I Improvements to fooling the intersection of halfspaces

In the proof of Theorem 6.2 in Section H, we presented a proof showing that $\Omega(m^6/\varepsilon^2)$-independence ε-fools the intersection of m halfspaces under the Gaussian measure. In fact, this dependence on m can be improved to quartic. One factor of m is shaved by using Lemma A.6, and another factor of m is shaved by a suitable change of basis. The argument used to shave the second factor of m is specific to the Gaussian case, and does not carry over to the Bernoulli setting.

Theorem I.1. *Let $m > 1$ be an integer. Let $H_i = \{x : \langle a_i, x\rangle > \theta_i\}$ for $i \in [m]$, with $\|a_i\|_2 = 1$ for all i. Let X be a vector of n independent standard normals, and Y be a vector of k-wise independent Gaussians. Then for $k = \Omega(m^4/\varepsilon^2)$ and even,*

$$|\mathbf{Pr}[X \in \cap_{i=1}^m H_i] - \mathbf{Pr}[Y \in \cap_{i=1}^m H_i]| < \varepsilon$$

Proof. Let $v_1, \ldots, v_m \in \mathbb{R}^n$ be an orthonormal basis for a linear space containing the a_i. Define the region $R = \{x : \forall i \in [m] \ \sum_{j=1}^m \langle a_i, v_j\rangle x_j > \theta_i\}$ in $\mathbb{R}^m$. Note R is itself the intersection of m halfspaces in $\mathbb{R}^m$, with the ith halfspace having normal vector $b_i \in \mathbb{R}^m$ with $(b_i)_j = \langle a_i, v_j\rangle$.

We now define the map $F : \mathbb{R}^n \to \mathbb{R}^m$ by $F(x) = (\langle x, v_1\rangle, \ldots, \langle x, v_m\rangle)$. It thus suffices to show that $\mathbf{E}[I_R(F(X))] \approx_\varepsilon \mathbf{E}[I_R(F(Y))]$. We do this by a chain of inequalities, similarly as in the proof of Theorem 6.2. Below we set $c = m^2/\varepsilon$.

$$\mathbf{E}[I_R(F(X))] \approx_\varepsilon \mathbf{E}[\tilde{I}_R^c(F(X))] \approx_\varepsilon \mathbf{E}[\tilde{I}_R^c(F(Y))] \approx_\varepsilon \mathbf{E}[I_R(F(Y))]. \tag{I.1}$$

For the first inequality and last inequalities, since we performed an orthonormal change of basis the $F(X)_i$ remain independent standard normals, and we can reuse the same analysis from the proof of Theorem 6.2 without modification.

For the middle inequality we use Taylor's theorem. Let P_{k-1} the degree-$(k-1)$ Taylor polynomial approximating $\tilde{I}_R^c$. Then by Lemma A.6,

$$|\tilde{I}_R^c(F(x)) - P_{k-1}(F(x))| \le \frac{2^{O(k)} \cdot c^k \cdot \|F(x)\|_2^k}{k^k} \tag{I.2}$$

Since the $F(X)_i$ are independent standard normal random variables, $\sum_{i=1}^m F(X)_i^2$ follows a chi-squared distribution with m degrees of freedom, and its $k/2$th moment is determined by k-wise independence, and thus

$$\mathbf{E}\left[\left(\sum_{i=1}^m F(X)_i^2\right)^{k/2}\right] = 2^{k/2} \cdot \frac{\Gamma(k/2 + m/2)}{\Gamma(m/2)} = 2^{O(k)} \cdot k^m \cdot k^{k/2} \le 2^{O(k)} \cdot k^{k/2}. \tag{I.3}$$

Thus, the expected value of our Taylor error is $2^{O(k)} \cdot (c/\sqrt{k})^k = O(\varepsilon)$ for $k = \Omega(c^2)$. ■

J An FT-mollification proof of a multivariate Jackson's theorem

We remind the reader of the setup in Section 6.1. We have $F : \mathbb{R}^m \to \mathbb{R}$ and define

$$\omega(F, \delta) = \sup_{\substack{\|x\|_2, \|y\|_2 \le 1 \\ \|x-y\|_2 \le \delta}} |F(x) - F(y)|.$$

Given some positive integer k, we would like to construct a polynomial p_k of degree k such that $\sup_{\|x\|_2 \le 1} |F(x) - p_k(x)| = O(\omega(F, m/k))$. We show here how this can be achieved via FT-mollification followed by Taylor's theorem.

Define the function G to be $F - F(0)$ in the ball of radius 2 about the origin and 0 otherwise. Now, consider the FT-mollification $\tilde{G}^c$, and suppose $\|x\|_2 \le 1$. Then,

$$\begin{aligned}
\left|G(x) - \tilde{G}^c(x)\right| &= \left|\int_{\mathbb{R}^m} (G(x) - G(x-y))B_c(y)\right| \\
&\le \mathbf{E}_{y \sim B_c}[|G(x) - G(x-y)|] \\
&\le \omega(F, m/k) + \sum_{s=0}^{\infty} \mathbf{Pr}_{y \sim B_c}[2^s m/k < \|y\|_2 \le 2^{s+1} m/k] \cdot 2^{s+1} \cdot \omega(F, m/k) \\
&\le \omega(F, m/k) + \sum_{s=0}^{\infty} \left(\frac{\mathbf{E}_{y \sim B_c}[\|y\|_2^2]}{(2^s m/k)^2}\right) \cdot 2^{s+1} \cdot \omega(F, m/k) \\
&= O(\omega(F, m/k) \cdot (k/c)^2)
\end{aligned}$$

Now, let q_k be the degree-k Taylor expansion of $\tilde{G}^c$ about 0. By Lemma A.6, $|\tilde{G}^c(x) - q_k(x)| \le (2k/m) \cdot \omega(F, m/k) \cdot (D \cdot c/k)^k$ for some absolute constant D, since $\|G\|_\infty \le (2k/m) \cdot \omega(F, m/k)$. We thus obtain $|G(x) - q_k(x)| \le O(\omega(F, m/k))$ by setting $c = k/(2D)$. Finally, set $p_k = q_k + F(0)$.