Detecting Low-Degree Truncation

Anindya De
University of Pennsylvania
Philadelphia, USA
de.anindya@gmail.com

Huan Li
University of Pennsylvania
Philadelphia, USA
huanli@cis.upenn.edu

Shivam Nadimpalli
Columbia University
New York, USA
sn2855@columbia.edu

Rocco A. Servedio
Columbia University
New York, USA
ras2105@columbia.edu

ACM Reference Format:
Anindya De, Huan Li, Shivam Nadimpalli, and Rocco A. Servedio. 2024. Detecting Low-Degree Truncation. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC '24), June 24–28, 2024, Vancouver, BC, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3618260.3649633

1 INTRODUCTION

One of the most basic and natural ways that a probability distribution $\mathcal{D}$ can be altered is by *truncating* it, i.e. conditioning on some subset of possible outcomes. Indeed, the study of truncated distributions is one of the oldest topics in probability and statistics: already in the 19th century, Galton [Gal97] attempted to estimate the mean and standard deviation of the running times of horses on the basis of sample data that did not include data for horses that were slower than a particular cutoff value. Since the running times were assumed to be normally distributed, this was an early attempt to infer the parameters of an unknown normal distribution given samples from a truncated version of the distribution. Subsequent early work by other statistical pioneers applied the method of moments [Pea02, Lee14] and maximum likehood techniques [Fis31] to the same problem of estimating the parameters of an unknown univariate normal distribution from truncated samples. The study of truncation continues to be an active area in contemporary statistics (see [Sch86, BC14, Coh16] for recent books on this topic).

Quite recently, a number of research works in theoretical computer science have tackled various algorithmic problems that deal with *high-dimensional* truncated data. Much of this work attempts to *learn* a parametric description of an unknown distribution that has been subject to truncation. For example, in [DGTZ19] Daskalakis, Gouleakis, Tzamos and Zampetakis gave an efficient algorithm for high-dimensional truncated linear regression, and in [DGTZ18] Daskalakis, Gouleakis, Tzamos and Zampetakis gave an efficient algorithm for estimating the mean and covariance of an unknown multivariate normal distribution from truncated samples given access to a membership oracle for the truncation set. In [FKT20] Fotakis, Kalavasis and Tzamos gave a similar result for the setting in which the unknown background distribution is a product distribution over $\{0,1\}^d$ instead of a multivariate normal distribution, and in [KTZ19] Kontonis, Tzamos and Zampetakis extended the results of [DGTZ18] to the case of an unknown truncation set satisfying certain restrictions. In summary, the recent work described above has focused on learning (parameter estimation) of truncated normal distributions and product distributions over $\{0,1\}^d$.

This Work: *Detecting* Truncation. In the current paper, rather than the learning problem we study what is arguably the most basic problem that can be considered in the context of truncated data — namely, detecting whether or not truncation has taken place at all. A moment's thought shows that some assumptions are required in order for this problem to make sense: for example, if the truncation set is allowed to be arbitrarily tiny (so that only an arbitrarily small fraction of the distribution is discarded by truncation), then it can be arbitrarily difficult to detect whether truncation has taken place. It is also easy to see that truncation cannot be detected if the unknown truncation set is allowed to be arbitrarily complex. Thus, it is natural to consider a problem formulation in which there is a fixed class of possibilities for the unknown truncation set; this is the setting we consider.

We note that the truncation detection problem we consider has a high-level resemblance to the standard *hypothesis testing* paradigm in statistics, in which the goal is to distinguish between a "null hypothesis" and an "alternate hypothesis." In our setting the null hypothesis corresponds to no truncation of the known distribution having taken place, and the alternate hypothesis is that the known distribution has been truncated by some unknown truncation set belonging to the fixed class of possibilities. However, there does not appear to be work in the statistics literature which deals with the particular kinds of truncation problems considered in this work, let alone computationally efficient algorithms for those problems.

Prior Work: Convex Truncation of Normal Distributions. Recent work [DNS23] considered the truncation detection problem in a setting where the background distribution $\mathcal{D}$ is the standard multidimensional normal distribution $N(0,1)^n$ and the truncation set is assumed to be an unknown *convex set* in $\mathbb{R}^n$. This specific problem formulation enabled the use of a variety of sophisticated tools and results from high-dimensional convex geometry, such as Gaussian

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
STOC '24, June 24–28, 2024, Vancouver, BC, Canada
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0383-6/24/06…$15.00
https://doi.org/10.1145/3618260.3649633

isoperimetry and the Brascamp-Lieb inequality [BL76b] and extensions thereof due to Vempala [Vem10]. Using these tools, [DNS23] gave several different algorithmic results and lower bounds. Chief among these were (i) a polynomial-time algorithm that uses $O(n/\varepsilon^2)$ samples and distinguishes the non-truncated standard normal distribution $N(0,1)^n$ from $N(0,1)^n$ conditioned on a convex set of Gaussian volume at most $1-\varepsilon$; and (ii) a $\tilde{\Omega}(\sqrt{n})$-sample lower bound for detecting truncation by a convex set of constant volume.

The results of [DNS23] provide a "proof of concept" that in sufficiently well-structured settings it can sometimes be possible to detect truncation in a computationally and statistically efficient way. This serves as an invitation for a more general study of truncation detection; in particular, it is natural to ask whether strong structural or geometric assumptions like those made in [DNS23] (normal distribution, a convex truncation set) are required in order to achieve nontrivial algorithmic results. Can efficient algorithms detect truncation for broader classes of "background" distributions beyond the standard normal distribution, or for other natural families of truncations besides convex sets? This question is the motivation for the current work.

This Work: "Low-Degree" Truncation of Hypercontractive Product Distributions. In this paper we consider

- A broader range of possibilities for the background distribution $\mathcal{D}$ over $\mathbb{R}^n$, encompassing many distributions which may be either continuous or discrete; and
- A family of non-convex truncation sets corresponding to *low-degree polynomial threshold functions*.

Recall that a Boolean-valued function $f : \mathbb{R}^n \to \{0,1\}$ is a *degree-d polynomial threshold function* (PTF) if there is a real multivariate polynomial $p(x)$ with $\deg(p) \le d$ such that $f(x) = 1$ if and only if $p(x) \ge 0$. Low-degree polynomial threshold functions are a well-studied class of Boolean-valued functions which arise naturally in diverse fields such as computational complexity, computational learning theory, and unconditional derandomization, see e.g. [BS92, GL94, HKM14, DRST14, Kan14a, DOSW11, CDS20, DKPZ21, BHYY22, DKN10, Kan11b, Kan11a, Kan12, KM13, MZ13, Kan14b, DDS14, DS14, Kan15, KKL17, KL18, KR18, ST18, BV19, OST20] among many other references.

Our main results, described in the next subsection, are efficient algorithms and matching information-theoretic lower bounds for detecting truncation by low-degree polynomial threshold functions for a wide range of background distributions and parameter settings.

1.1 Our Results

To set the stage for our algorithmic results, we begin with the following observation:

Observation 1. For any fixed ("known") background distribution $\mathcal{P}$ over $\mathbb{R}^n$, $O_d(n^d/\varepsilon^2)$ samples from an unknown distribution $\mathcal{D}$ are sufficient to distinguish (with high probability) between the two cases that (i) $\mathcal{D}$ is the original distribution $\mathcal{P}$, versus (ii) $\mathcal{D}$ is $\mathcal{P}|_f$, i.e. $\mathcal{P}$ conditioned on $f^{-1}(1)$, where f is an unknown degree-d PTF satisfying $\mathbf{Pr}_{\boldsymbol{x}\sim\mathcal{P}}[f(\boldsymbol{x}) = 1] \le 1-\varepsilon$.

This is an easy consequence of a standard uniform convergence argument using the well-known fact that the Vapnik-Chervonenkis dimension of the class of all degree-d polynomial threshold functions over $\mathbb{R}^n$ is $O(n^d)$. For the sake of completeness, we give a proof in Appendix A of the full version of this paper.

While the above observation works for any fixed background distribution $\mathcal{P}$, several drawbacks are immediately apparent. One is that a sample complexity of $O(n^d)$ is quite high, in fact high enough to information-theoretically learn an unknown degree-d PTF; are this many samples actually required for the much more modest goal of merely *detecting* whether truncation has taken place? A second and potentially more significant issue is that the above VC-based algorithm is computationally highly inefficient, involving a brute-force enumeration over "all" degree-d PTFs; for a sense of how costly this may be, recall that even in the simple discrete setting of the uniform distribution over the Boolean cube $\{-1,1\}^n$ there are $2^{\Omega(n^{d+1})}$ distinct degree-d PTFs over $\{-1,1\}^n$ for constant d [Sak93, Theorem 2.34]. So it is natural to ask whether there exist more efficient (either in terms of running time or sample complexity) algorithms for interesting cases of the truncation detection problem, and to ask about lower bounds for this problem.

On the lower bounds side, it is natural to first consider arguably the simplest case, in which $\mathcal{P}$ is the uniform distribution $\mathcal{U}$ over the Boolean hypercube $\{-1,+1\}^n$. In this setting we have the following observation:

Observation 2. If the truncating PTF f is permitted to have as few as $n^{d/2}$ satisfying assignments, then any algorithm that correctly decides whether its samples come from $\mathcal{D} = \mathcal{U}$ versus from $\mathcal{D} = \mathcal{U}|_f$ must use $\Omega(n^{d/4})$ samples.

This lower bound can be established using only basic linear algebra and simple probabilistic arguments; it is inspired by the "voting polynomials" lower bound of Aspnes et al. [ABFR94] against MAJ-of-AC^0 circuits. We give the argument in Appendix B of the full version of this paper.

Taken together, there is a quartic gap between the (computationally inefficient) upper bound given by Observation 1 and the information-theoretic lower bound of Observation 2 for PTFs with extremely few satisfying assignments. Our main result is a proof that the true complexity of the truncation distinguishing problem lies exactly in the middle of these two extremes. We:

(i) Give a *computationally efficient* distinguishing algorithm which has sample complexity $O(n^{d/2})$ for a wide range of product distributions and values of $\mathrm{Vol}(f)$, and
(ii) Show that even for the uniform background distribution $\mathcal{U}$ over $\{-1,+1\}^n$, distinguishing whether or not $\mathcal{U}$ has been truncated by a degree-d PTF of volume $\approx 1/2$ requires $\Omega(n^{d/2})$ samples.

We now describe our results in more detail.

An Efficient Algorithm. We give a truncation distinguishing algorithm which succeeds if $\mathcal{P}$ is any multivariate i.i.d. product distribution $\mathcal{P} = \mu^{\otimes n}$ over $\mathbb{R}^n$ satisfying a natural *hypercontractivity* property and if $\mathrm{Vol}(f)$ is "not too small." We defer the precise technical definition of the (fairly standard) hypercontractivity property that we require to Section 2.2, and here merely remark that a wide range of i.i.d. product distributions satisfy the required condition, including the cases where μ is

- any fixed distribution over $\mathbb{R}$ that is supported on a finite (independent of n) number of points;
- any normal distribution $N(c, \sigma^2)$ where c, σ are independent of n;
- any uniform distribution over a continuous interval $[a, b]$;
- any distribution which is supported on an interval $[a, b]$ for which there are two constants $0 < c < C$ such that everywhere on $[a, b]$ the pdf is between $c/(b-a)$ and $C/(b-a)$.

An informal statement of our main positive result is below:

Theorem 3 (Efficiently detecting PTF truncation, informal theorem statement)**.** Let $0 < \varepsilon < 1$. Fix any constant d and any hypercontractive i.i.d. product distribution $\mu^{\otimes n}$ over $\mathbb{R}^n$. Let $f : \mathbb{R}^n \to \{0, 1\}$ be an unknown degree-d PTF such that

$$1 - \varepsilon \geq \Pr_{\boldsymbol{x} \sim \mu^{\otimes n}}[f(\boldsymbol{x}) = 1] \geq 2^{-O(\sqrt{n})}.$$

There is an efficient algorithm that uses $\Theta(n^{d/2}/\varepsilon^2)$ samples from $\mathcal{D}$ and successfully (w.h.p.) distinguishes between the following two cases:

(i) $\mathcal{D}$ is $\mu^{\otimes n}$, i.e. the "un-truncated" distribution; versus
(ii) $\mathcal{D}$ is $\mu^{\otimes n}|_f$, i.e. $\mu^{\otimes n}$ truncated by f.

Note that ε is a lower bound on the probability mass of the distribution $\mu^{\otimes n}$ which has been "truncated;" as remarked earlier, without a lower bound on ε, it can be arbitrarily difficult to distinguish the truncated distribution. Thus, as long as the background distribution is a "nice" i.i.d. product distribution and the truncating PTF's volume is "not too tiny", in polynomial time we can achieve a square-root improvement in sample complexity over the naive brute-force computationally inefficient algorithm.

A Matching Lower Bound. It is natural to wonder whether Theorem 3 is optimal: Can we establish lower bounds on the sample complexity of determining whether a "nice" distribution has been truncated by a PTF? And can we do this when the truncating PTF (unlike in Observation 2) has volume which is not extremely small?

Our main lower bound achieves these goals; it shows that even for the uniform distribution $\mathcal{U}$ over $\{-1, 1\}^n$ and for PTFs of volume $\approx 1/2$, the sample complexity achieved by our algorithm in Theorem 3 is best possible up to constant factors.

Theorem 4 (Lower bound for detecting PTF truncation, informal theorem statement)**.** Fix any constant d. Let $f : \{-1, 1\}^n \to \mathbb{R}$ be an unknown degree-d PTF such that $\mathbf{Pr}_{\boldsymbol{x} \sim \mathcal{U}}[f(\boldsymbol{x}) = 1] \in [0.49, 0.51]$. Any algorithm that uses samples from $\mathcal{D}$ and successfully (w.h.p.) distinguishes between the cases that (i) $\mathcal{D}$ is $\mathcal{U}$, versus (ii) $\mathcal{D}$ is $\mathcal{U}|_f$, must use $\Omega(n^{d/2})$ samples.

1.2 Techniques

We now give a technical overview of both the upper bound (Theorem 3) and the lower bound (Theorem 4), starting with the former.

1.2.1 Overview of Theorem 3. For simplicity, we start by considering the case when the background distribution $\mathcal{P} = \mu^{\otimes n}$ is the uniform measure on the Boolean hypercube.

The Boolean Hypercube $\{-1, +1\}^n$. Let us denote the uniform measure over $\{-1, +1\}^n$ by $\mathcal{U}_n$. Recall that our goal is to design an algorithm with the following performance guarantee: Given i.i.d. sample access to an unknown distribution $\mathcal{D}$, the algorithm w.h.p.

(i) Outputs "un-truncated" when $\mathcal{D} = \mathcal{U}_n$; and
(ii) Outputs "truncated" when $\mathcal{D} = \mathcal{U}_n|_{f^{-1}(1)}$ for a degree-d PTF $f : \{-1, +1\}^n \to \{0, 1\}$, where $1 - \varepsilon \geq \mathbf{Pr}[f(\boldsymbol{x}) = 1]$ for $\boldsymbol{x} \sim \mathcal{U}_n$.

To avoid proliferation of parameters, we set $\varepsilon = 0.1$ for the rest of the discussion. We thus have

$$\Pr_{\boldsymbol{x} \sim \mathcal{U}_n}\left[f(\boldsymbol{x}) = 1\right] \leq 0.9.$$

For any point $x \in \{-1, +1\}^n$, let $\widetilde{x} \in \{-1, +1\}^{\binom{n}{1}+\ldots+\binom{n}{d}}$ be the vector given by

$$\widetilde{x} := \left(\widetilde{x}_\alpha\right)_{\substack{\alpha \subseteq [n] \\ 0 < |\alpha| \leq d}} \qquad \text{where } \widetilde{x}_\alpha := \prod_{i \in \alpha} x_i.$$

In other words, every coordinate of $\widetilde{x}$ corresponds to a non-constant monomial in x of (multilinear) degree at most d. Note that the map $x \mapsto \widetilde{x}$ can be viewed as a *feature map* corresponding to the "polynomial kernel" in learning theory.

The main idea underlying our algorithm, which is given in Theorem 3, is the following:

(1) When $\mathcal{D} = \mathcal{U}_n$, then it is easy to see that $\mathbf{E}\left[\widetilde{\boldsymbol{x}}\right] = \overline{0}$ (the all-0 vector). This is immediate from the fact that the expectation of any non-constant monomial under $\mathcal{U}_n$ is 0.
(2) On the other hand, suppose $\mathcal{D} = \mathcal{U}_n|_{f^{-1}(1)}$ for a degree-d PTF f as above. In this case, it can be shown that

$$\left\| \mathop{\mathbf{E}}_{\mathcal{U}_n|_{f^{-1}(1)}}\left[\widetilde{\boldsymbol{x}}\right] \right\|_2 \geq 2^{-\Theta(d)} =: c_d.$$

This is done by relating the quantity in the LHS above to the Fourier spectrum of degree-d PTFs, which has been extensively studied in concrete complexity theory (see for example [GL94, DRST14, HKM14, Kan13]). In particular, we obtain this lower bound on $\left\|\mathbf{E}\left[\widetilde{\boldsymbol{x}}\right]\right\|_2$ from an anti-concentration property of low-degree polynomials over the Boolean hypercube. This in turn is a consequence of hypercontractivity of the uniform measure over $\{-1, +1\}^n$, a fundamental tool in discrete Fourier analysis (see Section 9.5 of [O'D14]).

Items 1 and 2 above together imply that estimating $\left\|\mathbf{E}\left[\widetilde{\boldsymbol{x}}\right]\right\|_2^2$ up to an additive error of $\pm c_d^2/2$ suffices to distinguish between $\mathcal{D} = \mathcal{U}_n$ and $\mathcal{D} = \mathcal{U}_n|_{f^{-1}(1)}$. Next, note that

$$\left\| \mathop{\mathbf{E}}_{\mathcal{D}}\left[\widetilde{\boldsymbol{x}}\right] \right\|_2^2 = \mathop{\mathbf{E}}_{\boldsymbol{x}, \boldsymbol{y} \sim \mathcal{D}}\left[\left\langle \widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}} \right\rangle\right].$$

Using the idea of "U-statistics" [Hoe94], this suggests a natural unbiased estimator, namely drawing $2T$ points $\widetilde{\boldsymbol{x}}^{(1)}, \ldots, \widetilde{\boldsymbol{x}}^{(T)}$ and $\widetilde{\boldsymbol{y}}^{(1)}, \ldots, \widetilde{\boldsymbol{y}}^{(T)}$ for some T which we will fix later, and then setting

$$\mathbf{M} := \frac{1}{T^2}\left\langle \sum_{i=1}^{T} \widetilde{\boldsymbol{x}}^{(i)}, \sum_{j=1}^{T} \widetilde{\boldsymbol{y}}^{(j)} \right\rangle.$$

In particular, we have $\mathbf{E}[\mathbf{M}] = \left\|\mathbf{E}\left[\widetilde{\boldsymbol{x}}\right]\right\|_2^2$.

In order to be able to distinguish between the un-truncated and truncated distributions by estimating $\mathbf{M}$ (and then appealing to

Chebyshev's inequality), it therefore suffices to upper bound the variance of $\mathbf{M}$ in both the un-truncated and the truncated settings. When $\mathcal{D} = \mathcal{U}_n$, then $\mathbf{Var}[\mathbf{M}]$ is straightforward to calculate and it turns out that

$$\mathop{\mathbf{Var}}_{\mathcal{U}_n}[\mathbf{M}] = \frac{m}{T^2} \qquad \text{where } m := \#\{\alpha \subseteq [n] : 0 < |\alpha| \le d\} = O_d(n^d).$$

However, in the truncated setting, $\mathbf{Var}[\mathbf{M}]$ is significantly trickier to analyze; at a high-level, our analysis expresses $\mathbf{Var}[\mathbf{M}]$ in terms of the "weights" of various levels of the Fourier spectrum of f.[1] The key technical ingredient we use to control the variance is the so-called "level-k inequality" for Boolean functions, which states that for any Boolean function $f : \{-1,+1\}^n \to \{0,1\}$, writing $\mathbf{W}^{=k}[f]$ for the "Fourier weight at level-k", we have

$$\mathbf{W}^{=k}[f] \le O_k\left(\mathbf{W}^{=0}[f] \cdot \log^k\left(\frac{1}{\mathbf{W}^{=0}[f]}\right)\right).$$

Recall that $\mathbf{W}^{=0}[f] = \mathbf{E}_{\mathcal{U}_n}[f]^2$, and so the level-$k$ inequality bounds higher-level Fourier weight in terms of the mean of the function. We remark that the level-k inequality is also a consequence of hypercontractivity over the Boolean hypercube (as before, see Section 9.5 of [O'D14]). With this in hand, we can show that

$$\mathop{\mathbf{Var}}_{\mathcal{U}_n|_{f^{-1}(1)}}[\mathbf{M}] = \frac{O_d(n^d)}{T^2}$$

as long as $\mathbf{E}_{\mathcal{U}_n}[f] \ge 2^{-\sqrt{n}}$.

Finally, taking $T = \Theta_d(n^{d/2})$ implies that the standard deviation of each of our estimators is comparable to the difference in means (which was c_d^2), allowing us to distinguish between the un-truncated and truncated settings.

Hypercontractive Product Distributions $\mu^{\otimes n}$. We use the same high-level approach (as well as the same estimator $\mathbf{M}$) in order to distinguish low-degree truncation of a hypercontractive product measure $\mu^{\otimes n}$, but the analysis becomes more technically involved. To explain the principal challenge, note that over the Boolean hypercube $\{-1,+1\}^n$, the Fourier basis functions $(\chi_\alpha)_{\alpha\subseteq[n]}$,

$$\chi_\alpha(x) := \prod_{i\in\alpha} x_i,$$

form a multiplicative group. This group structure is useful because it means that the product of two basis functions is another basis function: For $\alpha, \beta \subseteq [n]$, we have the product formula

$$\chi_\alpha \cdot \chi_\beta = \chi_{\alpha \,\triangle\, \beta}$$

where $\alpha \triangle \beta$ denotes the symmetric difference of α and β.

Over an arbitrary hypercontractive measure $\mu^{\otimes n}$, this may no longer be the case; as a concrete example, this fails for the Gaussian measure and the Hermite basis (cf. Chapter 11 of [O'D14]). Over a general product space $\mu^{\otimes n}$ the Fourier basis functions are now indexed by *multi-subsets* of $[n]$ (as opposed to subsets of $[n]$ over $\{-1,+1\}^n$)—see the discussion following Definition 5. More importantly, there is no simple formula for the product of two Fourier basis functions, and this makes the analysis technically more involved. We remark that this problem, which is known as the *linearization problem*, has been well studied for various classes of orthogonal polynomials (see Section 6.8 of [AAR99]). Lemmas 16 and 17 establish a weak version of a "product formula" between two Fourier basis functions for $\mu^{\otimes n}$ that suffices for our purposes and lets us carry out an analysis similar to the above sketch for the Boolean hypercube $\{-1,+1\}^n$.

[1] See Section 2.1 for a formal definition.

1.2.2 Overview of Theorem 4. We turn to an overview of our lower bound, Theorem 4. As in the previous section, we write $\mathcal{U}_n$ to denote the uniform distribution over the n-dimensional Boolean hypercube $\{-1,+1\}^n$ and $(\chi_S)_S$ for the Fourier basis over $\{-1,+1\}^n$. To prove Theorem 4, it suffices to construct a distribution $\mathcal{F}_d$ over degree-d PTFs over $\{-1,+1\}^n$ with the following properties:

(1) The distribution $\mathcal{F}_d$ is supported on thresholds of homogenous degree-d polynomials over $\{-1,+1\}^n$. Note that such polynomials are necessarily multilinear; in particular, each PTF $\boldsymbol{f} \sim \mathcal{F}_d$ can be expressed as

$$\boldsymbol{f}(x) := \mathbf{1}\left\{\sum_{S:|S|=d} \widehat{\boldsymbol{p}}(S)\chi_S(x) \ge 0\right\}.$$

The coefficients $\widehat{\boldsymbol{p}}(S)$ will be i.i.d. random variables drawn from the standard Gaussian distribution $N(0,1)$.

(2) Let $m = \Omega(n^{d/2})$ and consider the distributions
- $\mathcal{D}_1$, obtained by drawing m independent samples from $\mathcal{U}_n$; and
- $\mathcal{D}_2$, obtained by first drawing $\boldsymbol{f} \sim \mathcal{F}_d$, and then drawing m independent samples from $\mathcal{U}_n|_{f^{-1}(1)}$.

Then distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ are $o(1)$-close to each other in variation distance.

Polynomials of the form

$$\sum_S \widehat{\boldsymbol{p}}(S)\chi_S(x) \qquad \text{for } \widehat{\boldsymbol{p}}(S) \sim N(0,1)$$

are known in the literature as *Gaussian random polynomials*, and have been extensively studied (with an emphasis on the behavior of their roots) [IZ97, Ham56, BS14]. We will however be interested in a certain "pseudorandom-type" behavior of these polynomials.

In particular, we first reduce the problem of proving indistinguishability of $\mathcal{D}_1$ and $\mathcal{D}_2$ to proving the following: Suppose $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ are m randomly chosen points from $\{-1,1\}^n$ (which we fix). Then, with probability $1 - o(1)$ over the choice of these m points, the distribution of

$$\big(\boldsymbol{f}(\boldsymbol{u}_1), \ldots, \boldsymbol{f}(\boldsymbol{u}_m)\big) \qquad \text{is } o(1)\text{-close to that of} \qquad \big(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m\big)$$

for $\boldsymbol{f} \sim \mathcal{F}_d$ and where each $\boldsymbol{b}_i$ is an independent unbiased random bit. In other words, we aim to show that if the evaluation points $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ are randomly chosen (but subsequently known to the algorithm), then $\boldsymbol{f}(\boldsymbol{u}_1), \ldots, \boldsymbol{f}(\boldsymbol{u}_m)$ is $o(1)$-indistinguishable from random.

We establish this last statement by proving something even stronger. Namely, we first observe that the $\mathbb{R}^m$-valued random variable $(\boldsymbol{p}(\boldsymbol{u}_m), \ldots, \boldsymbol{p}(\boldsymbol{u}_m))$ is an m-dimensional normal random variable for any fixed outcome of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$. Subsequently, we show that this random variable $(\boldsymbol{p}(\boldsymbol{u}_m), \ldots, \boldsymbol{p}(\boldsymbol{u}_m))$ is $o(1)$-close to the standard m-dimensional normal random variable $N(0, I_m)$ where I_m is the identity matrix in m dimensions. This exploits a recent bound on total variation distance between multivariate normal distributions [DMR20] in terms of their covariance matrices,

and involves bounding the trace of the Gram matrix generated by random points on the hypercube; details are deferred to the main body of the paper.

1.3 Related Work

As mentioned earlier, "truncated statistics" has been a topic of central interest to statisticians for more than a century and recently in theoretical computer science as well. Starting with the work of Daskalakis et al. [DGTZ18], several works have looked at the problem of learning an unknown high-dimensional distribution in settings where the algorithm only gets samples from a truncated set [FKT20, KTZ19, BDNP21]. We note here that in the recent past, there have also been several works on truncation in the area of statistics related to supervised learning scenarios [DSYZ21, DGTZ19, DRZ20], but the models and techniques in those works are somewhat distant from the topic of the current paper. Finally, in retrospect, some earlier works on "learning from positive samples" [CDS20, DDS15, DGL05] also have a similar flavor. In particular, the main result of [CDS20] is a poly(n) time algorithm which, given access to samples from a Gaussian truncated by an unknown degree-2 PTF, approximately recovers the truncation set; and one of the main results of [DDS15] is an analogous poly(n)-time algorithm but for degree-1 PTF (i.e. LTF) truncations of the uniform distribution over $\{-1,1\}^n$. Note that while the settings of [CDS20, DDS15] are somewhat related to the current paper, the goals and results of those works are quite different; in particular, the focus is on learning (as opposed to testing / determining whether truncation has taken place), and the sample complexities of the algorithms in [CDS20, DDS15], albeit polynomial in n, are polynomials of high degree.

In terms of the specific problem we study, the work most closely related to the current paper is that of [DNS23]. In particular, as noted earlier, in [DNS23], the algorithm gets access to samples from either (i) $N(0,1)^n$ or (ii) $N(0,1)^n$ conditioned on a convex set. Besides the obvious difference in the truncation sets which are considered—convex sets in [DNS23] vis-a-vis PTFs in the current paper—the choice of the background distribution in [DNS23] is far more restrictive. Namely, [DNS23] requires the background distribution to be the normal distribution $N(0,1)^n$, whereas the results in current paper hold for the broad family of hypercontractive product distributions (which includes many other distributions as well as the normal distribution). The difference in the problem settings is also reflected in the techniques employed in these two papers. In particular, the algorithm and analysis of [DNS23] heavily rely on tools from convex geometry including Gaussian isoperimetry [Bor85], the Brascamp-Lieb inequality [BL76a, BL76b], and recent structural results for convex bodies over Gaussian space [DNS21, DNS22]. In our setting, truncation sets defined by PTFs even of degree two need not be convex, so we must take a very different approach. The algorithm in the current paper uses techniques originating from the study of PTFs in concrete complexity theory, in particular on the hypercontractivity of low-degree polynomials, anti-concentration, and the "level-k" inequalities [O'D14]. So to summarize the current work vis-a-vis [DNS23], the current work studies a different class of truncations under a significantly less restrictive assumption on the background distribution, and our main algorithm, as well as its analysis, are completely different from those of [DNS23].

Our lower bound argument extends and strengthens a $\tilde{\Omega}(n^{1/2})$ lower bound, given in [DNS23], for distinguishing the standard normal distribution $N(0,1)^n$ from $N(0,1)^n|_{f^{-1}(1)}$ where f is an unknown origin-centered LTF (i.e. a degree-1 PTF); both arguments use a variation distance lower bound between a standard multivariate normal distribution and a multivariate normal distribution with a suitable slightly perturbed covariance matrix. Our lower bound argument in the current paper combines tools from the LTF lower bound mentioned above with ingredients (in particular, the use of a "shadow sample"; see Section 4 of the full version) from a different lower bound from [DNS23] for symmetric slabs; extends the [DNS23] analysis from degree-1 to degree-d for any constant d; and gives a tighter analysis than [DNS23] which does not lose any log factors.

We end this discussion of related work with the following overarching high-level question, which we hope will be investigated in future work: Suppose $\mathcal{P}$ is a background distribution and $\mathcal{F}$ is a class of Boolean functions. *Under what conditions can we distinguish between $\mathcal{D} = \mathcal{P}$ versus $\mathcal{D} = \mathcal{P}|_f$ (for some $f \in \mathcal{F}$) with sample complexity asymptotically smaller than the sample complexity of learning $\mathcal{F}$?* We view our results on distinguishing truncation by PTFs as a step towards answering this question.

2 PRELIMINARIES

We write $\mathbb{N} := \{0, 1, \ldots\}$ and $\mathbf{1}\{\cdot\}$ for the 0/1 indicator function. We will write

$$\binom{[n]}{d} := \{S \subseteq [n] : |S| = d\}.$$

Let $(\mathbb{R}, \mu)$ be a probability space. For $n \in \mathbb{N}$, we write $L^2(\mathbb{R}^n, \mu^{\otimes n})$ for the (real) inner-product space of functions $f : \mathbb{R}^n \to \mathbb{R}$ with the inner product

$$\langle f, g \rangle := \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mu^{\otimes n}} \big[f(\boldsymbol{x}) \cdot g(\boldsymbol{x}) \big].$$

Here $\mu^{\otimes n}$ denotes the product probability distribution on $\mathbb{R}^n$. For $q > 0$ we write

$$\|f\|_q := \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mu^{\otimes n}} \big[|f(\boldsymbol{x})|^q \big]^{1/q}.$$

In particular, for $f : \mathbb{R}^n \to \{0,1\}$, we write $\mathrm{Vol}(f) := \|f\|_1 = \mathbf{E}[f(\boldsymbol{x})]$ where $\boldsymbol{x} \sim \mu^{\otimes n}$.

We say that a function $f : \mathbb{R}^n \to \{0,1\}$ is a *degree-d polynomial threshold function (PTF)* if there exists a polynomial $p : \mathbb{R}^n \to \mathbb{R}$ of degree at most d such that

$$f(x) = \mathbf{1}\{p(x) \geq 0\}.$$

The primary class of distributions we will consider throughout is that of truncations of an i.i.d. product distribution $\mu^{\otimes n}$ by a degree-d PTF of at least some minimum volume; more precisely, we will consider the following class of truncations:

$$C_{\mathrm{PTF}}(d, \alpha) := \left\{ \mu^{\otimes n} |_{f^{-1}(1)} : f \text{ is a degree-}d \text{ PTF with } \mathrm{Vol}(f) \geq \alpha \right\} \tag{1}$$

where $\alpha = \alpha(n)$ may depend on n (in fact we will be able to take α as small as $2^{-\Theta(\sqrt{n})}$). Throughout the paper we will assume that d

(the degree of the PTFs we consider) is a fixed constant independent of the ambient dimension n.

2.1 Harmonic Analysis over Product Spaces

Our notation and terminology in this section closely follow those of O'Donnell [O'D14]; in particular, we refer the reader to Chapter 8 of [O'D14] for further background.

Definition 5. A *Fourier basis* for $L^2(\mathbb{R}, \mu)$ is an orthonormal basis $\mathcal{B} = \{\chi_0, \chi_1, \ldots\}$ with $\chi_0 \equiv 1$.

It is well known that if $L^2(\mathbb{R}, \mu)$ is separable,[2] then it has a Fourier basis (see for e.g. Section I.4 of [Con19]). Note that we can obtain a Fourier basis for $L^2(\mathbb{R}^n, \mu^{\otimes n})$ by taking all possible n-fold products of elements of $\mathcal{B}$; more formally, for a multi-index $\alpha \in \mathbb{N}^n$, we define

$$\chi_\alpha(x) := \prod_{i=1}^{n} \chi_{\alpha_i}(x_i).$$

Then the collection $\mathcal{B}_n := \{\chi_\alpha : \alpha_i \in \mathbb{N}^n\}$ forms a Fourier basis for $L^2(\mathbb{R}^n, \mu^{\otimes n})$; this lets us write $f \in L^2(\mathbb{R}^n, \mu^{\otimes n})$ as $f = \sum_{\alpha \in \mathbb{N}^n} \widehat{f}(\alpha)\chi_\alpha$ where

$$\widehat{f}(\alpha) := \langle f, \chi_\alpha \rangle$$

is the *Fourier coefficient of f on α*.

We can assume without loss of generality that the basis elements of $L^2(\mathbb{R}, \mu)$, namely $\{\chi_0, \chi_1, \ldots\}$, are polynomials with $\deg(\chi_i) = i$. This is because a polynomial basis can be obtained for $L^2(\mathbb{R}, \mu)$ by running the Gram-Schmidt process. By extending this basis to $L^2(\mathbb{R}^n, \mu^{\otimes n})$ by taking products, it follows that we may assume without loss of generality that for a multi-index $\alpha \in \mathbb{N}^n$, we have $\deg(\chi_\alpha) = |\alpha|$ where

$$|\alpha| := \sum_{i=1}^{n} \alpha_i.$$

We will also write $\#\alpha := \big|\mathrm{supp}(\alpha)\big|$ where $\mathrm{supp}(\alpha) := \{i : \alpha_i \neq 0\}$.

Remark 6. While the Fourier coefficients $\{\widehat{f}(\alpha)\}$ depend on the choice of basis $\{\chi_\alpha\}$, we will always work with some fixed (albeit arbitrary) polynomial basis, and hence there should be no ambiguity in referring to the coefficients as though they were unique. We assume that the orthogonal basis $\{\chi_\alpha\}$ is "known" to the algorithm; this is certainly a reasonable assumption for natural examples of hypercontractive distributions (e.g. distributions with finite support, the uniform distribution on intervals, the Gaussian distribution, etc.), and is in line with our overall problem formulation of detecting whether a known background distribution has been subjected to truncation.

As a consequence of orthonormality, we get that for any $f \in L^2(\mathbb{R}^n, \mu^{\otimes n})$, we have

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mu^{\otimes n}}\big[f(\boldsymbol{x})\big] = \widehat{f}(0^n) \qquad \text{and} \qquad \|f\|_2^2 = \sum_{\alpha \in \mathbb{N}^n} \widehat{f}(\alpha)^2,$$

with the latter called *Parseval's formula*. We also have *Plancharel's formula*, which says that

$$\langle f, g \rangle = \sum_{\alpha \in \mathbb{N}^n} \widehat{f}(\alpha)\widehat{g}(\alpha).$$

[2] Recall that a metric space is *separable* if it contains a countable dense subset.

Product Distribution $\mathcal{D}$	C_q
Gaussian Distribution $N(0,1)^n$ [AS17]	$\sqrt{q-1}$
Uniform Measure on $\{-1,+1\}^n$ [O'D14]	$\sqrt{q-1}$
Finite Product Domains $(\Omega^n, \mu^{\otimes n})$ [Wol07]	$\sqrt{\frac{q}{2\min(\mu)}}$

Table 1: Examples of hypercontractive distributions, along with their accompanying hypercontractivity constants. Here $\min(\mu)$ denotes the minimal non-zero probability of any element in the support of the (finitely supported) distribution μ.

Finally, we write

$$\mathbf{W}^{=k}[f] := \sum_{|\alpha|=k} \widehat{f}(\alpha)^2 \qquad \text{and} \qquad \mathbf{W}^{\leq k}[f] := \sum_{|\alpha|\leq k} \widehat{f}(\alpha)^2$$

for the *Fourier weight of f at level k* and the *Fourier weight of f up to level k* respectively.

2.2 Hypercontractive Distributions

The primary analytic tools we will require in both our upper and lower bounds are consequences of *hypercontractive* estimates for functions in $L^2(\mathbb{R}^n, \mu^{\otimes n})$; we refer the reader to Chapters 9 and 10 of [O'D14] for further background on hypercontractivity and its applications.

Definition 7. We say that $(\mathbb{R}, \mu)$ is *hypercontractive* if for every $q \geq 2$, there is a fixed constant $C_q(\mu)$ such that for every $n \geq 1$ and every multivariate degree-d polynomial $p : \mathbb{R}^n \to \mathbb{R}$ we have

$$\|p\|_q \leq C_q(\mu)^d \cdot \|p\|_2 \tag{2}$$

where $C_q(\mu)$ is independent of n and satisfies

$$C_q(\mu) \leq K\sqrt{q} \tag{3}$$

for an absolute constant K. When the product distribution $\mu^{\otimes n}$ is clear from context, we will sometimes simply write $C_q := C_q(\mu)$ instead.

It is clear from the monotonicity of norms that $C_q \geq 1$; see Table 1 for examples of hypercontractive distributions with accompanying hypercontractivity constants $C_q(\mu)$.

Remark 8. We note that Definition 7 is not the standard definition of a hypercontractive product distribution (cf. Chapters 9 and 10 of [O'D14]), but is in fact an easy consequence of hypercontractivity that is sometimes referred to as the "Bonami lemma." The guarantees of Equations (2) and (3) are all we require for our purposes, and so we choose to work with this definition instead.

Remark 9. While Equation (3) may seem extraneous, we note that the "level-k inequalities" (Proposition 11) crucially rely on this bound on the hypercontractivity constant C_q.

We turn to record several useful consequences of hypercontractivity which will be crucial to the analysis of our estimator as well

as to our lower bound. We defer the proofs of Propositions 10 and 11 to the full version of this paper.

The following *anti-concentration* inequality is a straightforward consequence of hypercontractivity. A similar result for arbitrary product distributions with finite support was obtained by Austrin–Håstad [AH09], and a similar result for functions over $\{-1,+1\}^n$ with the uniform measure was obtained by Dinur et al. [DFKO06]. The proof of the following proposition closely follows that of Proposition 9.7 in [O'D14]:

Proposition 10 (Anti-concentration of low-degree polynomials)**.** Suppose $(\mathbb{R}^n, \mu^{\otimes n})$ is a hypercontractive probability space. Then for any degree-d polynomial $p : \mathbb{R}^n \to \mathbb{R}$ with $\mathbf{E}[p] = 0$ and $\mathbf{Var}[p] = 1$, we have

$$\Pr_{\boldsymbol{x}\sim\mu^{\otimes n}}\left[|p(\boldsymbol{x})| \geq \frac{1}{2}\right] \geq 0.5625 \cdot c^d$$

for a constant $c := c(\mu)$ independent of n.

The following proposition bounds Fourier weight up to level k (i.e. $\mathbf{W}^{\leq k}[f]$) in terms of the bias (i.e. the degree-0 Fourier coefficient) of the function. We note that an analogous result for functions over $\{-1,+1\}^n$ with the uniform measure is sometimes known as "Chang's Lemma" or "Talagrand's Lemma" [Cha02, Tal96]; see also Section 9.5 of [O'D14].

Proposition 11 (Level-k inequalities)**.** Suppose $(\mathbb{R}, \mu)$ is hypercontractive and $f : \mathbb{R}^n \to \{0,1\}$ is a Boolean function. Then for all $1 \leq k \leq 2\log\left(\frac{1}{\mathrm{Vol}(f)}\right)$ we have

$$\mathbf{W}^{\leq k}[f] \leq K^k \cdot \mathrm{Vol}(f)^2 \cdot \left(\log\left(\frac{1}{\mathrm{Vol}(f)}\right)\right)^k$$

where K is a constant independent of n.

Remark 12. We note that the Proposition 11 also holds for bounded functions $f : \mathbb{R}^n \to [-1,1]$ with $\mathrm{Vol}(f) := \mathbf{E}[|f|]$, although we will not require this.

3 AN $O(n^{d/2})$-SAMPLE ALGORITHM FOR DEGREE-d PTFS

In this section, we present a $O(n^{d/2})$-sample algorithm for distinguishing a hypercontractive product distribution $\mu^{\otimes n}$ from $\mu^{\otimes n}$ truncated by the satisfying assignments of a degree-d PTF. More precisely, we prove the following in Section 3.2:

Theorem 13. Let $\varepsilon > 0$ and let $(\mathbb{R}, \mu)$ be hypercontractive. There is an algorithm, PTF-Distinguisher (Algorithm 1), with the following performance guarantee: Given access to independent samples from any unknown distribution $\mathcal{D} \in \{\mu^{\otimes n}\} \cup C_{\mathrm{PTF}}(d, 2^{-\sqrt{n}})$, the algorithm uses T samples where

$$T := \Theta_d\left(\frac{n^{d/2}}{\min\left\{1, \varepsilon/(1-\varepsilon), c^{\Theta(d)}/(1-\varepsilon)\right\}^2}\right)$$

with $c := c(\mu)$ as in Proposition 10, runs in $O_d(T \cdot n^d)$ time, and

(1) If $\mathcal{D} = \mu^{\otimes n}$, then with probability at least 9/10 the algorithm outputs "un-truncated;"
(2) If $\mathrm{d_{TV}}(\mu^{\otimes n}, \mathcal{D}) \geq \varepsilon$ (equivalently, $\mathcal{D} = \mu^{\otimes n}|_f$ for some degree-d PTF f with $2^{-\sqrt{n}} \leq \mathrm{Vol}(f) \leq 1-\varepsilon$), then with probability at least 9/10 the algorithm outputs "truncated."

Input: $\mathcal{D} \in \{\mu^{\otimes n}\} \cup C_{\mathrm{PTF}}(d, 2^{-\sqrt{n}})$, $\varepsilon > 0$

Output: "Un-truncated" or "truncated"

PTF-Distinguisher($\mathcal{D}$):

(1) Draw $2T$ independent sample points $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)}, \boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(T)} \sim \mathcal{D}$, where
$$T := \Theta_d\left(\frac{n^{d/2}}{\min\left\{1, \varepsilon/(1-\varepsilon), c^{\Theta(d)}/(1-\varepsilon)\right\}^2}\right).$$
(2) Compute the statistic $\mathbf{M}$ where
$$\mathbf{M} := \frac{1}{T^2}\left\langle \sum_{i=1}^{T} \widetilde{\boldsymbol{x}}^{(i)}, \sum_{j=1}^{T} \widetilde{\boldsymbol{y}}^{(j)} \right\rangle$$
with
$$\widetilde{\boldsymbol{x}}^{(i)} := \left(\chi_\alpha(\boldsymbol{x}^{(i)})\right)_{1\leq|\alpha|\leq d}$$
and $\widetilde{\boldsymbol{y}}^{(j)}$ defined similarly.
(3) Output "truncated" if $\mathbf{M} \geq \min\left\{1, \left(\frac{\varepsilon}{1-\varepsilon}\right), \frac{c^{\Theta(d)}}{(1-\varepsilon)}\right\}$ and "un-truncated" otherwise.

Algorithm 1: Distinguisher for degree-d PTFs. Throughout the algorithm the constant $c := c(\mu)$ is as in the proof of Lemma 15.

Before proceeding to the proof of Theorem 13, we give a brief high-level description of Algorithm 1. The algorithm draws $2T$ independent samples $\{\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}\}_{i\in T}$ where T is as above, and then performs a *feature expansion* to obtain the $2T$ vectors $\{\widetilde{\boldsymbol{x}}^{(i)}, \widetilde{\boldsymbol{y}}^{(i)}\}_{i\in T}$ where

$$\widetilde{\boldsymbol{x}}^{(i)} := \left(\chi_\alpha(\boldsymbol{x}^{(i)})\right)_{1\leq|\alpha|\leq d}$$

and $\widetilde{\boldsymbol{y}}^{(i)}$ is defined similarly. The statistic $\mathbf{M}$ employed by the algorithm to distinguish between the un-truncated and truncated is then given by

(1) First computing the average of the kernelized $\widetilde{\boldsymbol{x}}^{(i)}$ vectors and the kernelized $\widetilde{\boldsymbol{y}}^{(i)}$ vectors; and then
(2) Taking the inner product between the two averaged kernel vectors.

An easy calculation, given below, relates the statistic $\mathbf{M}$ to the low-degree (but not degree-0) Fourier weight of the truncation function (note that if no truncation is applied then the truncation function is identically 1). The analysis then proceeds by using anti-concentration of low-degree polynomials to show that the means of the estimators differ by $\Omega_\varepsilon(1)$ between the two settings. We bound the variance of the estimator in both the un-truncated and truncated setting (using the level-k inequalities at a crucial point in the analysis of the truncated setting), and given a separation between

the means and a bound on the variances, it is straightforward to distinguish between the two settings using Chebyshev's inequality.

Remark 14. We note that the trick of drawing a "bipartite" set of samples, i.e. drawing $2T$ samples $\{\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}\}_{i\in T}$, was recently employed in the algorithm of Diakonikolas, Kane, and Pensia [DKP22] for the problem of Gaussian mean testing. For our problem we could have alternately used the closely related estimator $\mathbf{M}'$ given by

$$\mathbf{M}' := \binom{T}{2}^{-1} \sum_{i\neq j} \left\langle \widetilde{\boldsymbol{x}}^{(i)}, \widetilde{\boldsymbol{x}}^{(j)} \right\rangle$$

to distinguish between the un-truncated and truncated distributions via a similar but slightly more cumbersome analysis. We note that the main technical tool used in the analysis of Diakonikolas, Kane, and Pensia [DKP22] is the Carbery-Wright inequality for degree-2 polynomials in Gaussian random variables, whereas our argument uses the above-mentioned kernelization approach and other consequences of hypercontractivity, namely the level-k inequalities, beyond just anti-concentration.

3.1 Useful Preliminaries

The following lemma will be crucial in obtaining a lower-bound for the expectation of our test statistic $\mathbf{E}[\mathbf{M}]$ in the truncated setting; we note that an analogous statement in the setting of the Boolean hypercube was obtained by Gotsman and Linial [GL94].

Lemma 15. Suppose $(\mathbb{R}, \mu)$ is hypercontractive. If $f : \mathbb{R}^n \to \{0, 1\}$ is a degree-d PTF, then

$$\sum_{1\leq|\alpha|\leq d} \widehat{f}(\alpha)^2 \geq \Omega\left(\min\left\{\mathrm{Vol}(f), 1-\mathrm{Vol}(f), c^{\Theta(d)}\right\}\right)^2$$

for an absolute constant $c := c(\mu) \in (0, 1]$.

Proof. We may assume that

$$f(x) = \mathbf{1}\{p(x) \geq \theta\}$$

where $p : \mathbb{R}^n \to \mathbb{R}$ is a degree-d polynomial with $\mathbf{E}[p(\boldsymbol{x})] = \widehat{p}(0^n) = 0$ and $\|p\|_2^2 = \mathbf{Var}[p] = \sum_\alpha \widehat{p}(\alpha)^2 = 1$. By Cauchy–Schwarz and Plancherel, we get

$$\begin{aligned}\sum_{1\leq|\alpha|\leq d} \widehat{f}(\alpha)^2 &= \left(\sum_{1\leq|\alpha|\leq d} \widehat{f}(\alpha)^2\right)\left(\sum_{1\leq|\alpha|\leq d} \widehat{p}(\alpha)^2\right)\\ &\geq \left(\sum_{1\leq|\alpha|\leq d} \widehat{f}(\alpha)\cdot\widehat{p}(\alpha)\right)^2\\ &= \left(\mathbf{E}\left[f(\boldsymbol{x})\cdot p(\boldsymbol{x})\right]\right)^2. \end{aligned} \tag{4}$$

where we made use of the the fact that p is a degree-d polynomial with $\widehat{p}(0^n) = 0$ and $\mathbf{Var}\left[p(\boldsymbol{x})\right] = 1$. Note that by Proposition 10, we have that either

$$\Pr_{\boldsymbol{x}\sim\mu^{\otimes n}}\left[p(\boldsymbol{x}) \geq \frac{1}{2}\right] \geq \Omega\left(c^d\right) \text{ or } \Pr_{\boldsymbol{x}\sim\mu^{\otimes n}}\left[p(\boldsymbol{x}) \leq -\frac{1}{2}\right] \geq \Omega\left(c^d\right). \tag{5}$$

Suppose that it is the former. (As we will briefly explain later, the argument is symmetric in the latter case.) We further break the analysis into cases depending on the magnitude of θ.

Case 1: $\theta \geq \frac{1}{2}$. In this case, we have by Equation (4) that

$$\begin{aligned}\sum_{1\leq|\alpha|\leq d} \widehat{f}(\alpha)^2 &\geq \left(\mathbf{E}\left[f(\boldsymbol{x})\cdot p(\boldsymbol{x})\right]\right)^2\\ &= \left(\mathbf{E}\left[\mathbf{1}_{p(\boldsymbol{x})\geq\theta}\cdot p(\boldsymbol{x})\right]\right)^2\\ &\geq (\mathrm{Vol}(f)\cdot\theta)^2\\ &\geq \Omega\left(\mathrm{Vol}(f)^2\right),\end{aligned}$$

and so the result follows.

Case 2: $0 \leq \theta < \frac{1}{2}$. In this case, we have by Proposition 10 that

$$\mathrm{Vol}(f) = \Pr_{\boldsymbol{x}\sim\mu^{\otimes n}}\left[p(\boldsymbol{x}) \geq \theta\right] \geq \Pr_{\boldsymbol{x}\sim\mu^{\otimes n}}\left[p(\boldsymbol{x}) \geq \frac{1}{2}\right] \geq \Omega\left(c^d\right).$$

Once again by Equation (4), we have

$$\begin{aligned}\sum_{1\leq|\alpha|\leq d} \widehat{f}(\alpha)^2 &\geq \left(\mathbf{E}\left[f(\boldsymbol{x})\cdot p(\boldsymbol{x})\right]\right)^2\\ &\geq \left(\frac{1}{2}\Pr\left[p(\boldsymbol{x}) \geq \frac{1}{2}\right]\right)^2\\ &\geq \Omega\left(c^{\Theta(d)}\right),\end{aligned}$$

where the second inequality follows from $f\cdot p$ being always non-negative and at least $\frac{1}{2}$ with probability $\mathbf{Pr}[p(\boldsymbol{x}) \geq \frac{1}{2}]$.

Case 3: $\theta < 0$. Consider the degree-d PTF $f^\dagger := 1 - f$ given by

$$f^\dagger(x) = \mathbf{1}\{p(x) < \theta\}.$$

It is easy to check that $|\widehat{f^\dagger}(\alpha)| = |\widehat{f}(\alpha)|$ for all $S \neq \emptyset$ and that $\mathrm{Vol}(f^\dagger) = 1 - \mathrm{Vol}(f)$. Repeating the above analysis then gives that

$$\begin{aligned}\sum_{1\leq|\alpha|\leq d} \widehat{f}(\alpha)^2 &= \sum_{1\leq|\alpha|\leq d} \widehat{f^\dagger}(\alpha)^2\\ &\geq \Omega\left(\mathrm{Vol}(f^\dagger)^2\cdot c^{\Theta(d)}\right)\\ &= \Omega\left(\left(1-\mathrm{Vol}(f)\right)^2\cdot c^{\Theta(d)}\right).\end{aligned}$$

Putting Cases 1 through 3 together, we get that

$$\sum_{1\leq|\alpha|\leq d} \widehat{f}(\alpha)^2 \geq \Omega\left(\min\left\{\mathrm{Vol}(f), 1-\mathrm{Vol}(f), c^{\Theta(d)}\right\}\right)^2, \tag{6}$$

completing the proof. Recall, however, that we assumed that

$$\Pr_{\boldsymbol{x}\sim\mu^{\otimes n}}\left[p(\boldsymbol{x}) \geq \frac{1}{2}\right] \geq \Omega\left(c^d\right)$$

in Equation (5). Suppose that we instead have

$$\Pr_{\boldsymbol{x}\sim\mu^{\otimes n}}\left[p(\boldsymbol{x}) \leq -\frac{1}{2}\right] \geq \Omega\left(c^d\right).$$

Then note that the same trick used in Case 3 by considering $f^\dagger$ instead of f and repeating the three cases completes the proof. □

We will also require the following two lemmas which are closely linked to the *linearization problem for orthogonal polynomials* (see Section 6.8 of [AAR99]). The first lemma bounds the magnitude of the Fourier coefficients of the *product* of basis functions; while the

estimate below relies on hypercontractivity, we note that exact expressions for the Fourier coefficients are known for various classes of orthogonal polynomials including the Chebyshev, Hermite, and Laguerre polynomials.

Lemma 16. Suppose $(\mathbb{R}, \mu)$ is hypercontractive. Let $\alpha, \beta, \gamma \in \mathbb{N}^n$. Then

$$\left\langle \chi_\alpha, \chi_\beta \cdot \chi_\gamma \right\rangle \le C_4(\mu)^{|\beta|+|\gamma|}.$$

Proof. Using Cauchy–Schwarz, we have that

$$\begin{aligned}\left\langle \chi_\alpha, \chi_\beta \cdot \chi_\gamma \right\rangle &= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mu^{\otimes n}}\left[\chi_\alpha(\boldsymbol{x}) \cdot \left(\chi_\beta(\boldsymbol{x}) \cdot \chi_\gamma(\boldsymbol{x})\right)\right] \\ &\le \sqrt{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mu^{\otimes n}}\left[\chi_\alpha(\boldsymbol{x})^2\right] \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mu^{\otimes n}}\left[\chi_\beta(\boldsymbol{x})^2 \cdot \chi_\gamma(\boldsymbol{x})^2\right]} \\ &= \sqrt{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mu^{\otimes n}}\left[\chi_\beta(\boldsymbol{x})^2 \cdot \chi_\gamma(\boldsymbol{x})^2\right]}\end{aligned}$$

due to the orthonormality of χ_α. Using Cauchy–Schwarz once again, we have that

$$\begin{aligned}\left\langle \chi_\alpha, \chi_\beta \cdot \chi_\gamma \right\rangle &\le \left(\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mu^{\otimes n}}\left[\chi_\beta(\boldsymbol{x})^4\right] \cdot \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mu^{\otimes n}}\left[\chi_\gamma(\boldsymbol{x})^4\right]\right)^{1/4} \\ &= \|\chi_\beta\|_4 \cdot \|\chi_\gamma\|_4 \\ &\le C_4(\mu)^{|\beta|+|\gamma|}\end{aligned}$$

where the final inequality uses hypercontractivity and the fact that χ_β (χ_γ respectively) is a polynomial of two-norm 1 and degree at most $|\beta|$ (at most $|\gamma|$ respectively). □

Finally, we also require the following combinatorial lemma:

Lemma 17. Given $d \in \mathbb{N}$ and a fixed multi-index $\alpha \in \mathbb{N}^n$ with $|\alpha| := k \le 2d$, we have

$$\left|\left\{(\beta, \gamma) : |\beta|, |\gamma| \le d, \left\langle \chi_\alpha, \chi_\beta \chi_\gamma \right\rangle \ne 0\right\}\right| \le O_d\left(n^{d-\lceil k/2\rceil}\right)$$

where $\beta, \gamma \in \mathbb{N}^n$.

Proof. We will upper bound the number of pairs (β, γ) such that $|\beta|, |\gamma| \le d$ and $\left\langle \chi_\alpha, \chi_\beta \chi_\gamma \right\rangle \ne 0$; so fix such a pair (β, γ).

We first note that for any $i \notin \mathrm{supp}(\alpha)$, we must have $\beta_i = \gamma_i$; for otherwise $\left\langle \chi_\alpha, \chi_\beta \chi_\gamma \right\rangle = 0$ due to orthonormality of χ_{β_i} and χ_{γ_i}. Also, for each $i \in \mathrm{supp}(\alpha)$, we must have $\beta_i + \gamma_i \ge \alpha_i$, since if $\beta_i + \gamma_i < \alpha_i$, then

$$\left\langle \chi_{\beta_i} \cdot \chi_{\gamma_i}, \chi_{\alpha_i} \right\rangle = 0$$

by orthonormality and the fact that $\chi_{\beta_i} \cdot \chi_{\gamma_i}$ is a linear combination of basis functions $\{\chi_j\}_{j<\alpha_i}$.

Summing over all $i \in \mathrm{supp}(\alpha)$, we get that

$$\sum_{i\in\mathrm{supp}(\alpha)} \beta_i + \gamma_i \ge \sum_{i\in\mathrm{supp}(\alpha)} \alpha_i = |\alpha|,$$

and so it follows that either

$$\sum_{i\in\mathrm{supp}(\alpha)} \beta_i \ge \lceil k/2\rceil \qquad \text{or} \qquad \sum_{i\in\mathrm{supp}(\alpha)} \gamma_i \ge \lceil k/2\rceil;$$

without loss of generality we suppose it is the former. It follows that the total number of ways of choosing such a pair (β, γ) is bounded by

$$O_d(1) \cdot O_d(1) \cdot \sum_{j=0}^{d-\lceil k/2\rceil} (n - |\mathrm{supp}(\alpha)|)^j$$

which in turn is

$$O_d(1) \cdot \sum_{j=0}^{d-\lceil k/2\rceil} n^j \le O_d(n^{d-\lceil k/2\rceil}),$$

where

- The first two $O_d(1)$ factors bound the number of possible outcomes of $(\beta_i)_{i\in\mathrm{supp}(\alpha)}$ (recall that each $\beta_i \in [0, d]$ and $|\mathrm{supp}(\alpha)| \le 2d$) and $(\gamma)_{i\in\mathrm{supp}(\alpha)}$ respectively); and
- The third term on the LHS upper bounds the number of choices for $(\beta_i)_{i\notin\mathrm{supp}(\alpha)}$. Recall that for $i \notin \mathrm{supp}(\alpha)$ we must have $\beta_i = \gamma_i$, and so

$$(\beta_i)_{i\notin\mathrm{supp}(\alpha)} = (\gamma_i)_{i\notin\mathrm{supp}(\alpha)}$$

so we need only bound the number of choices of $(\beta_i)_{i\notin\mathrm{supp}(\alpha)}$; moreover, as

$$\sum_i \beta_i \le d \quad \text{and} \quad \sum_{i\in\mathrm{supp}(\alpha)} \beta_i \ge \lceil k/2\rceil,$$

we must have

$$\sum_{i\notin\mathrm{supp}(\alpha)} \beta_i \le d - \lceil k/2\rceil,$$

completing the proof. □

3.2 Proof of Theorem 13

Proof of Theorem 13. Throughout, we will write

$$m := \#\left\{\alpha \in \mathbb{N}^n : 1 \le |\alpha| \le d\right\}.$$

Recalling that there are at most $\binom{n+k-1}{k}$ multi-indices $\alpha \in \mathbb{N}^n$ with $|\alpha| = k$, we have that

$$m \le O_d(n^d).$$

We compute the mean and variance of the estimator $\mathbf{M}$ in each case (i.e. when $\mathcal{D} = \mu^{\otimes n}$ and when $\mathcal{D} \in C_{\mathrm{PTF}}(d, 2^{-\sqrt{n}})$) separately.

Case 1: $\mathcal{D} = \mu^{\otimes n}$. For brevity (and to distinguish the calculations in this case from the following one), we denote the un-truncated distribution by

$$\mathcal{D}_u := \mu^{\otimes n}.$$

When $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)}, \boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(T)} \sim \mathcal{D}_u$, we have by bi-linearity of the inner product that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathcal{D}_u}[\mathbf{M}] &= \frac{1}{T^2}\sum_{i,j=1}^{T} \mathop{\mathbf{E}}_{\mathcal{D}_u}\left[\left\langle \widetilde{\boldsymbol{x}}^{(i)}, \widetilde{\boldsymbol{y}}^{(j)} \right\rangle\right] \\ &= \sum_{1\le|\alpha|\le d} \mathop{\mathbf{E}}_{\mathcal{D}_u}\left[\chi_\alpha(\boldsymbol{x}) \cdot \chi_\alpha(\boldsymbol{y})\right] \\ &= \sum_{1\le|\alpha|\le d} \mathop{\mathbf{E}}_{\mathcal{D}_u}\left[\chi_\alpha(\boldsymbol{x})\right] \cdot \mathop{\mathbf{E}}_{\mathcal{D}_u}\left[\chi_\alpha(\boldsymbol{y})\right] \\ &= 0 \end{aligned} \tag{7}$$

as $\boldsymbol{x}, \boldsymbol{y} \sim \mathcal{D}_u$ are independent samples and also because of the orthonormality of $\{\chi_\alpha\}$.

Turning to the variance, we have

$$\operatorname*{\mathbf{Var}}_{\mathcal{D}_u}[\mathbf{M}] = \frac{1}{T^4}\left(\sum_{i,j=1}^{T} \operatorname*{\mathbf{Var}}_{\mathcal{D}_u}\left[\left\langle \widetilde{\boldsymbol{x}}^{(i)}, \widetilde{\boldsymbol{y}}^{(j)}\right\rangle\right]\right) \tag{8}$$

$$= \frac{1}{T^2}\left(\operatorname*{\mathbf{E}}_{\mathcal{D}_u}\left[\left\langle \widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}\right\rangle^2\right] - \operatorname*{\mathbf{E}}_{\mathcal{D}_u}\left[\left\langle \widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}\right\rangle\right]^2\right) \tag{9}$$

where $\boldsymbol{x}, \boldsymbol{y} \sim \mathcal{D}$ are independent samples. In this case, we have by our previous calculation (Equation (7)) that this is in fact equal to

$$\begin{aligned}
\operatorname*{\mathbf{Var}}_{\mathcal{D}_u}[\mathbf{M}] &= \frac{1}{T^2}\left(\operatorname*{\mathbf{E}}_{\mathcal{D}_u}\left[\left\langle \widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}\right\rangle\left\langle \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}\right\rangle\right]\right) \\
&= \frac{1}{T^2}\left(\operatorname*{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{D}_u}\left[\widetilde{\boldsymbol{x}}^{\mathsf{T}} \operatorname*{\mathbf{E}}_{\boldsymbol{y}\sim\mathcal{D}_u}\left[\widetilde{\boldsymbol{y}}\cdot\widetilde{\boldsymbol{y}}^{\mathsf{T}}\right]\widetilde{\boldsymbol{x}}\right]\right) \\
&= \frac{1}{T^2}\left(\operatorname*{\mathbf{E}}_{\mathcal{D}_u}\left[\sum_{1\le|\alpha|\le d}\chi_\alpha(\boldsymbol{x})^2\right]\right) \\
&= \frac{m}{T^2} \\
&\le O_d\left(\frac{n^d}{T^2}\right)
\end{aligned} \tag{10}$$

where we used orthonormality as well as the fact that $\mathbf{E}_{\mathcal{D}_u}\left[\widetilde{\boldsymbol{y}}\cdot\widetilde{\boldsymbol{y}}^{\mathsf{T}}\right] = \mathrm{Id}_m$.

Case 2: $\mathrm{d}_{\mathrm{TV}}\left(\mu^{\otimes n}, \mathcal{D}\right) \ge \varepsilon$ *with* $\mathcal{D} \in \mathcal{C}_{\mathrm{PTF}}(d, 2^{-\sqrt{n}})$. For brevity, we denote the truncated distribution by

$$\mathcal{D}_t := \mu^{\otimes n}|_{f^{-1}(1)}.$$

In this case, $f : \mathbb{R}^n \to \{0,1\}$ is a degree-d PTF with

$$\mathrm{Vol}(f) \in \left[2^{-\sqrt{n}}, 1-\varepsilon\right].$$

We may assume that $f(x) = \mathbf{1}\{p(x) \ge \theta\}$ where $p : \mathbb{R}^n \to \mathbb{R}$ is a degree-d polynomial with $\widehat{p}(0^n) = 0$ and $\|p\|_2^2 = \mathbf{Var}[p] = \sum_\alpha \widehat{p}(\alpha)^2 = 1$.

We have the following easy relation between the Fourier coefficients of f and the means of the characters $\{\chi_\alpha\}$ under the truncated distribution $\mathcal{D}_t$:

$$\operatorname*{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{D}_t}\left[\chi_\alpha(\boldsymbol{x})\right] = \frac{1}{\mathrm{Vol}(f)}\operatorname*{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{D}_u}\left[f(\boldsymbol{x})\chi_\alpha(\boldsymbol{x})\right] = \frac{\widehat{f}(\alpha)}{\mathrm{Vol}(f)}. \tag{11}$$

We thus have

$$\begin{aligned}
\operatorname*{\mathbf{E}}_{\mathcal{D}_t}[\mathbf{M}] &= \sum_{1\le|\alpha|\le d}\operatorname*{\mathbf{E}}_{\mathcal{D}_t}\left[\chi_\alpha(\boldsymbol{x})\right]\cdot\operatorname*{\mathbf{E}}_{\mathcal{D}_t}\left[\chi_\alpha(\boldsymbol{y})\right] \\
&= \frac{1}{\mathrm{Vol}(f)^2}\sum_{1\le|\alpha|\le d}\widehat{f}(\alpha)^2 \\
&\ge \frac{1}{\mathrm{Vol}(f)^2}\cdot\Omega\left(\min\left\{\mathrm{Vol}(f), 1-\mathrm{Vol}(f), c^{\Theta(d)}\right\}\right)^2 \\
&\ge \Omega\left(\min\left\{1, \left(\frac{1-\mathrm{Vol}(f)}{\mathrm{Vol}(f)}\right), \frac{c^{\Theta(d)}}{\mathrm{Vol}(f)}\right\}\right)^2
\end{aligned} \tag{12}$$

where the second line follows from Equation (11) and the final inequality is due to Lemma 15; here $c := c(\mu)$ is as in Proposition 10.

Turning to the variance of the estimator, we have as before by Equation (9) that

$$\begin{aligned}
\operatorname*{\mathbf{Var}}_{\mathcal{D}_t}[\mathbf{M}] &= \frac{1}{T^2}\left(\operatorname*{\mathbf{E}}_{\mathcal{D}_t}\left[\left\langle \widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}\right\rangle^2\right] - \operatorname*{\mathbf{E}}_{\mathcal{D}_t}\left[\left\langle \widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}\right\rangle\right]^2\right) \\
&\le \frac{1}{T^2}\left(\operatorname*{\mathbf{E}}_{\mathcal{D}_t}\left[\left\langle \widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}\right\rangle\left\langle \widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}\right\rangle\right]\right) \\
&= \frac{1}{T^2}\left(\operatorname*{\mathbf{E}}_{\mathcal{D}_t}\left[\left(\sum_{1\le|\alpha|\le d}\chi_\alpha(\boldsymbol{x})\chi_\alpha(\boldsymbol{y})\right)^2\right]\right) \\
&= \frac{1}{T^2}\operatorname*{\mathbf{E}}_{\mathcal{D}_t}\left[\sum_{1\le|\beta|,|\gamma|\le d}\chi_\beta(\boldsymbol{x})\chi_\gamma(\boldsymbol{x})\chi_\beta(\boldsymbol{y})\chi_\gamma(\boldsymbol{y})\right] \\
&= \frac{1}{T^2}\sum_{1\le|\beta|,|\gamma|\le d}\operatorname*{\mathbf{E}}_{\mathcal{D}_t}\left[\chi_\beta\chi_\gamma\right]^2.
\end{aligned}$$

Similar to Equation (11) we can write

$$\begin{aligned}
&\sum_{1\le|\beta|,|\gamma|\le d}\operatorname*{\mathbf{E}}_{\mathcal{D}_t}\left[\chi_\beta\chi_\gamma\right]^2 \\
&= \sum_{1\le|\beta|,|\gamma|\le d}\left(\frac{1}{\mathrm{Vol}(f)}\operatorname*{\mathbf{E}}_{\mathcal{D}_u}\left[f\cdot\chi_\beta\chi_\gamma\right]\right)^2 \\
&= \frac{1}{\mathrm{Vol}(f)^2}\sum_{1\le|\beta|,|\gamma|\le d}\left(\operatorname*{\mathbf{E}}_{\mathcal{D}_u}\left[f\cdot\left(\sum_{|\alpha|\le 2d}\left\langle\chi_\beta\chi_\gamma,\chi_\alpha\right\rangle\chi_\alpha\right)\right]\right)^2 \\
&= \frac{1}{\mathrm{Vol}(f)^2}\sum_{1\le|\beta|,|\gamma|\le d}\left(\sum_{|\alpha|\le 2d}\left\langle\chi_\beta\chi_\gamma,\chi_\alpha\right\rangle\widehat{f}(\alpha)\right)^2 \\
&\le \frac{O_d(1)}{\mathrm{Vol}(f)^2}\sum_{1\le|\beta|,|\gamma|\le d}\left(\sum_{|\alpha|\le 2d}|\widehat{f}(\alpha)|\cdot\mathbf{1}\left(\left\langle\chi_\beta\chi_\gamma,\chi_\alpha\right\rangle\neq 0\right)\right)^2
\end{aligned} \tag{13}$$

where the $O_d(1)$ factor in Equation (13) comes from the RHS of Lemma 16. Now, observe that for a fixed (β,γ) pair with $1 \le |\beta|, |\gamma| \le d$ there are $O_d(1)$ many multi-indices α such that

$$\left\langle\chi_\beta\chi_\gamma,\chi_\alpha\right\rangle \neq 0$$

as we must have $\mathrm{supp}(\alpha) \subseteq \mathrm{supp}(\beta)\cup\mathrm{supp}(\gamma)$. Combining this observation with the Cauchy-Schwarz inequality

$$\left(\sum_{i=1}^{t} a_i\right)^2 \le t\cdot\left(\sum_{i=1}^{t} a_i^2\right),$$

we have

$$\begin{aligned}
&\sum_{1\le|\beta|,|\gamma|\le d} \mathop{\mathbf{E}}_{\mathcal{D}_t}\left[\chi_\beta\chi_\gamma\right]^2 \\
&\le \frac{O_d(1)}{\mathrm{Vol}(f)^2}\sum_{1\le|\beta|,|\gamma|\le d}\sum_{|\alpha|\le 2d}\widehat{f}(\alpha)^2\cdot\mathbf{1}\left(\left\langle\chi_\beta\chi_\gamma,\chi_\alpha\right\rangle\neq 0\right) \\
&= \frac{O_d(1)}{\mathrm{Vol}(f)^2}\sum_{|\alpha|\le 2d}\sum_{1\le|\beta|,|\gamma|\le d}\widehat{f}(\alpha)^2\cdot\mathbf{1}\left(\left\langle\chi_\beta\chi_\gamma,\chi_\alpha\right\rangle\neq 0\right) \\
&= \frac{O_d(1)}{\mathrm{Vol}(f)^2}\sum_{k=0}^{2d}\sum_{|\alpha|=k}\sum_{1\le|\beta|,|\gamma|\le d}\widehat{f}(\alpha)^2\cdot\mathbf{1}\left(\left\langle\chi_\beta\chi_\gamma,\chi_\alpha\right\rangle\neq 0\right) \\
&\le \frac{O_d(1)}{\mathrm{Vol}(f)^2}\sum_{k=0}^{2d}\sum_{|\alpha|=k}\widehat{f}(\alpha)^2\cdot O_d\left(n^{d-\lceil k/2\rceil}\right)
\end{aligned}$$

where the final inequality is due to Lemma 17. Finally, we have that

$$\begin{aligned}
&\sum_{1\le|\beta|,|\gamma|\le d} \mathop{\mathbf{E}}_{\mathcal{D}_t}\left[\chi_\beta\chi_\gamma\right]^2 && (14)\\
&\le \frac{O_d(1)}{\mathrm{Vol}(f)^2}\sum_{k=0}^{2d} O_d\left(n^{d-\lceil k/2\rceil}\right)\cdot\mathbf{W}^{=k}[f] \\
&\le \frac{O_d(1)}{\mathrm{Vol}(f)^2}\sum_{k=0}^{2d} O_d\left(n^{d-\lceil k/2\rceil}\right)\cdot\left(\mathrm{Vol}(f)^2\left(\log\left(\frac{1}{\mathrm{Vol}(f)}\right)\right)^k\right) && (15)\\
&\le O_d(1)\left(\sum_{k=0}^{2d} O_d\left(n^{d-\lceil k/2\rceil}\right)\cdot n^{k/2}\right) \\
&\le O_d(1)\cdot n^d, && (16)
\end{aligned}$$

where Equation (15) is by the level-k inequalities (Proposition 11) and Equation (16) uses the fact that $\mathrm{Vol}(f)\ge 2^{-\sqrt{n}}$ (which is by assumption). Putting everything together, we get that

$$\mathop{\mathbf{Var}}_{\mathcal{D}_t}[\mathbf{M}]\le O_d\left(\frac{n^d}{T^2}\right). \tag{17}$$

To summarize, when $\mathcal{D}=\mu^{\otimes n}$, we have from Equations (7) and (10) that

$$\mathop{\mathbf{E}}_{\mathcal{D}_u}[\mathbf{M}]=0 \qquad \text{and} \qquad \mathop{\mathbf{Var}}_{\mathcal{D}_u}[\mathbf{M}]=O_d\left(\frac{n^d}{T^2}\right). \tag{18}$$

On the other hand, when $\mathcal{D}\in\mathcal{C}_{\mathrm{PTF}}(d,2^{-\sqrt{n}})$ with $\mathrm{d_{TV}}(\mathcal{D},\mu^{\otimes n})\ge\varepsilon$, we have from Equations (12) and (17) that

$$\mathop{\mathbf{E}}_{\mathcal{D}_t}[\mathbf{M}]\ge\Omega\left(\min\left\{1,\left(\frac{\varepsilon}{1-\varepsilon}\right),\left(\frac{c^{\Theta(d)}}{1-\varepsilon}\right)\right\}\right)^2 \tag{19}$$

and

$$\mathop{\mathbf{Var}}_{\mathcal{D}_t}[\mathbf{M}]\le O_d\left(\frac{n^d}{T^2}\right) \tag{20}$$

For the number of samples being

$$T=\Theta_d\left(\frac{n^{d/2}}{\min\left\{1,\varepsilon/(1-\varepsilon),c^{\Theta(d)}/(1-\varepsilon)\right\}^2}\right),$$

the correctness of the distinguishing algorithm Algorithm 1 follows directly from Equation (18) and Equations (19) and (20) by a simple application of Chebyshev's inequality. □

ACKNOWLEDGEMENTS

A.D. is supported by NSF grants CCF-1910534, CCF-1926872, and CCF-2045128. H.L. is supported by NSF grants CCF-1910534, CCF-1934876, and CCF-2008305. S.N. is supported by NSF grants IIS-1838154, CCF-2106429, CCF-2211238, CCF-1763970, and CCF-2107187. Part of this work was completed while S.N. was participating in the program on "Meta Complexity" at the Simons Institute for the Theory of Computing. R.A.S. is supported by NSF grants IIS-1838154, CCF-2106429, and CCF-2211238. The authors would like to thank the anonymous STOC reviewers for helpful comments.

REFERENCES

[AAR99] George E Andrews, Richard Askey, and Ranjan Roy. *Special functions*, volume 71. Cambridge University Press, Cambridge, 1999.

[ABFR94] J. Aspnes, R. Beigel, M. Furst, and S. Rudich. The expressive power of voting polynomials. *Combinatorica*, 14(2):1–14, 1994.

[AH09] Per Austrin and Johan Håstad. Randomly supported independence and resistance. In *Proc. 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 483–492. ACM, 2009.

[AS17] Guillaume Aubrun and Stanisław J. Szarek. *Alice and Bob meet Banach*, volume 223 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2017. The interface of asymptotic geometric analysis and quantum information theory.

[BC14] N. Balakrishnan and Erhard Cramer. *The art of progressive censoring*. Springer, 2014.

[BDNP21] Arnab Bhattacharyya, Rathin Desai, Sai Ganesh Nagarajan, and Ioannis Panageas. Efficient statistics for sparse graphical models from truncated samples. In *International Conference on Artificial Intelligence and Statistics*, pages 1450–1458. PMLR, 2021.

[BHYY22] Omri Ben-Eliezer, Max Hopkins, Chutong Yang, and Hantao Yu. Active learning polynomial threshold functions. *CoRR*, abs/2201.09433, 2022.

[BL76a] H. Brascamp and E. Lieb. Best constants in Young's Inequality, its converse and its generalization to more than three functions. *Advances in Mathematics*, 20:151–172, 1976.

[BL76b] H. Brascamp and E. Lieb. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log-concave functions and with an application to the diffusion equation. *Journal of Functional Analysis*, 22:366–389, 1976.

[Bor85] C. Borell. Geometric bounds on the Ornstein-Uhlenbeck velocity process. *Probability Theory and Related Fields*, 70:1–13, 1985.

[BS92] J. Bruck and R. Smolensky. Polynomial threshold functions, AC^0 functions and spectral norms. *SIAM Journal on Computing*, 21(1):33–42, 1992.

[BS14] T. Bloom and B. Shiffman. Zeros of random polynomials on $\mathbb{C}^m$. *Math Res. Lett.*, 14:469–479, 2014.

[BV19] Pierre Baldi and Roman Vershynin. Polynomial threshold functions, hyperplane arrangements, and random tensors. *SIAM Journal on Mathematics of Data Science*, 1(4):699–729, 2019.

[CDS20] Clément L. Canonne, Anindya De, and Rocco A. Servedio. Learning from satisfying assignments under continuous distributions. In Shuchi Chawla, editor, *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 82–101. SIAM, 2020.

[Cha02] Mei-Chu Chang. A polynomial bound in Freiman's theorem. *Duke Mathematical Journal*, 113(3):399 – 419, 2002.

[Coh16] A. Clifford Cohen. *Truncated and censored samples: theory and applications*. CRC Press, 2016.

[Con19] John B Conway. *A course in functional analysis*, volume 96. Springer, 2019.

[DDS14] Anindya De, Ilias Diakonikolas, and Rocco A. Servedio. Deterministic approximate counting for juntas of degree-2 polynomial threshold functions. In *Proceedings of the 29th Annual Conference on Computational Complexity (CCC)*, pages 229–240. IEEE, 2014.

[DDS15] A. De, I. Diakonikolas, and R. A. Servedio. Learning from satisfying assignments. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, pages 478–497, 2015.

[DFKO06] Irit Dinur, Ehud Friedgut, Guy Kindler, and Ryan O'Donnell. On the Fourier tails of bounded functions over the discrete cube. In *Proc. 38th ACM Symp. on Theory of Computing*, pages 437–446, 2006.

[DGL05] F. Denis, R. Gilleron, and F. Letouzey. Learning from positive and unlabeled examples. *Theor. Comput. Sci.*, 348(1):70–83, 2005.

[DGTZ18] C. Daskalakis, T. Gouleakis, C. Tzamos, and M. Zampetakis. Efficient statistics, in high dimensions, from truncated samples. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018*, pages 639–649. IEEE Computer Society, 2018.

[DGTZ19] Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Computationally and statistically efficient truncated regression. In *Conference on Learning Theory (COLT)*, volume 99 of *Proceedings of Machine Learning Research*, pages 955–960, 2019.

[DKN10] Ilias Diakonikolas, Daniel M. Kane, and Jelani Nelson. Bounded independence fools degree-2 threshold functions. In *Proc. 51st IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 11–20, 2010.

[DKP22] Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. Gaussian mean testing made simple. *CoRR*, abs/2210.13706, 2022.

[DKPZ21] Ilias Diakonikolas, Daniel M. Kane, Thanasis Pittas, and Nikos Zarifis. The optimality of polynomial regression for agnostic learning under gaussian marginals in the SQ model. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 1552–1584. PMLR, 2021.

[DMR20] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians. arXiv:1810.08693v5, 22 May 2020, 2020.

[DNS21] Anindya De, Shivam Nadimpalli, and Rocco A. Servedio. Quantitative correlation inequalities via semigroup interpolation. In James R. Lee, editor, *12th Innovations in Theoretical Computer Science Conference, ITCS 2021, January 6-8, 2021, Virtual Conference*, volume 185 of *LIPIcs*, pages 69:1–69:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

[DNS22] Anindya De, Shivam Nadimpalli, and Rocco A. Servedio. Convex influences. In Mark Braverman, editor, *13th Innovations in Theoretical Computer Science Conference, ITCS*, volume 215 of *LIPIcs*, pages 53:1–53:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

[DNS23] Anindya De, Shivam Nadimpalli, and Rocco A. Servedio. Testing Convex Truncation. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 4050–4082. 2023.

[DOSW11] I. Diakonikolas, R. O'Donnell, R. Servedio, and Y. Wu. Hardness results for agnostically learning low-degree polynomial threshold functions. In *SODA*, pages 1590–1606, 2011.

[DRST14] Ilias Diakonikolas, Prasad Raghavendra, Rocco A Servedio, and Li-Yang Tan. Average sensitivity and noise sensitivity of polynomial threshold functions. *SIAM Journal on Computing*, 43(1):231–253, 2014.

[DRZ20] C. Daskalakis, D. Rohatgi, and E. Zampetakis. Truncated linear regression in high dimensions. *Advances in Neural Information Processing Systems*, 33:10338–10347, 2020.

[DS14] Anindya De and Rocco A. Servedio. Efficient deterministic approximate counting for low-degree polynomial threshold functions. In *Proceedings of the 46th Annual Symposium on Theory of Computing (STOC)*, pages 832–841, 2014.

[DSYZ21] C. Daskalakis, P. Stefanou, R. Yao, and E. Zampetakis. Efficient truncated linear regression with unknown noise variance. *Advances in Neural Information Processing Systems*, 34:1952–1963, 2021.

[Fis31] R.A. Fisher. Properties and applications of HH functions. In *Mathematical tables*, pages 815–852, 1931.

[FKT20] Dimitris Fotakis, Alkis Kalavasis, and Christos Tzamos. Efficient parameter estimation of truncated boolean product distributions. In *Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pages 1586–1600, 2020.

[Gal97] Francis Galton. An examination into the registered speeds of American trotting horses, with remarks on their value as hereditary data. *Proceedings of the Royal Society of London*, 62(379-387):310–315, 1897.

[GL94] C. Gotsman and N. Linial. Spectral properties of threshold functions. *Combinatorica*, 14(1):35–50, 1994.

[Ham56] J. Hammersley. The zeros of a random polynomial. In *Proc. Third Berkeley Symposium on Probability and Statistics*, volume 2, pages 89–111, 1956.

[HKM14] Prahladh Harsha, Adam Klivans, and Raghu Meka. Bounding the sensitivity of polynomial threshold functions. *Theory of Computing*, 10(1):1–26, 2014.

[Hoe94] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Collected Works of Wassily Hoeffding*, pages 171–204, 1994.

[IZ97] I. Ibragimov and O. Zeitouni. On roots of random polynomials. *Transactions of the American Mathematical Society*, 349(6):2427–2441, 1997.

[Kan11a] Daniel Kane. k-independent Gaussians fool polynomial threshold functions. In *Proceedings of the 26th Conference on Computational Complexity (CCC)*, pages 252–261, 2011.

[Kan11b] Daniel Kane. A small PRG for polynomial threshold functions of Gaussians. In *Proceedings of the 52nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 257–266, 2011.

[Kan12] D. Kane. A Structure Theorem for Poorly Anticoncentrated Gaussian Chaoses and Applications to the Study of Polynomial Threshold Functions. In *53rd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2012, New Brunswick, NJ, USA, October 20-23, 2012*, pages 91–100, 2012.

[Kan13] Daniel M Kane. The correct exponent for the gotsman-linial conjecture. In *2013 IEEE Conference on Computational Complexity*, pages 56–64. IEEE, 2013.

[Kan14a] D. M. Kane. The correct exponent for the Gotsman-Linial Conjecture. *Computational Complexity*, 23(2):151–175, 2014.

[Kan14b] Daniel Kane. A pseudorandom generator for polynomial threshold functions of Gaussians with subpolynomial seed length. In *Proceedings of the 29th Annual Conference on Computational Complexity (CCC)*, pages 217–228, 2014.

[Kan15] D. M. Kane. A Polylogarithmic PRG for Degree 2 Threshold Functions in the Gaussian Setting. In *30th Conference on Computational Complexity, CCC 2015, June 17-19, 2015, Portland, Oregon, USA*, pages 567–581, 2015.

[KKL17] Valentine Kabanets, Daniel M Kane, and Zhenjian Lu. A polynomial restriction lemma with applications. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 615–628, 2017.

[KL18] Valentine Kabanets and Zhenjian Lu. Satisfiability and Derandomization for Small Polynomial Threshold Circuits. In *Proceedings of the 22nd International Conference on Randomization and Computation (RANDOM)*, volume 116, pages 46:1–46:19, 2018.

[KM13] Daniel M. Kane and Raghu Meka. A prg for lipschitz functions of polynomials with applications to sparsest cut. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '13, page 1–10, New York, NY, USA, 2013. Association for Computing Machinery.

[KR18] Daniel Kane and Sankeerth Rao. A PRG for Boolean PTF of degree 2 with seed length subpolynomial in ε and logarithmic in n. In *Proceedings of the 33rd Computational Complexity Conference (CCC)*, pages 2:1–2:24, 2018.

[KTZ19] Vasilis Kontonis, Christos Tzamos, and Manolis Zampetakis. Efficient truncated statistics with unknown truncation. In David Zuckerman, editor, *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA, November 9-12, 2019*, pages 1578–1595. IEEE Computer Society, 2019.

[Lee14] Alice Lee. Table of the Gaussian "Tail" Functions; When the "Tail" is Larger than the Body. *Biometrika*, 10(2/3):208–214, 1914.

[MZ13] R. Meka and D. Zuckerman. Pseudorandom Generators for Polynomial Threshold Functions. *SIAM J. Comput.*, 42(3):1275–1301, 2013.

[O'D14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[OST20] Ryan O'Donnell, Rocco A. Servedio, and Li-Yang Tan. Fooling gaussian ptfs via local hyperconcentration. In Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy, editors, *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 1170–1183. ACM, 2020.

[Pea02] Karl Pearson. On the systematic fitting of frequency curves. *Biometrika*, 2:2–7, 1902.

[Sak93] M. Saks. Slicing the hypercube. In Keith Walker, editor, *Surveys in Combinatorics 1993*, pages 211–257. London Mathematical Society Lecture Note Series 187, 1993.

[Sch86] Helmut Schneider. *Truncated and censored samples from normal populations*. Marcel Dekker, Inc., 1986.

[ST18] Rocco A. Servedio and Li-Yang Tan. Luby-velickovic-wigderson revisited: Improved correlation bounds and pseudorandom generators for depth-two circuits. In Eric Blais, Klaus Jansen, José D. P. Rolim, and David Steurer, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, (APPROX/RANDOM)*, volume 116 of *LIPIcs*, pages 56:1–56:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018.

[Tal96] M. Talagrand. How much are increasing sets positively correlated? *Combinatorica*, 16(2):243–258, 1996.

[Vem10] Santosh S. Vempala. Learning convex concepts from gaussian distributions with PCA. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 124–130. IEEE Computer Society, 2010.

[Wol07] Paweł Wolff. Hypercontractivity of simple random variables. *Studia Mathematica*, 3:219–236, 2007.