

PTF Testing Lower Bounds for Non-Gaussian Component Analysis

Ilias Diakonikolas^{*}

University of Wisconsin-Madison
ilias@cs.wisc.edu

Daniel M. Kane[†]

University of California, San Diego
dakane@cs.ucsd.edu

Sihan Liu[‡]

University of California, San Diego
si1046@ucsd.edu

Thanasis Pittas[§]

University of Wisconsin-Madison
pittas@wisc.edu

November 25, 2025

Abstract

This work studies information-computation gaps for statistical problems. A common approach for providing evidence of such gaps is to show sample complexity lower bounds (that are stronger than the information-theoretic optimum) against natural models of computation. A popular such model in the literature is the family of *low-degree polynomial tests*. While these tests are defined in such a way that make them easy to analyze, the class of algorithms that they rule out is somewhat restricted. An important goal in this context has been to obtain lower bounds against the stronger and more natural class of low-degree *Polynomial Threshold Function (PTF) tests*, i.e., any test that can be expressed as comparing some low-degree polynomial of the data to a threshold. Proving lower bounds against PTF tests has turned out to be challenging. Indeed, we are not aware of any non-trivial PTF testing lower bounds in the literature.

In this paper, we establish the first non-trivial PTF testing lower bounds for a range of statistical tasks. Specifically, we prove a near-optimal PTF testing lower bound for Non-Gaussian Component Analysis (NGCA). Our NGCA lower bound implies similar lower bounds for a number of other statistical problems. Our proof leverages a connection to recent work on pseudorandom generators for PTFs and recent techniques developed in that context. At the technical level, we develop several tools of independent interest, including novel structural results for analyzing the behavior of low-degree polynomials restricted to random directions.

^{*}Supported by NSF Medium Award CCF-2107079, and an H.I. Romnes Faculty Fellowship.

[†]Supported by NSF Medium Award CCF-2107547 and NSF Award CCF-1553288 (CAREER).

[‡]Supported by NSF Medium Award CCF-2107547 and NSF Award CCF-1553288 (CAREER).

[§]Supported by NSF Medium Award CCF-2107079.

1 Introduction

In classical statistical estimation, the focus has primarily been on determining the minimum amount of information required to estimate the parameters of an unknown distribution to a desired level of accuracy. Classical statistical theory provides general methodology to characterize this quantity for a range of estimation and inference tasks. When taking computational aspects into account, the situation becomes more subtle. Statistically optimal estimators often entail an exhaustive search. On the other hand, known computationally efficient estimators often require more data than is necessary. A fundamental question is whether these observed gaps are inherent. An information-computation gap describes a scenario where no computationally efficient method can achieve the information-theoretic limits.

A key question is how to formally establish the existence of an information-computation gap for a particular problem. Traditional methods from complexity theory, such as NP-hardness, seem inadequate for this purpose; see, e.g., [ABX08]. Over the past decade, a line of work in theoretical computer science has made progress in our understanding of this broad question. A prominent approach to establishing information-computation gaps involves showing unconditional lower bounds within natural (yet restricted) computational models—such as Statistical Query (SQ) algorithms [Kea93, FGR⁺13], low-degree polynomials (LDP) [Hop18, KWB19], and Sum-of-Squares (SOS) algorithms (see, e.g., [BS16]). These methodologies have provided rigorous evidence of information-computation tradeoffs for a range of fundamental and well-studied statistical tasks.

In this paper, we consider the class of algorithms based on low-degree Polynomial Threshold Functions (PTFs). A degree- k PTF $f : \mathbb{R}^N \rightarrow \{0, 1\}$ is a function of the form $f(\mathbf{x}) = \text{sign}(p(\mathbf{x}))$, where $p : \mathbb{R}^N \rightarrow \mathbb{R}$ is a polynomial of degree at most k and $\text{sign}(u)$ denotes the function which is equal to 1 whenever $u \geq 0$ and 0 otherwise. PTFs is a natural class of Boolean functions that has been extensively studied in complexity theory and machine learning over the past six decades; see, e.g., [Ros58, Cho61, MTT61, Der65, MP88] for some early work and [DGJ⁺10, DHK⁺10, DKN10, MZ10, Kan11b, Kan14, DS14, DDFS14, DRST14, DKS18a, DK19, OST20, KM22, DKK⁺24].

We will use the term “PTF tests” for the associated class of algorithms. As we will explain below, PTF tests are strictly stronger than LDP tests. Perhaps surprisingly, prior to this work, no non-trivial information-computation gaps were known against this class.

Background Before defining the family of PTF tests, we provide some background. We focus on hypothesis testing problems, as lower bounds for more complex tasks (like learning or parameter estimation) often stem from testing lower bounds. In particular, the null hypothesis is a single distribution D_\emptyset , and the alternative is sampled from a prior μ on a family \mathcal{D}_{alt} of distributions (μ , D , and \mathcal{D}_{alt} are known to the testing algorithm).

Problem 1.1 (Hypothesis Testing). *We are given n samples in \mathbb{R}^d generated in one of two ways:*

- *(Null Hypothesis) The samples are drawn i.i.d. from a known distribution D_\emptyset .*
- *(Alternative Hypothesis) A member D_{alt} is sampled according to a known prior distribution μ on a family of alternative distributions \mathcal{D}_{alt} , and then the n samples are drawn i.i.d. from D_{alt} .*

Given the samples, the goal is to distinguish between the two cases with high constant probability.

A natural class of tests is based on PTFs. As we review below, the well-studied class of Low-Degree Polynomial Tests consists of PTFs but they take a specific form.

Low-Degree Polynomial (LDP) Tests We now define the family of LDP tests and discuss existing testing lower bounds in this model. Informally, the family contains tests of the following form: For a polynomial p that satisfies a separation in terms of its expected values under the null and the alternative distributions, the test is the thresholded version of p at the midpoint of these expected values. These tests are usually parameterized by two numbers: the maximum degree k of the polynomial, which quantifies the runtime of

evaluating the test, and, the number n of samples used. In slightly more formal language, given a polynomial p , a null distribution D_\emptyset and an alternative distribution family \mathcal{D}_{alt} that we aim to distinguish, the “advantage” γ of the polynomial is defined as the difference in expected values of p under D_\emptyset versus under a random distribution D_{alt} from \mathcal{D}_{alt} , relative to the variance of the polynomial. We will use the notation $\mathbf{x}^{(1:n)}$ as a shorthand notation for $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ throughout the paper.

Definition 1.2 (γ -advantageous polynomial). *Let $\gamma > 0$, $p : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ be a degree- k , n -sample polynomial.*

Let D_\emptyset be a distribution in \mathbb{R}^d , \mathcal{D}_{alt} be a family of distributions in \mathbb{R}^d , μ be a distribution on \mathcal{D}_{alt} , and H be the hypothesis testing of distinguishing between D_\emptyset and \mathcal{D}_{alt} with prior μ . We say that p is a degree- k , n -sample, γ -advantageous polynomial with respect to the testing problem H if.¹

$$\left| \mathbf{E}_{\mathbf{x}^{(1:n)} \sim D_\emptyset} [p(\mathbf{x}^{(1:n)})] - \mathbf{E}_{\substack{D_{\text{alt}} \sim \mu \\ \mathbf{y}^{(1:n)} \sim D_{\text{alt}}}} [p(\mathbf{y}^{(1:n)})] \right| > \gamma \max \left(\mathbf{Var}_{\mathbf{x}^{(1:n)} \sim D_\emptyset} [p(\mathbf{x}^{(1:n)})], \mathbf{Var}_{\substack{D_{\text{alt}} \sim \mu \\ \mathbf{y}^{(1:n)} \sim D_{\text{alt}}}} [p(\mathbf{y}^{(1:n)})] \right)^{1/2}.$$

The family of n -sample, k -degree polynomial tests includes all tests $h : \mathbb{R}^{n \times d} \rightarrow \{0, 1\}$ of the following form. For every polynomial $p : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}$ of degree at most k , the family contains a test $h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) = \text{sign}(p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) - \kappa)$ that thresholds the polynomial at the point κ which is defined to be the midpoint of the two expectations $\mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim D_\emptyset} [p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})]$ and $\mathbf{E}_{D_{\text{alt}} \sim \mu, \mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \sim D_{\text{alt}}} [p(\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)})]$.

It then follows immediately from Chebyshev’s inequality that if there is a γ -advantageous polynomial p , then the test h in the family corresponding to p has bounded error probability:

$$\Pr_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim D_\emptyset} [h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) = 1] + \Pr_{D_{\text{alt}} \sim \mu, \mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \sim D_{\text{alt}}} [h(\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)}) = 0] \leq 8/\gamma^2.$$

That is, a polynomial with large advantage γ directly translates to an effective tester with low error probability. Conversely, an upper bound on γ against all low-degree polynomials indicates that no tests (with high success probability) exist under this design framework.

A convenient fact about [Definition 1.2](#) is that the optimal advantage for a testing problem is relatively easy to analyze.² In particular, if one squares the defining equation for γ -advantage, it becomes a bound on the relative size of two explicit quadratic forms over the space of all degree at most k -polynomials, which can be explicitly optimized to find the optimal value of γ . This yields a convenient framework for analyzing the power of Low-Degree Polynomial tests, in which many quantitatively tight lower bounds have been established; see, e.g., [[HS17](#), [BKW19](#), [KWB19](#), [BBH⁺21](#), [MW21](#), [SW22](#), [DKWB24](#)] for a variety of statistical problems.

(General) Polynomial Threshold Function Tests Following the above discussion, one could see that a key limitation of existing hardness results for the LDP tests family is that they only yield lower bounds against the *specific* proof technique based on second-moment Chebyshev’s inequality while leaving the general power of all *polynomial threshold function* tests, namely all tests of the form $\text{sign}(p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}))$, where p is an arbitrary low-degree polynomial, and $\mathbf{x}^{(i)}$ are the samples drawn, poorly understood; see the definition below for a formal definition of a low-degree PTF test.

Definition 1.3 (β -good PTF test). *Let D_\emptyset be a distribution in \mathbb{R}^d , \mathcal{D}_{alt} be a family of distributions in \mathbb{R}^d , μ be a distribution on \mathcal{D}_{alt} , and H be the hypothesis testing problem whose null distribution, alternative distributions family, and prior distribution are given by D_\emptyset , \mathcal{D}_{alt} , and μ accordingly. Let $h : \mathbb{R}^{n \times d} \mapsto \{0, 1\}$ be a polynomial threshold function of degree- k . We say h is a β -good PTF test for H if it satisfies that*

$$\left| \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim D_\emptyset} [h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})] - \mathbf{E}_{D_{\text{alt}} \sim \mu, \mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \sim D_{\text{alt}}} [h(\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)})] \right| \geq \beta. \quad (1)$$

¹Many works focusing on lower bounds, use only the variance under D_\emptyset in the RHS of the condition in [Definition 1.2](#).

²This holds for the variant where the RHS in [Definition 1.2](#) includes only the variance with respect to D_\emptyset , which is usually a much simpler distribution. Using this variant suffices for the purpose of proving upper bounds on γ .

There is no clear reason to expect that the failure of this specific technique on a test problem would rule out all PTFs. Indeed, as we show in [Section B](#), there exist simple examples of pairs of null and alternative distributions for which no polynomial satisfies the separation condition of [Definition 1.2](#), yet PTF tests effectively solve the testing problem with high success probability.

Understanding the power of general PTF tests is considered as a prominent research direction within the relevant community [[Hop24](#), [Wei24](#)]. Specifically, a recent workshop [[HSW24](#)] on information-computation tradeoffs highlighted PTF tests as one of the main directions in the frontier of this field. In his new survey [[Wei](#)], Wein writes: “[...] it is an interesting open problem to rule out other notions of success such as thresholding, but this seems beyond our current capabilities.” In this work, we therefore ask the question:

Can we rigorously prove information-computation gaps within the family of all PTF tests?

As our main contribution, we answer this question in the affirmative. In particular, we give the first PTF testing lower bound for the fundamental task of *Non-Gaussian Component Analysis*. As an immediate corollary, we obtain PTF lower bounds for a number of other statistical problems.

Non-Gaussian Component Analysis Historically, NGCA is a problem that originates from the signal processing literature [[BKS⁺06](#)], and has since attracted much attention from the algorithmic statistics and theoretical machine learning communities; see [[SKBM08](#), [SNS16](#), [DJNS13](#), [GS19](#), [DH24](#)]. Informally, the problem corresponds to the task of searching for a non-Gaussian direction of some high-dimensional distribution. The testing version of NGCA aims at distinguishing between a high dimensional standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and a distribution that is non-Gaussian along an unknown direction, but behaves like standard Gaussian in every other orthogonal direction. The alternative distribution, commonly referred to as the hidden-direction distribution is defined below:

Definition 1.4 (High-Dimensional Hidden Direction Distribution). *For a distribution A on \mathbb{R} and a unit vector v of \mathbb{R}^d , we denote by $\mathcal{M}_{A,v}$ the distribution of the random variable $x + \xi v$, where $x \sim \mathcal{N}(\mathbf{0}, \mathbf{I} - vv^\top)$ and $\xi \sim A$. That is, $\mathcal{M}_{A,v}$ is the distribution which coincides with A on the direction of v and is standard Gaussian in the orthogonal subspace.*

The standard assumption is that the non-Gaussian component A is also similar to $\mathcal{N}(0, 1)$ in the sense that their first m moments match. The higher the m , the harder it becomes to distinguish $\mathcal{M}_{A,v}$, for random v , from the standard multivariate Gaussian. On the other hand, for algorithms to exist, one needs to assume that the $(m + 1)$ st moment differs by a non-trivial amount.

Problem 1.5 (Non-Gaussian Component Analysis). *Let A be a distribution in \mathbb{R} that matches the first m moments with $\mathcal{N}(0, 1)$. We are given n samples generated in one of the following two ways:*

- (Null Hypothesis) *The samples are drawn i.i.d. from $D_\emptyset = \mathcal{N}(\mathbf{0}, \mathbf{I})$.*
- (Alternative Hypothesis) *First, a unit vector $v \in \mathbb{R}^d$ is drawn uniformly at random from the unit sphere, then the samples are drawn i.i.d. from $D_{\text{alt}} = \mathcal{M}_{A,v}$ from [Definition 1.4](#).*

Given the samples, the goal is to distinguish between the two cases with high constant probability.

A concrete motivation in studying NGCA is that it exhibits *information-computation gaps* when A is carefully constructed to match many moments with the standard Gaussian. Information theoretically, it is known that the sample complexity of this problem is $O(d)$ under some mild assumptions on the distribution A (see, e.g., [[VX11](#)]). Nonetheless, all known efficient algorithms require significantly larger resources (see e.g., [[DH24](#)]). Prior work has given formal evidence that solving NGCA requires either access to a large amount of information (specifically, many i.i.d. samples from the test distribution) or significant computational resources. Concretely, this phenomenon has been established for several well-studied families of algorithms, including Statistical Query (SQ) algorithms [[DKS17](#), [DKRS23](#)], Sum-of-Squares (SoS) algorithms [[DKPP24](#)], and

Low-Degree Polynomial tests [MW21, BBH⁺21]. Since NGCA can be used to embed hard instances of several other statistical tasks [DKS17], a lower bound against NGCA directly implies lower bounds for several other problems. In this work, we give the first lower bound of NGCA against the family of PTF tests.

Connection to Pseudorandom Generators There is a close connection between lower bounds for hypothesis testing and the theory of pseudorandom generators (PRGs). Specifically, for a test to effectively distinguish between the null and alternative hypotheses, the probability of accepting each hypothesis must differ significantly. In the language of PRGs, Definition 1.3 failing implies that $D_{\text{alt}}^{\otimes n}$ “fools” h with respect to $D_{\emptyset}^{\otimes n}$ with error at most β . Thus, proving lower bounds against degree- k PTF tests with n samples reduces to showing that $D_{\text{alt}}^{\otimes n}$ fools low-degree PTFs with respect to $D_{\emptyset}^{\otimes n}$.

There is an extensive literature on PRGs for PTFs. These works generally aim to construct low-entropy distributions that fool PTFs under structured high-entropy distributions such as Gaussians, see, e.g., [DGJ⁺09, DKN10, MZ10, Kan11a, Kan11b, Kan12, OST20, KM22]. Despite this distinction, some techniques developed in the PRG literature (in particular [KM22]) can be leveraged in our setting.

For our specific NGCA problem, the null distribution is precisely the standard Gaussian—for which some of the strongest PRGs for PTFs are known. Moreover, the alternative distributions are assumed to match many moments with the Gaussian, a condition that is essentially necessary for indistinguishability and one that is also required by SQ and Low-degree lower bounds for establishing information-computation gaps in the literature.

1.1 Our Result

Our main theorem establishes that when the non-Gaussian component A from Problem 1.5 matches the first m moments with $\mathcal{N}(0, 1)$, there is no degree- k PTF that satisfies Definition 1.3 with $\beta = 0.11$ for the NGCA hypothesis testing problem of Problem 1.5, unless at least one of the following holds: The sample complexity n is at least $d^{\Omega(m)}$ or the degree k of the PTF is at least $d^{\Omega(1)}$ (suggesting a runtime of $(nd)^{d^{\Omega(1)}}$). Notably, this matches quantitatively with the computation-statistic tradeoff established for the weaker model of LDP tests.

Theorem 1.6 (Main Result). *There exists a sufficiently large absolute constant C^* such that the following holds. For any $c^* \in (0, 1/4)$, $d, k, n, m \in \mathbb{Z}_+$ such that (i) m is even, (ii) $\max(k, m) < d^{c^*/C^*}$, and (iii) $n < d^{(1/4-c^*)m}$, we have that if $p : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ is a degree- k polynomial, and A is a distribution on \mathbb{R} that matches the first m moments with $\mathcal{N}(0, 1)$, then:*

$$\left| \mathbf{E}_{\substack{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1}) \\ \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{M}_{A, \mathbf{v}}}} [\text{sign}(p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}))] - \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\text{sign}(p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}))] \right| \leq 0.11. \quad (2)$$

where $\mathcal{M}_{A, \mathbf{v}}$ denotes the hidden direction distribution from Definition 1.4, and $\text{sign} : \mathbb{R} \rightarrow \{0, 1\}$ is the sign function with $\text{sign}(x) = 1$ if and only if $x \geq 0$.

One way of interpreting Theorem 1.6 is as follows. Let c^* be an arbitrary constant chosen from $(0, 1/4)$. If the PTF test uses $n < d^{(1/4-c^*)m}$ many samples, then the degree of the tester k must be at least some polynomial in the sample dimension d in order for the tester to be effective. The degree k can be interpreted as a parameter controlling the runtime of the tester. For an arbitrary polynomial p , the runtime for this computation is on the order of $\text{poly}((nd)^k)$ —as this is the space required for storing all relevant coefficients of p . Consequently, Theorem 1.6 implies an inherent trade-off between the exponential runtime $(nd)^{d^{\Theta(c^*)}}$, and the sample complexity $d^{(1/4-c^*)m}$ for the family of PTF tests.

One might wonder to what extent is the tradeoff established in Theorem 1.6 optimal, and how our result compares quantitatively to the hardness results shown in other restricted computation models. For hardness results against Statistical Query algorithms and LDP tests, existing lower bounds imply that there are no “efficient” testers within these families if the number of samples drawn is less than $d^{(1/2-c^*)m}$ (see

[Theorem D.1](#) and [Theorem D.3](#) for the formal statements). Interestingly, the constant $1/2$ in the exponent of these results is better than the constant $1/4$ that appears in [Theorem 1.6](#). Surprisingly, this gap—rather than being an artifact of our proof technique—is inherent for the family of PTF tests. In particular, if one fixes m to be some constant, we show in [Theorem C.1](#) that there exists a degree- $\Theta(\log d)$ PTF that draws only $\tilde{\Theta}(d^{m/4})$ many samples, and effectively solves the NGCA problem for some specific moment-matching distribution A . As a sharp contrast, the SQ, and LDP test lower bounds predict that no such test should exist. This suggests that the constant $1/4$ that appears in [Theorem 1.6](#) is indeed worst-case optimal, and that PTF tests are (slightly) more powerful than SQ and LDP tests for certain NGCA problem instances. We leave it as an interesting open question whether the lower bound against PTF tests can be improved by making further structural assumptions on the non-Gaussian component A beyond the moment-matching condition.

Before we end this subsection, we briefly comment on the subtle condition that $m < d^{c^*/C^*}$. This condition is commonly used in the literature for lower bounds against NGCA (see [[DKS17](#), [DKRS23](#)]). Informally, the necessity of such a condition can be seen as follows. If $m \approx d$, one cannot hope to prove a computation lower bound of $(nd)^m \approx 2^{d \log d}$. There is always a simple test that runs in time $2^{O(d)}$: construct an exponential size cover of all possible directions \mathbf{v} , project the distribution along each possible direction and reduce to a one-dimensional testing problem.

Application to Other Statistical Tasks For many important statistical problems—such as robust mean estimation, list-decodable mean estimation, and learning Gaussian mixture models—that are seemingly unrelated to each other, one can construct hard instances that can be encoded as NGCA instances; see Chapter 8 of [[DK23](#)] or Section 1 of [[DKPP24](#)] for a more thorough treatment. This means that proving an information-computation gap for NGCA within a given computational model translates to information-computation gap for all these tasks. See [Table 1](#) for a summary of some PTF testing lower bounds obtained as corollaries of [Theorem 1.6](#).

Information-Computation Gaps for PTF tests		
Statistical Estimation Task	Information-Theoretic	Sample Complexity for low-degree PTFs
Robust Mean Estimation up to ℓ_2 -error $O(\tau\sqrt{\log(1/\tau)/B^2})$ with Isotropic Gaussians	$O_{\tau,B}(d)$	$\Omega(d^{B(1-c^*)/4})$
Robust Mean Estimation up to ℓ_2 -error $O(\frac{1}{m}\tau^{1-1/m})$ with bounded m -th moments	$O_{\tau}(d)$	$\Omega(d^{m(1-c^*)/4})$
List-decodable Mean Estimation to error $O((m\tau)^{-1/m})$	$O_{\tau}(d)$	$\Omega(d^{m(1-c^*)/4})$
Learning the mixture of m Gaussians	$\tilde{O}(md)$	$\Omega(d^{m(1-c^*)/2})$

Table 1: Comparison of our PTF lower bounds with the information-theoretic sample complexity for various tasks. The parameter τ is the rate of contamination, and the parameter c^* can be set to any arbitrarily small constant. See [Section E](#) for the formal statements of the results that appear in the table.

Comparison with Existing Lower Bounds In general, lower bounds against PTF tests are incomparable to SQ and SoS lower bounds, as they capture different structural limitations of learning algorithms. For the specific problem of Non-Gaussian Component Analysis, SQ lower bounds are effectively equivalent to Low-Degree Polynomial (LDP) lower bounds as established in [[BBH⁺21](#)]. Our result therefore strengthens prior SQ and LDP lower bounds by demonstrating hardness under the more general PTF testing framework.

Future Directions Our work proves the first lower bounds against general PTF tests, a strengthening of the well-studied low-degree polynomial test family, for multiple statistical problems. Several open questions remain for future research. A key technical question is whether our lower bounds can be quantitatively improved under additional structural assumptions on the non-Gaussian component distribution A that arise

naturally in learning-theoretic settings, like bounded chi-square distance between A and $\mathcal{N}(0, 1)$. Identifying such refinements could lead to sharper hardness results and a deeper understanding of the limitations of PTF tests. Another important direction is to obtain PTF testing lower bounds for other fundamental statistical estimation problems not covered in this paper, such as planted clique [BHK⁺16] and sparse principal component analysis [ZHT06].

1.2 Technical Overview

We will use the notation $\mathbf{x}^{(1:n)}$ to denote the sequence of vectors $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$. We want to show that for any low degree polynomial $p : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$, we have that

$$\mathbf{E}[\text{sign}(p(\mathbf{x}^{(1:n)}))] \approx \mathbf{E}[\text{sign}(p(\mathbf{y}^{(1:n)}))], \quad (3)$$

where each $\mathbf{x}^{(i)} \in \mathbb{R}^d$ follows an independent standard Gaussian, and each $\mathbf{y}^{(i)} \in \mathbb{R}^d$ follows our hidden direction distribution $\mathcal{M}_{A,\mathbf{v}}$ (see [Definition 1.4](#)) using a fixed vector \mathbf{v} , which is sampled once uniformly at random from the unit sphere at the start.

We begin with a brief recap of the setup in the PRG literature, and present a high level comparison between these approaches and ours. These works consider a generator distribution $\bar{\mathbf{y}} = n^{-1/2} \sum_{i=1}^n \bar{\mathbf{y}}^{(i)}$ in \mathbb{R}^d , where each $\bar{\mathbf{y}}^{(i)}$ is chosen to be some low-entropy distribution whose low-degree moments match with the standard Gaussian.³ Given an arbitrary low-degree polynomial $q : \mathbb{R}^d \mapsto \mathbb{R}$, the goal is then to show that $\mathbf{E}[\text{sign}(q(\mathbf{x}))] \approx \mathbf{E}[\text{sign}(q(n^{-1/2} \sum_{i=1}^n \bar{\mathbf{y}}^{(i)}))]$, where \mathbf{x} is a standard Gaussian vector in \mathbb{R}^d . Note that the standard Gaussian \mathbf{x} can be alternatively written as $\mathbf{x} = n^{-1/2} \sum_{i=1}^n \mathbf{x}^{(i)}$, where the $\mathbf{x}^{(i)}$'s are themselves independent standard Gaussian distributions. Under this setup, their objective can be alternatively formulated

$$\mathbf{E}[\text{sign}(\bar{p}(\mathbf{x}^{(1:n)}))] \approx \mathbf{E}[\text{sign}(\bar{p}(\bar{\mathbf{y}}^{(1:n)}))] \quad \text{where } \bar{p}(\mathbf{z}^{(1:n)}) = q\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{z}^{(i)}\right). \quad (4)$$

Under this new formulation, one can see that our objective [Equation \(3\)](#) looks particularly similar to theirs except for two subtle differences. First, in the PRG setup, the polynomial $\bar{p} : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ has the specific form of being a lifted version of a significantly *lower-dimensional* polynomial $q : \mathbb{R}^d \mapsto \mathbb{R}$ (see [Equation \(4\)](#)). In fact, the structure turns out to be quite convenient for controlling the higher order derivatives of \bar{p} in terms of each individual variable $\mathbf{z}^{(i)}$, which are subsequently exploited in the PRG literature to establish the left part of [Equation \(4\)](#). In our setup, the polynomial p has a much more complex structure as it can be an *arbitrary* polynomial with $n \times d$ many inputs. This makes the task seemingly intractable at a first glance. The twist lies in the second difference between our setups: the underlying distribution of each $\mathbf{y}^{(i)}$ in our objective exhibits significantly more structure compared to the pseudorandom distributions of $\bar{\mathbf{y}}^{(i)}$. Specifically, instead of being an arbitrary distribution whose low-degree moments match with the standard Gaussian, the distribution of $\mathbf{y}^{(i)}$ is *identical* to the Gaussian distribution in all but some *randomly* chosen direction \mathbf{v} . This turns out to be a valuable property that counteracts the complexity of the polynomial p in our setup. More concretely, our main insight is that the specific properties of the hidden direction distribution essentially allow us to restrict our attention to higher order *directional* derivatives of p (projected onto the hidden direction \mathbf{v}). By exploiting the fact that \mathbf{v} is chosen randomly, we manage to show that the higher order directional derivatives of p can be effectively bounded in a way that is qualitatively similar to (but quantitatively different from) the case in the PRG literature, allowing us to escape from the full complexity of the polynomial p in all directions.

In what follows, we give a more detailed sketch of our arguments interleaved with technical comparisons with the prior work in the PRG literature.

³In fact, each Y_i is independently from a k -wise independent family of Gaussian distributions, which effectively satisfies the low-entropy and moment-matching conditions.

Overall Framework: Hybrid Argument and Mollification The basic proof strategy is via the so called *hybrid argument* developed in the PRG literature. At a high level, the framework is motivated by the wishful thinking that the task may be significantly easier if one progressively replaces the samples $\mathbf{x}^{(1:n)}$ into $\mathbf{y}^{(1:n)}$. In particular, in a single “replacement step”, the goal is to show the following intermediate approximation steps for all $i = 1, \dots, n$:

$$\mathbf{E} \left[\text{sign} \left(p \left(\mathbf{x}^{(1:i-1)}, \mathbf{x}^{(i)}, \mathbf{y}^{(i+1:n)} \right) \right) \right] \approx \mathbf{E} \left[\text{sign} \left(p \left(\mathbf{x}^{(1:i-1)}, \mathbf{y}^{(i)}, \mathbf{y}^{(i+1:n)} \right) \right) \right] \pm o(1/n). \quad (5)$$

If so, applying the triangle inequality $n - 1$ times would complete the proof of [Equation \(3\)](#).

In both the PRG and the NGCA settings, one would like to exploit the assumption that the alternative distribution matches many moments with the standard Gaussian. Hence, a natural attempt to show [Equation \(5\)](#) would be to rewrite the expected values in terms of the moments of the corresponding distributions. If the PTF function were a smooth function, this can be achieved by considering the Taylor expansion of the function. However, it is not hard to see that the function is discontinuous when the polynomial p evaluates to 0, making Taylor’s theorem not directly applicable.

To circumvent the issue, the PRG literature uses the idea of *mollification*. Specifically, we would like to construct a sufficiently smooth function $h : \mathbb{R}^{n \times d} \mapsto [0, 1]$ that approximates the PTF $\text{sign}(p(\cdot))$ well. As we have said, the PTF function $\text{sign}(p(\cdot))$ is discontinuous when $p(\mathbf{x}) = 0$. Hence, we cannot hope to have a smooth and pointwise close approximation near the zeros of the polynomial. Instead, we want h to have the same behavior as the PTF when the polynomial is “large” (in a technical sense that will be specified later), and then smoothly interpolate in the other case. On the one hand, this effectively ensures the smoothness property of h globally. On the other hand, h indeed approximates the PTF well under the Gaussian distribution⁴ as the only disagreement region is when the polynomial p is small, whose probability mass can be bounded by the Gaussian anti-concentration properties or some variants (that will be discussed later on). For convenience, we refer to this smoothed approximation h of the PTF as the *mollified PTF*, and our task is now reduced to showing

$$\mathbf{E} \left[h \left(p \left(\mathbf{x}^{(1:i-1)}, \mathbf{x}^{(i)}, \mathbf{y}^{(i+1:n)} \right) \right) \right] \approx \mathbf{E} \left[h \left(p \left(\mathbf{x}^{(1:i-1)}, \mathbf{y}^{(i)}, \mathbf{y}^{(i+1:n)} \right) \right) \right] \pm o(1/n). \quad (6)$$

It then remains for us to (1) construct a smooth mollified PTF h that closely approximates $\text{sign}(p(\cdot))$ under Gaussian inputs, and (2) show a single replacement step ([Equation \(6\)](#)) for this smooth function h (the formal version of this is in [Proposition 3.5](#)).

Mollification by Strong Anti-Concentration The difficulty of (1) is as follows: since the PTF function is discontinuous and h is smooth, we cannot expect $|h(\mathbf{x}^{(1:n)}) - \text{sign}(p(\mathbf{x}^{(1:n)}))|$ to be small for all inputs. Alternatively, one must rely on some type of *anti-concentration* result saying that the probability of $\mathbf{x}^{(1:n)}$ lying in the disagreement region between h and $\text{sign}(p(\cdot))$ is small under the Gaussian distribution. A natural idea to do so would be to construct a smooth approximation of the indicator function $g(\mathbf{x}^{(1:n)}) \approx \mathbb{1}\{|p(\mathbf{x}^{(1:n)})| > \varepsilon\}$, and define the mollified PTF to be the product $g(\mathbf{x}^{(1:n)}) \text{sign}(p(\mathbf{x}^{(1:n)}))$. On the one hand, this ensures that the function will smoothly interpolate between 0 and 1 for inputs $\mathbf{x}^{(1:n)}$ near the zeros of p . On the other hand, the disagreement region will be roughly the same as the set $\{\mathbf{x}^{(1:n)} \mid p(\mathbf{x}^{(1:n)}) < \varepsilon\}$, which is guaranteed to have small mass by the famous Gaussian anti-concentration theorem from [\[CW01\]](#). However, for an arbitrary polynomial p , the anti-concentration property decays rapidly as the degree of the polynomial

⁴A technical detail we omit here is that naively it seems like one also needs to show a similar approximation result under the distribution of $\mathbf{y}^{(1:n)}$, which is technically challenging. Fortunately, this can be solved by a standard sandwich trick that reduces the task into constructing two mollified PTFs h_+, h_- such that (i) $h_+(\cdot) \geq \text{sign}(p(\cdot)) \geq h_-(\cdot)$, and (ii) h_+, h_- approximates the PTF function well under the Gaussian distribution. This saves us from the trouble of showing how well the mollified PTF approximates the PTF under the less structured distributions of $\mathbf{y}^{(1:n)}$. We refer the reader to [Section 3.2](#), and more specifically [Lemma 3.4](#) for more detail on this.

increases, i.e., there exists a degree- k polynomial p such that $\Pr[p(\mathbf{x}^{(1:n)}) < \varepsilon] \approx \varepsilon^{1/k}$. In the context of the PRG literature, this leads to an exponential dependency on the seed length⁵ while in our context the approach might break entirely once the polynomial degree k becomes larger than the logarithm of the sample dimension d . Needless to say, such an assumption would significantly weaken the lower bound on k compared to the target of $k = d^{\Omega(1)}$ in [Theorem 1.6](#).

To tackle the issue, we borrow the ideas from [[OST20](#), [Kan11b](#), [KM22](#)] that take advantage of the *strong anti-concentration* properties of polynomials. In particular, strong anti-concentration (see [Equation \(13\)](#)) is a relative notion of anti-concentration on the sizes of the derivatives of the polynomials: given a polynomial p of degree k , for any $0 \leq t \leq k-1$ and $\varepsilon \in (0, 1)$, it holds that $\|\nabla^t p(\mathbf{x}^{(1:n)})\|_F > \varepsilon \|\nabla^{t+1} p(\mathbf{x}^{(1:n)})\|_F$ with probability at least $1 - O(k^2\varepsilon)$, where $\nabla^t p(\mathbf{x}^{(1:n)})$ denotes the tensor containing all t -th order partial derivatives of p . Notably, unlike the Carbery-Wright anti-concentration theorem, the failure probability here is only a polynomial in k , and it is precisely this polynomial dependency that makes the stronger lower bound on k from [Theorem 1.6](#) possible. By setting $\varepsilon = k^{-2.1}$, applying the union bound, and chaining the inequalities obtained, we obtain that

$$\left| p(\mathbf{x}^{(1:n)}) \right| > k^{-2.1} \left\| \nabla p(\mathbf{x}^{(1:n)}) \right\|_F > k^{-6.2} \left\| \nabla^2(\mathbf{x}^{(1:n)}) \right\|_F > \dots > k^{-2.1k} \left\| \nabla^k p(\mathbf{x}^{(1:n)}) \right\|_F, \quad (7)$$

with high constant probability. Note that $\nabla^k p$ is a constant tensor. Thus, as long as [Equation \(7\)](#) holds approximately (within a constant factor), we can infer that $p(\mathbf{x}^{(1:n)})$ must be non-zero and the PTF sign ($p(\mathbf{x}^{(1:n)})$) will be smooth. Therefore, it suffices to modify the function on inputs $\mathbf{x}^{(1:n)}$ where [Equation \(7\)](#) is not approximately true. In particular, we can define a function $\rho : \mathbb{R} \mapsto [0, 1]$ that serves as a smooth approximation of the indicator function $\mathbb{1}\{|z| \leq 1\}$, and set the mollified PTF to be

$$\tilde{h}(\mathbf{x}^{(1:n)}) := \prod_{t=1}^k \rho \left(\frac{k^{-\Theta(1)} \|\nabla^t p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F^2}{\|\nabla^{t-1} p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F^2} \right) \text{sign}(p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})). \quad (8)$$

When [Equation \(7\)](#) holds, \tilde{h} is simply the same as the smooth part of the original PTF function. As the value of $|p(\mathbf{x}^{(1:n)})|$ approaches 0, $\frac{k^{-\Theta(1)} \|\nabla^t p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F^2}{\|\nabla^{t-1} p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F^2}$ for some t must begin to exceed 1. Consequently, the function $\rho(\cdot)$, which we define to be a smooth approximation of the indicator function $\mathbb{1}\{|x| \leq 1\}$ will decay smoothly until it reaches 0, ensuring the smoothness of the function \tilde{h} globally. Moreover, since [Equation \(7\)](#) holds with high constant probability, this ensures \tilde{h} approximates $\text{sign}(p(\cdot))$ up to a small constant error over Gaussian inputs.

As promising as it may seem, there are still substantial technical difficulties in showing the replacement step ([Equation \(6\)](#)) for this particular mollified PTF \tilde{h} . At a high level, the difficulty stems from a design flaw in which [Equation \(8\)](#) fails to leverage the hidden directional structure of the underlying distribution of $\mathbf{y}^{(1:n)}$. In the rest of the subsection, we will present a natural attempt to prove the replacement step for this mollified PTF \tilde{h} , illustrate the difficulty encountered, and present a simple modification on top of \tilde{h} to obtain our actual mollified PTF h (see [Equation \(10\)](#)).

Replacement Step by Taylor Expansion Consider the following natural attempt in showing the i -th replacement step ([Equation \(6\)](#)) for the mollified PTF \tilde{h} . Thanks to the smoothness property of \tilde{h} , we can rewrite \tilde{h} using its Taylor expansion in terms of the variable \mathbf{z} to be replaced:

$$\tilde{h}(\mathbf{x}^{(1:i-1)}, \mathbf{z}, \mathbf{y}^{(i+1:n)}) = \sum_{t=0}^{\infty} \left\langle \nabla_i^t \tilde{h}(\mathbf{x}^{(1:i-1)}, \mathbf{0}, \mathbf{y}^{(i+1:n)}), \mathbf{z}^{\otimes t} \right\rangle, \quad (9)$$

⁵This was indeed the case for the early work [[MZ10](#)] based on such naive mollification procedures.

where $\nabla_i^t \tilde{h}$ denotes the tensor containing all t -th order partial derivatives of \tilde{h} with respect to its i -th argument. Suppose that the first degree- m moments of $\mathbf{z} = \mathbf{x}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{z} = \mathbf{y}^{(i)} \sim \mathcal{M}_{A,\mathbf{v}}$ match exactly. We then have that the difference $\mathbf{E}_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\tilde{h}(\mathbf{x}^{(1:i-1)}, \mathbf{z}, \mathbf{y}^{(i+1:n)})] - \mathbf{E}_{\mathbf{z} \sim \mathcal{M}_{A,\mathbf{v}}} [\tilde{h}(\mathbf{x}^{(1:i-1)}, \mathbf{z}, \mathbf{y}^{(i+1:n)})]$ comes only from the higher order terms (the terms with $t > m$ from Equation (9)). As we have said, the function \tilde{h} is carefully constructed to be as smooth as the polynomial p . Concretely, one can show by some straightforward computation that $\|\nabla_i^t \tilde{h}(\mathbf{x}^{(1:n)})\|_F$ should be roughly the same as $\|\nabla_i^t p(\mathbf{x}^{(1:n)})\|_F / |p(\mathbf{x}^{(1:n)})|$. In the PRG literature, the particular polynomial p showing up is of the form $p(\mathbf{x}^{(1:n)}) = q\left(\frac{1}{\sqrt{n}}(\mathbf{x}^{(1)} + \dots + \mathbf{x}^{(n)})\right)$. Due to the specific form of p , one can see that p has only a mild dependence on the i -th sample. Consequently, it is not hard to show that $\|\nabla_i^t p(\mathbf{x}^{(1:n)})\|_F / |p(\mathbf{x}^{(1:n)})|$ is at most $n^{-\Theta(t)}$. If we were to have the same bound on higher-order derivatives in our case, we could then consider the Taylor expansion truncated to the first degree- m terms, which gives that

$$\tilde{h}(\mathbf{x}^{(1:i-1)}, \mathbf{z}, \mathbf{y}^{(i+1:n)}) = \sum_{t=0}^{m-1} \left\langle \nabla_i^t \tilde{h}(\mathbf{x}^{(1:i-1)}, \mathbf{0}, \mathbf{y}^{(i+1:n)}), \mathbf{z}^{\otimes t} \right\rangle + \left\langle \nabla_i^m \tilde{h}(\mathbf{x}^{(1:i-1)}, \hat{\mathbf{z}}, \mathbf{y}^{(i+1:n)}), \mathbf{z}^{\otimes m} \right\rangle,$$

where $\hat{\mathbf{z}}$ is some vector that lies in the line between \mathbf{z} and $\mathbf{0}$. After that, we note that the expected values of the first $(m-1)$ terms match exactly under $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathcal{M}_{A,\mathbf{v}}$ while the last term is on the order of $n^{-\Theta(m)} \ll o(1/n)$. This would then readily conclude the proof of the replacement step.

However, in our case, since p can be an arbitrary low-degree polynomial, we cannot say that p and subsequently \tilde{h} have only a weak dependence on the i -th sample. In particular, the best bound on $\|\nabla_i^t p(\mathbf{x}^{(1:n)})\|_F / |p(\mathbf{x}^{(1:n)})|$ (and therefore $\|\nabla_i^t \tilde{h}\|_F$) will be on the order of $k^{\Theta(t)}$ (due to the tightness of Equation (7)), which is an *increasing* function in t . As a result, the mollified PTF \tilde{h} simply cannot be approximated by its low-degree Taylor expansion, and this appears to be a substantial barrier in showing the replacement step (Equation (6)) for \tilde{h} .

Our key insight is that we can instead leverage the specific structure of the underlying hidden direction distribution. In particular, we can exploit the fact that the distribution of $\mathbf{y}^{(i)}$ only differs from that of $\mathbf{x}^{(i)}$ in some randomly chosen direction \mathbf{v} . As a result, instead of taking the Taylor expansion of \tilde{h} viewed as a multivariate function in terms of the entire i -th sample, we can now consider the *directional* Taylor expansion of \tilde{h} restricted to the \mathbf{v} direction. That is, we view \mathbf{z} as $\bar{\mathbf{z}} + \xi \mathbf{v}$ for some $\bar{\mathbf{z}}$ orthogonal to \mathbf{v} , and we use the Taylor expansion with respect to the scalar variable ξ :

$$\tilde{h}(\mathbf{x}^{(1:i-1)}, \bar{\mathbf{z}} + \xi \mathbf{v}, \mathbf{y}^{(i+1:n)}) = \sum_{t=0}^{m-1} D_{i,\mathbf{v}}^t \tilde{h}(\mathbf{x}^{(1:i-1)}, \bar{\mathbf{z}}, \mathbf{y}^{(i+1:n)}) \xi^t + D_{i,\mathbf{v}}^m \tilde{h}(\mathbf{x}^{(1:i-1)}, \bar{\mathbf{z}} + \hat{\xi}, \mathbf{y}^{(i+1:n)}) \xi^m,$$

where $\hat{\xi}$ is some point in $[0, \xi]$, and $D_{i,\mathbf{v}}^t \tilde{h}$ denotes the t -th order directional derivative of \tilde{h} with respect to the i -th sample, i.e., $D_{i,\mathbf{v}} \tilde{h} = \mathbf{v}^\top \nabla_i \tilde{h}$. Using the directional Taylor expansion, one can then write

$$\begin{aligned} & \mathbf{E}_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\tilde{h}(\mathbf{x}^{(1:i-1)}, \mathbf{z}, \mathbf{y}^{(i+1:n)})] - \mathbf{E}_{\mathbf{z} \sim \mathcal{M}_{A,\mathbf{v}}} [\tilde{h}(\mathbf{x}^{(1:i-1)}, \mathbf{z}, \mathbf{y}^{(i+1:n)})] \\ &= \mathbf{E}_{\bar{\mathbf{z}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I} - \mathbf{v}\mathbf{v}^\top)} \left[\sum_{t=0}^{\infty} D_{i,\mathbf{v}}^t \tilde{h}(\mathbf{x}^{(1:i-1)}, \bar{\mathbf{z}}, \mathbf{y}^{(i+1:n)}) \left(\mathbf{E}_{\xi \sim A} [\xi^t] - \mathbf{E}_{\xi \sim \mathcal{N}(0,1)} [\xi^t] \right) \right]. \end{aligned}$$

Since we assume that the first degree m moments of A match exactly with $\mathcal{N}(0, 1)$, it suffices to argue that the size of the t -th order directional derivative $D_{i,\mathbf{v}}^t \tilde{h}$ is of diminishing size as a function of t .

Derivative Decay in a Random Direction Towards this goal, let $\mathbf{x}^{(1:n)}$ be input samples following the standard Gaussian distribution, and we will show that the directional derivative $D_{i,\mathbf{v}}^t \tilde{h}(\mathbf{x}^{(1:n)})$ will have small

size with high constant probability, where the randomness is over \mathbf{v} and $\mathbf{x}^{(1:n)}$. With some straightforward computation, one can show that the above derivative has about the same size as the tensor containing all t -th order directional derivatives of the polynomial p along the direction of \mathbf{v} . Specifically, we consider the tensor $p^{[t],\mathbf{v}}(\mathbf{x}^{(1:n)}) \in (\mathbb{R}^n)^{\otimes t}$ defined as follows:

$$p_{i_1,\dots,i_t}^{[t],\mathbf{v}}(\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)}) = D_{i_1,\mathbf{v}} D_{i_2,\mathbf{v}} \cdots D_{i_t,\mathbf{v}} p(\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)}).$$

Alternatively, this directional derivative of p can be written as a product between the gradient tensor $\nabla^t p(\mathbf{x}^{(1:n)})$ and the “random direction tensor” $\mathbf{v}^{\otimes t}$. By the strong anti-concentration property of Gaussian, $\mathbf{x}^{(1:n)}$ satisfies Equation (7), and $\nabla^t p(\mathbf{x}^{(1:n)})$ must have size at most $k^{\Theta(t)} |p(\mathbf{x}^{(1:n)})|$ with high constant probability. Therefore, the directional derivative can be large only if the full gradient tensor $\nabla^t p(\mathbf{x}^{(1:n)})$ correlates well with $\mathbf{v}^{\otimes t}$. Since the distribution of $\mathbf{x}^{(1:n)}$ is rotationally invariant (as a property of the standard Gaussian) and \mathbf{v} is chosen randomly, it can be intuitively seen that the correlation will be small with high constant probability. In particular, via some technical tensor computation, we show in Lemma 3.1 that $\left\| p_{i_1,\dots,i_t}^{[t],\mathbf{v}}(\mathbf{x}^{(1:n)}) \right\|_F / |p(\mathbf{x}^{(1:n)})|$ will be on the order of $d^{-t/4} k^{O(t)}$ with high constant probability over the random choice of \mathbf{v} and $\mathbf{x}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Consequently, if we assume that the degree k is a sufficiently small polynomial in d , i.e., $k \ll d^{\Theta(1)}$, $d^{-t/4} k^{O(t)}$ should then be an exponentially decreasing function in t , implying that \tilde{h} is indeed well-approximated by its low-degree directional Taylor expansion.

However, one subtle technical issue remains. That is, the above argument works only for the first replacement step when the input samples all follow the standard Gaussian distribution. In later replacement steps, some of the samples have already been replaced with the hidden direction distribution, which can make the gradient correlate strongly with the hidden direction \mathbf{v} . As a result, we can no longer effectively control the size of the directional derivative even for $\mathbf{x}^{(1:n)}$ satisfying the derivative decay condition in Equation (7).

Controlling Derivatives by Fine-Tuning the Mollifier To circumvent the issue, we proceed with the following simple modification to the mollifier function to explicitly check for derivative decay along the \mathbf{v} -direction: instead of zeroing out the inputs on which the full derivative tensor $\nabla^t p$ violates the strong anti-concentration property stated in Equation (7), we now define a new mollified PTF defined directly with respect to the directional derivatives of p :

$$h(\mathbf{x}^{(1:n)}) := \prod_{t=1}^k \rho \left(\frac{d^{\Theta(t)} \|p^{[t],\mathbf{v}}(\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)})\|_F^2}{p(\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)})^2} \right) \text{sign}(p(\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)})) , \quad (10)$$

where ρ is as before a smooth approximation of the indicator function $\mathbb{1}\{|z| \leq 1\}$. On the one hand, as a direct implication of Lemma 3.1, we have that the disagreement region between h and the original PTF will be small for a randomly chosen \mathbf{v} under the Gaussian distribution. On the other hand, this simple modification ensures smoothness along the \mathbf{v} direction for an arbitrary input $\mathbf{x}^{(1:n)}$ that makes the mollifier PTF h non-zero.

By some tedious but straightforward computation (Lemma 3.12), one can show that the size of the t -th order directional derivative of h is still roughly comparable to $\|p^{[t],\mathbf{v}}(\mathbf{x}^{(1:n)})\|_F / |p(\mathbf{x}^{(1:n)})|$. We can then do a simple case analysis. If the ratio is large, then the term $\rho \left(\frac{d^{\Theta(t)} \|p^{[t],\mathbf{v}}(\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)})\|_F^2}{p(\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)})^2} \right)$ ensures that the entire mollified PTF evaluates to 0. Otherwise, we can conclude that the t -th order directional derivative $D_{i,\mathbf{v}}^t h$ will be roughly on the order of $d^{-\Theta(t)}$ (cf. Lemma 3.10). Due to the moment matching condition, when one computes the difference $\mathbf{E}_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [h(\mathbf{x}^{(1:i-1)}, \mathbf{z}, \mathbf{y}^{(i+1:n)})] - \mathbf{E}_{\mathbf{z} \sim \mathcal{M}_{A,\mathbf{v}}} [h(\mathbf{x}^{(1:i-1)}, \mathbf{z}, \mathbf{y}^{(i+1:n)})]$ via the directional Taylor expansion, the first degree m terms in ξ cancel exactly, leaving us with the dominating term

$$D_{i,\mathbf{v}}^{m+1} h(\mathbf{x}^{(1:i-1)}, \bar{\mathbf{z}}, \mathbf{y}^{(i+1:n)}) \left(\mathbf{E}_{\xi \sim A} [\xi^{m+1}] - \mathbf{E}_{\xi \sim \mathcal{N}(0,1)} [\xi^{m+1}] \right) ,$$

which can be appropriately bounded by $d^{-\Theta(m)}$.⁶ As long as we have $d^{-\Theta(m)} \ll o(1/n)$, the replacement step will go through, and this concludes the sketch of our proof of [Theorem 1.6](#).

2 Notation

2.1 Basic Notation

We use \mathbb{Z}_+ for the set of positive integers and $[n]$ to denote $\{1, \dots, n\}$. We use bold lowercase letters for vectors and bold uppercase letters for tensors. We use $\mathbf{1}_d$ for the d -dimensional all-one vector, $\mathbf{0}_d$ for the d -dimensional all-zero vector, and \mathbf{I}_d for the $d \times d$ identity matrix. When the dimension is clear from the context, we will drop the subscript. For a set S , we use $\mathcal{U}(S)$ to denote the uniform distribution on S . Given a distribution D in \mathbb{R}^d , we write $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim D$ to denote n i.i.d. samples from D . For the sake of saving space, we often write $\mathbf{x}^{(1:n)}$ to denote the sequence of vectors $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$. We write $a \gg b$ to denote that $a \geq Cb$ for some sufficiently large constant $C > 0$.

2.2 Tensor Notation

We frequently use tensors in $(\mathbb{R}^n)^{\otimes k}$. For some $\mathbf{A} \in (\mathbb{R}^n)^{\otimes k}$, we denote by $\mathbf{A}_{i_1, \dots, i_k}$ the entry in \mathbf{A} indexed by $i_1, \dots, i_k \in [n]$. For two tensors $\mathbf{A}, \mathbf{B} \in (\mathbb{R}^n)^{\otimes k}$, we define the inner product (or dot product) between them as $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i_1, \dots, i_k \in [n]} \mathbf{A}_{i_1, \dots, i_k} \mathbf{B}_{i_1, \dots, i_k}$. We use \mathbf{A}^\flat to denote the flattened version of \mathbf{A} , i.e., the vector in \mathbb{R}^{n^k} obtained by stacking all entries of \mathbf{A} into a single vector in lexicographic order. We define the Frobenius norm of tensor \mathbf{A} to be $\|\mathbf{A}\|_F := \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle}$. For two tensors $\mathbf{A} \in (\mathbb{R}^n)^{\otimes k}$ and $\mathbf{B} \in (\mathbb{R}^n)^{\otimes \ell}$ with $k > \ell$ we write \mathbf{AB} to denote the tensor in $(\mathbb{R}^n)^{\otimes(k-\ell)}$ defined as follows:

$$(\mathbf{AB})_{i_{\ell+1}, \dots, i_k} := \sum_{i_1, \dots, i_\ell \in [n]} \mathbf{A}_{i_1, \dots, i_\ell} \mathbf{B}_{i_1, \dots, i_\ell}.$$

Note that for $\mathbf{A}, \mathbf{B} \in (\mathbb{R}^n)^{\otimes k}$ it holds that \mathbf{AB} defined as above is the same as the inner product $\langle \mathbf{A}, \mathbf{B} \rangle$. Moreover, using the tensor product notation, if $\mathbf{e}(1), \dots, \mathbf{e}(n) \in \mathbb{R}^n$ denote the standard basis vectors, then $\mathbf{T}\mathbf{e}(i)$ is simply the tensor \mathbf{T} restricted on the first index being equal to i . Similarly $\mathbf{T}(\mathbf{e}(i_1) \otimes \dots \otimes \mathbf{e}(i_k))$ selects the sub-tensor with the first k indices being i_1, \dots, i_k .

2.3 Derivative Notation

We write $p : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ to denote a multi-variable function with n variables, where each variable is a d -dimensional vector. We write $\nabla_i p$ to denote the vector of partial derivatives with respect to the i -th argument, i.e., $\nabla_i p : \mathbb{R}^{n \times d} \mapsto \mathbb{R}^d$ and $(\nabla_i p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}))_j = \frac{\partial}{\partial \mathbf{x}_j^{(i)}} p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$. Without the subscript, $\nabla p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$ denotes the usual gradient vector that takes derivatives with respect to all arguments at the same time, i.e., $\nabla p : \mathbb{R}^{n \times d} \mapsto \mathbb{R}^{n \times d}$ and $(\nabla p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}))_{i,j} = \frac{\partial}{\partial \mathbf{x}_j^{(i)}} p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$.

We also define the t -th order derivative tensor as follows:

$$\left(\nabla^t p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) \right)_{i_1, j_1, \dots, i_t, j_t} = \frac{\partial}{\partial \mathbf{x}_{j_t}^{(i_t)}} \cdots \frac{\partial}{\partial \mathbf{x}_{j_1}^{(i_1)}} p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) \quad (11)$$

⁶Technically, for the bound to hold, we will need to assume that the support of A is contained within $[d^{-c}, d^c]$ for some sufficiently small constant c . To circumvent the issue, we instead show the replacement step for the truncated distribution \bar{A} , and then relate the expected value of h under $\xi \sim \bar{A}$ back to the one under $\xi \sim A$ using the fact that h is bounded between $[0, 1]$ (cf. [Lemma 3.6](#)).

We will also make use of directional partial derivatives $D_{i,\mathbf{v}}p : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ defined as:

$$D_{i,\mathbf{v}}p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) = \mathbf{v}^\top \nabla_i p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$$

More generally, we denote by $p^{[k],\mathbf{v}} : \mathbb{R}^{n \times d} \mapsto (\mathbb{R}^n)^{\otimes k}$ the following directional derivative tensors:

$$p_{i_1, \dots, i_k}^{[k],\mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) = D_{i_1,\mathbf{v}} D_{i_2,\mathbf{v}} \cdots D_{i_k,\mathbf{v}} p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}).$$

When the unit vector \mathbf{v} is clear from the context (typically inside proofs), we will just write $p^{[k]}$.

Recall the convention of multiplying two tensors of different dimensions in [Section 2.2](#). Following that convention, we also note the following equivalence, which we will use sparingly in the paper:

$$D_{i,\mathbf{v}}p^{[k],\mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) = p^{[k+1],\mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) \mathbf{e}(i),$$

where $\mathbf{e}(i)$ is the i -th standard basis vector.

3 Proof of [Theorem 1.6](#)

3.1 Decay of Derivatives Restricted to a Random Direction

In this subsection, we show that the t -th order directional partial derivatives of a low-degree polynomial $p : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ along a random direction \mathbf{v} can be bounded from above by an exponentially decreasing function in t . The formal statement is given below.

Lemma 3.1 (Derivative Decay). *For any $k \in \mathbb{Z}_+$ and $\varepsilon \in (0, 1)$, if $p : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}$ is a degree- k polynomial and $\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})$, then the following holds with probability 0.99 over the randomness of \mathbf{v} :*

$$\Pr_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\forall t \in [k] : \|p^{[t],\mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F \leq (k/\varepsilon)^{4t} d^{-t/4} |p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})| \right] \geq 1 - \varepsilon. \quad (12)$$

Our starting point is the so called strong anti-concentration properties of Gaussian distributions used frequently in the PRG literature. Specifically, it states that the output of a polynomial is usually not too small compared to its derivative. As a simple corollary of this property, we obtain another interesting property of polynomials that come in handy in showing [Lemma 3.1](#). That is, the sizes of the higher order derivative tensor of a polynomial grow rather slowly (notably, the growth rate is independent of the input dimension).

Fact 3.2 (Slow Growth of Derivatives; Lemma 1.6 in [\[KM22\]](#)). *For any $k \in \mathbb{Z}_+$ and $\varepsilon \in (0, 1)$, if $p : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}$ is a degree- k polynomial, it holds that*

$$\Pr_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}} \left[\exists t \in [k] : \|\nabla^t p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F > (k^{3t}/\varepsilon^t) |p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})| \right] \leq O(\varepsilon). \quad (13)$$

As argued in [Section 1.2](#), a direct application of the strong anti-concentration property is not enough for our purposes. Instead, we need to leverage the specific structure of the distribution in the NGCA problem: the distribution $\mathcal{M}_{A,\mathbf{v}}$ is non-Gaussian only along a single randomly chosen direction \mathbf{v} , and is the same as the standard Gaussian in every orthogonal direction. Our main result in this subsection is that if we fix a low-degree polynomial p and take \mathbf{v} to be some random unit vector, the directional derivative tensor $p^{[k],\mathbf{v}}$ will instead be shrinking with high probability. Intuitively, this is because $p^{[t],\mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$ is just a *random* sub-part of the entire tensor $\nabla^t p(\mathbf{x}_1, \dots, \mathbf{x}_n)$ that appeared in [Fact 3.2](#).

Proof of Lemma 3.1. Fix some arbitrary $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \in \mathbb{R}^d$. For convenience, define the tensor $\mathbf{M} := \nabla^t p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$, and $\mathbf{N} := p^{[t], \mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$. Recall that the full degree- t gradient tensor \mathbf{M} lives in the space $(\mathbb{R}^{n \times d})^{\otimes t}$. We can therefore label each entry of \mathbf{M} by some indices $i_1, j_1, \dots, i_t, j_t$, where $i_1, \dots, i_t \in [n]$ and $j_1, \dots, j_t \in [d]$. On the other hand, the projected derivative tensor \mathbf{N} lives in the space $(\mathbb{R}^n)^{\otimes t}$. We can therefore label each entry of \mathbf{N} by some indices i_1, \dots, i_t , where $i_1, \dots, i_t \in [n]$. One can then verify using the definitions from Section 2.3 that the entries of \mathbf{N} and \mathbf{M} satisfy the following relationship:

$$\begin{aligned}\mathbf{N}_{i_1, \dots, i_t} &= \sum_{j_1 \in [d]} \mathbf{v}_{j_1} \frac{\partial}{\partial \mathbf{x}_{j_1}^{(i_1)}} \sum_{j_2 \in [d]} \mathbf{v}_{j_2} \frac{\partial}{\partial \mathbf{x}_{j_2}^{(i_2)}} \cdots \sum_{j_t \in [d]} \mathbf{v}_{j_t} \frac{\partial}{\partial \mathbf{x}_{j_t}^{(i_t)}} p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) \\ &= \sum_{j_1, \dots, j_t \in [d]} \mathbf{v}_{j_1} \cdots \mathbf{v}_{j_t} \frac{\partial}{\partial \mathbf{x}_{j_1}^{(i_1)}} \cdots \frac{\partial}{\partial \mathbf{x}_{j_t}^{(i_t)}} p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) \\ &= \sum_{j_1, \dots, j_t \in [d]} \mathbf{v}_{j_1} \cdots \mathbf{v}_{j_t} \mathbf{M}_{i_1, j_1, \dots, i_t, j_t}.\end{aligned}$$

Let $\mathbf{v}^{\otimes t}$ be the t -fold tensor product of \mathbf{v} with itself, i.e., the tensor with entries:

$$(\mathbf{v}^{\otimes t})_{j_1, \dots, j_t} := \mathbf{v}_{j_1} \cdots \mathbf{v}_{j_t},$$

and $\mathbf{M}^{(i_1, \dots, i_t)} \in (\mathbb{R}^d)^{\otimes t}$ be the sub-tensor of \mathbf{M} of the form:

$$\mathbf{M}_{j_1, \dots, j_t}^{(i_1, \dots, i_t)} := \mathbf{M}_{i_1, j_1, \dots, i_t, j_t}. \quad (14)$$

We can then write

$$\begin{aligned}\mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} [\|\mathbf{N}\|_F^2] &= \sum_{i_1, \dots, i_t \in [n]} \mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} \left[\left(\sum_{j_1, \dots, j_t} \mathbf{v}_{j_1} \cdots \mathbf{v}_{j_t} \mathbf{M}_{i_1, j_1, \dots, i_t, j_t} \right)^2 \right] \\ &= \sum_{i_1, \dots, i_t \in [n]} \mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} \left[\langle \mathbf{v}^{\otimes t}, \mathbf{M}^{(i_1, \dots, i_t)} \rangle^2 \right] \\ &= \sum_{i_1, \dots, i_t \in [n]} \left(\mathbf{M}^{(i_1, \dots, i_t)} \right)^{\flat, \top} \mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} \left[(\mathbf{v}^{\otimes t})^\flat (\mathbf{v}^{\otimes t})^{\flat, \top} \right] \left(\mathbf{M}^{(i_1, \dots, i_t)} \right)^\flat. \quad (15)\end{aligned}$$

We claim the following bound on the matrix $\mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} \left[(\mathbf{v}^{\otimes t})^\flat (\mathbf{v}^{\otimes t})^{\flat, \top} \right] \in \mathbb{R}^{d^t \times d^t}$ that appeared earlier:

Claim 3.3. Define $\mathbf{W}(\mathbf{v}) := \mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} \left[(\mathbf{v}^{\otimes t})^\flat (\mathbf{v}^{\otimes t})^{\flat, \top} \right] \in \mathbb{R}^{d^t \times d^t}$. Then we have

$$\|\mathbf{W}(\mathbf{v})\|_F \leq \left(\frac{2t}{e\sqrt{d}} \right)^t.$$

We defer the proof of the above claim to the end of this subsection. Before that, we show how to conclude the proof of Lemma 3.1 using the claim. Combining Claim 3.3 and eq. (15) gives the following (recall that $\mathbf{M}^{(i_1, \dots, i_t)}$ denotes the tensor from Equation (14)):

$$\begin{aligned}\mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} [\|\mathbf{N}\|_F^2] &\leq \sum_{i_1, \dots, i_t \in [n]} \|\mathbf{M}^{(i_1, \dots, i_t)}\|_F^2 \|\mathbf{W}(\mathbf{v})\|_F \\ &\leq (2t/e)^t d^{-t/2} \sum_{i_1, \dots, i_t \in [n]} \|\mathbf{M}^{(i_1, \dots, i_t)}\|_F^2 \quad (\text{Claim 3.3}) \\ &= (2t/e)^t d^{-t/2} \|\mathbf{M}\|_F^2.\end{aligned}$$

By Jensen's inequality, we have that

$$\mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} [\|\mathbf{N}\|_F] \leq (2t/e)^{t/2} d^{-t/4} \|\mathbf{M}\|_F. \quad (16)$$

Recall that at the beginning of the proof we fixed some arbitrary $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ and defined the quantities $\mathbf{N} := p^{[t], \mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$, $\mathbf{M} := \nabla^t p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$ accordingly. Therefore, Equation (16) implies that

$$\mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} \left[\|p^{[t], \mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F \right] \leq (2t/e)^{t/2} d^{-t/4} \|\nabla^t p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F. \quad (17)$$

Define the event $\mathcal{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}} = \{\|p^{[t], \mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F \neq 0, \nabla^t p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) \neq 0\}$ with respect to a sequence of points $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$. Then we have that

$$\begin{aligned} & \mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} \left[\mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\frac{\|p^{[t], \mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F}{\|\nabla^t p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F} \middle| \mathcal{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}} \right] \right] \\ &= \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} \left[\frac{\|p^{[t], \mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F}{\|\nabla^t p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F} \middle| \mathcal{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}} \right] \right] \\ &\leq (2t/e)^{t/2} d^{-t/4}. \quad (\text{using Equation (17) and the definition of } \mathcal{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}}) \end{aligned}$$

Applying Markov's inequality on \mathbf{v} then gives that

$$\mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\frac{\|p^{[t], \mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F}{\|\nabla^t p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F} \middle| \mathcal{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}} \right] \leq O(1) (2t/e)^{t/2} d^{-t/4} \quad (18)$$

with high constant probability. In the remaining analysis, we condition on some \mathbf{v} such that Equation (18) holds. Note that in the case that the event $\mathcal{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}}$ does not hold then we still have $\|p^{[t], \mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F \leq O((2t/e)^{t/2} d^{-t/4}) \|\nabla^t p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F$ as both sides of the inequality are zero. We thus have that

$$\begin{aligned} & \Pr_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\|p^{[t], \mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F \leq O(\varepsilon^{-1} (2t/e)^{t/2} d^{-t/4}) \|\nabla^t p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F \right] \\ &\geq \Pr_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\|p^{[t], \mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F \leq O(\varepsilon^{-1} (2t/e)^{t/2} d^{-t/4}) \|\nabla^t p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F \middle| \mathcal{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}} \right] \\ &\geq 1 - \varepsilon/2, \end{aligned}$$

where we used that the event $\mathcal{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}}$ happens with probability 1 (this is because the complement of the event amounts to $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ being the exactly equal to the roots of a polynomial), and Markov's inequality.

By a union bound over $t \in [k]$, we further get that

$$\begin{aligned} & \Pr_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\forall t \in [k] : \|p^{[t], \mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F \leq O(\varepsilon^{-1} k (2t/e)^{t/2} d^{-t/4}) \|\nabla^t p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F \right] \\ &\geq 1 - \varepsilon/2. \end{aligned}$$

Combining this with Equation (13) and the union bound then gives that

$$\begin{aligned} & \Pr_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\forall t \in [k] : \|p^{[t], \mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F \leq O(\varepsilon^{-1} k (2t/e)^{t/2} d^{-t/4}) k^{3t} (2/\varepsilon)^t |p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})| \right] \\ &\geq 1 - \varepsilon. \end{aligned}$$

We can finally simplify the expression $O\left(\varepsilon^{-1}k(2t/e)^{t/2}d^{-t/4}\right)$ that appears above as follows:

$$O\left(\varepsilon^{-1}k(2t/e)^{t/2}d^{-t/4}\right)k^{3t}(2/\varepsilon)^t \leq (k/\varepsilon)^4 t d^{-t/4}$$

which concludes the proof of [Lemma 3.1](#). \square

We conclude this section by showing [Claim 3.3](#):

Proof of Claim 3.3. To show that, we will first relate the expected value of $\mathbf{W}(\mathbf{v})$ under $\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})$ to that under $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. In particular, denote by $g : \mathbb{R} \mapsto \mathbb{R}$ the probability density function of the random variable $\|\mathbf{v}\|_2$, where $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then we have that

$$\begin{aligned} \left\| \mathbf{E}_{\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\mathbf{V}(\mathbf{v})^\flat \mathbf{V}(\mathbf{v})^{\flat, \top}] \right\|_F &= \left\| \int_0^{+\infty} \mathbf{E}_{\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\mathbf{V}(\mathbf{v})^\flat \mathbf{V}(\mathbf{v})^{\flat, \top} \mid \|\mathbf{v}\|_2 = b] g(\mathbf{v}) db \right\|_F \\ &= \left\| \mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} [\mathbf{V}(\mathbf{v})^\flat \mathbf{V}(\mathbf{v})^{\flat, \top}] \int_0^{+\infty} b^{2t} g(\mathbf{v}) db \right\|_F \\ &\geq \left\| \mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} [\mathbf{V}(\mathbf{v})^\flat \mathbf{V}(\mathbf{v})^{\flat, \top}] \right\|_F d^t = \|\mathbf{W}(\mathbf{v})\|_F d^t \end{aligned}$$

where the last line used that $\int_0^{+\infty} b^{2t} g(b) db = \mathbf{E}_{\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\|\mathbf{v}\|^{2t}] \geq \mathbf{E}_{\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\|\mathbf{v}\|^2]^t = d^t$. Rearranging this gives the following upper bound on the Frobenius norm of $\mathbf{W}(\mathbf{v})$:

$$\begin{aligned} \|\mathbf{W}(\mathbf{v})\|_F &= \left\| \mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} [\mathbf{V}(\mathbf{v})^\flat \mathbf{V}(\mathbf{v})^{\flat, \top}] \right\|_F \leq d^{-t} \left\| \mathbf{E}_{\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\mathbf{V}(\mathbf{v})^\flat \mathbf{V}(\mathbf{v})^{\flat, \top}] \right\|_F \\ &= d^{-t} \sqrt{\sum_{j_1, \dots, j_{2t} \in [d]} \left(\mathbf{E}_{\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\mathbf{v}_{j_1} \cdots \mathbf{v}_{j_{2t}}] \right)^2}. \quad (19) \end{aligned}$$

We can further bound the quantity $\mathbf{E}_{\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\mathbf{v}_{j_1} \cdots \mathbf{v}_{j_{2t}}]$ that appears in the right hand side above using Iserrlis' theorem ([Fact A.4](#)) as follows (P_k^2 below denotes the set of all matchings among $\{1, \dots, k\}$):

$$\begin{aligned} \sum_{j_1, \dots, j_{2t} \in [d]} (\mathbf{E} [\mathbf{v}_{j_1} \cdots \mathbf{v}_{j_{2t}}])^2 &= \sum_{j_1, \dots, j_{2t} \in [d]} \left(\sum_{p \in P_{2t}^2} \prod_{\{k, \ell\} \in p} \mathbf{E} [\mathbf{v}_{j_k} \mathbf{v}_{j_\ell}] \right)^2 \\ &\leq |P_{2t}^2| \sum_{j_1, \dots, j_{2t} \in [d]} \sum_{p \in P_{2t}^2} \prod_{\{k, \ell\} \in p} (\mathbf{E} [\mathbf{v}_{j_k} \mathbf{v}_{j_\ell}])^2 \\ &= |P_{2t}^2| \sum_{p \in P_{2t}^2} \prod_{\{k, \ell\} \in p} \sum_{j_k, j_\ell \in [d]} (\mathbf{E} [\mathbf{v}_{j_k} \mathbf{v}_{j_\ell}])^2 \\ &= |P_{2t}^2| \sum_{p \in P_{2t}^2} d^t \\ &= ((2t-1)!!)^2 d^t \leq (2t/e)^{2t} d^t, \quad (20) \end{aligned}$$

where the first line is an application of Iserrlis' theorem ([Fact A.4](#)), the second line uses the inequality $2ab \leq a^2 + b^2$, the fourth line uses that $\mathbf{E} [\mathbf{v}_{j_k} \mathbf{v}_{j_\ell}] = \mathbf{1}(j_k = j_\ell)$ and $|p|$ (the number of pairs within

the matching) is t , and the last line uses that the number of all possible matchings over $[2t]$ is $(2t - 1)!! < (2t)!! = 2^t t! < 2^t (t/e)^t = (2t/e)^t$. Combining [Equation \(19\)](#) and [Equation \(20\)](#) then gives that $\|\mathbf{W}(\mathbf{v})\|_F \leq d^{-t/2} (2t/e)^t$, concluding the proof of [Claim 3.3](#). \square

3.2 Framework: Mollification, Sandwiching, and Hybrid Argument

In this subsection, we lay out the high level proof strategy for our main theorem based on the ideas of mollification and the hybrid argument.

To begin with, we need a smooth function $\rho : \mathbb{R} \mapsto [0, 1]$ satisfying the following conditions:

$$\rho(x) = 1 \text{ if } |x| < 1, \rho(x) = 0 \text{ if } |x| \geq 3, \|\rho^{(t)}(x)\|_\infty \leq O(t^t). \quad (21)$$

There are standard ways to construct such a function, deferred to [Lemma A.5](#) in [Section A.2](#). We then use it to define the following mollifier function g :

$$g(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) = \prod_{t=1}^k \rho \left(\frac{d^{c_g t} \|p^{[t]}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F^2}{p^2(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})} \right), \quad (22)$$

where $c_g \in (0, 1/2)$ is some constant that we will specify later. Intuitively, g is constructed such that if some points $\mathbf{x}^{(1:n)}$ satisfy the derivative decay condition from [Lemma 3.1](#), then $g(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$ should evaluate to 1. Conversely, if $g(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$ evaluates to 1, we can infer from its definition that the weaker derivative decay condition $\|p^{[t]}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F^2 \leq 3d^{-c_g t} p^2(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$ must hold for the input points.

Finally, the mollified version of the PTF is the following function:

$$h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) := \text{sign}(p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})) g(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}). \quad (23)$$

Thanks to [Lemma 3.1](#) and our construction of g , we can show that $h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$ is a good approximation of $\text{sign}(p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}))$ under the Gaussian distribution. Moreover, since g is at most 1, we note that $h(\cdot)$ is bounded from above by $\text{sign}(p(\cdot))$ pointwise. Combining these two observations with a sandwiching argument allows us to show the following: if $\mathcal{M}_{A,\mathbf{v}}$ fools the mollified PTF h with respect to the Gaussian distribution, then $\mathcal{M}_{A,\mathbf{v}}$ also fools the original PTF.

Lemma 3.4 (Sandwiching). *Let $p : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ be a degree- k polynomial, $\mathbf{v} \in \mathbb{R}^d$ be some vector satisfying [Equation \(12\)](#) with $\varepsilon = 0.05$, and $h : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ be the mollified PTF of p defined as in [Equation \(23\)](#). Assume that $(k/0.05) < d^{(1/4 - c_g/2)/4}$, where c_g is the parameter used in [Equation \(22\)](#). The following statement holds: If*

$$\left| \mathbf{E}_{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \sim \mathcal{M}_{A,\mathbf{v}}} [h(\mathbf{y}^{(1:n)})] - \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [h(\mathbf{x}^{(1:n)})] \right| \leq \delta, \quad (24)$$

then it holds that

$$\left| \mathbf{E}_{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \sim \mathcal{M}_{A,\mathbf{v}}} [\text{sign}(p(\mathbf{y}^{(1:n)}))] - \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\text{sign}(p(\mathbf{x}^{(1:n)}))] \right| \leq \delta + 0.05. \quad (25)$$

Proof. For $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \in \mathbb{R}^d$ define the event

$$\mathcal{E}_\mathbf{v}(\mathbf{x}^{(1:n)}) = \left\{ \forall t \in [k] \quad \|p^{[t], \mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F \leq d^{-c_g t/2} |p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})| \right\}.$$

Fix $\varepsilon = 0.05$. Recall that we assume \mathbf{v} is chosen such that the derivative decay condition in [Equation \(12\)](#) holds. For convenience, we restate the condition below.

$$\Pr_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\forall t \in [k] \quad \|p^{[t], \mathbf{v}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})\|_F \leq (k/\varepsilon)^{4t} d^{-t/4} |p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})| \right] \geq 1 - \varepsilon. \quad (26)$$

Using our assumption $k/\varepsilon = (k/0.05) < d^{(1/4 - c_g/2)/4}$, we have that $(k/\varepsilon)^{4t} d^{-t/4} < d^{-c_g t/2}$. Thus a simplified version of [Equation \(26\)](#) holds, and in particular implies that:

$$\Pr_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\mathcal{E}_{\mathbf{v}}(\mathbf{x}^{(1:n)}) \right] \geq 1 - \varepsilon. \quad (27)$$

We will first show that $|\mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[h(\mathbf{x}^{(1:n)})] - \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[\text{sign}(p(\mathbf{x}^{(1:n)}))]| \leq \varepsilon$ for any degree- k polynomial $p : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$. Then, we will use that to show [Equation \(25\)](#). To see the first claim, using the definition of $g(\mathbf{x}^{(1:n)})$ ([Equation \(22\)](#)) and [Equation \(27\)](#) we have that

$$\left| \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[h(\mathbf{x}^{(1:n)})] - \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[\text{sign}(p(\mathbf{x}^{(1:n)}))] \right| \quad (28)$$

$$= \left| \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[(g(\mathbf{x}^{(1:n)}) - 1)\text{sign}(p(\mathbf{x}^{(1:n)}))] \right|$$

$$= \left| \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[(g(\mathbf{x}^{(1:n)}) - 1)\text{sign}(p(\mathbf{x}^{(1:n)})) \cdot \mathbf{1}(\mathcal{E}_{\mathbf{v}}^c(\mathbf{x}^{(1:n)}))] \right| \quad (29)$$

$$\leq \Pr_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\mathcal{E}_{\mathbf{v}}^c(\mathbf{x}^{(1:n)}) \right] \leq \varepsilon, \quad (30)$$

where [Equation \(29\)](#) used that $g(\mathbf{x}^{(1:n)})$ can be different than 1 only under the complement of the event $\mathcal{E}_{\mathbf{v}}(\mathbf{x}^{(1:n)})$.

We can now show [Equation \(25\)](#), which states that the two expectations (under $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathcal{M}_{A, \mathbf{v}}$) of $\text{sign}(p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}))$ are close to each other in absolute value. We will show both a lower bound and an upper bound on their difference, which together yield the bound in absolute value. We start with the lower bound. Using that $g(\mathbf{x}^{(1:n)}) \leq 1$ and $\text{sign}(p(\mathbf{x}^{(1:n)})) \in \{0, 1\}$, we have the pointwise relationship $\text{sign}(p(\mathbf{x}^{(1:n)})) \geq \text{sign}(p(\mathbf{x}^{(1:n)})) g(\mathbf{x}^{(1:n)})$. In particular, this implies that:

$$\begin{aligned} & \mathbf{E}_{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \sim \mathcal{M}_{A, \mathbf{v}}} [\text{sign}(p(\mathbf{y}^{(1:n)}))] \\ & \geq \mathbf{E}_{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \sim \mathcal{M}_{A, \mathbf{v}}} [(\text{sign}(p(\mathbf{y}^{(1:n)}))) g(\mathbf{x}^{(1:n)})] \\ & \geq \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\text{sign}(p(\mathbf{x}^{(1:n)})) g(\mathbf{x}^{(1:n)})] - \delta \quad (\text{by } \text{Equation (24)}) \\ & \geq \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\text{sign}(p(\mathbf{x}^{(1:n)}))] - \delta - \varepsilon. \quad (\text{by } \text{Equation (30)}) \end{aligned}$$

We can prove the other direction $\mathbf{E}[\text{sign}(p(\mathbf{y}^{(1:n)}))] \leq \mathbf{E}[\text{sign}(p(\mathbf{x}^{(1:n)}))] + \delta + \varepsilon$ by repeating the same argument with $-p$ in place of p . This concludes the proof of [Lemma 3.4](#). \square

Given the above sandwiching lemma, it then suffices for us to bound from above the difference $|\mathbf{E}_{\mathbf{y}^{(1:n)} \sim \mathcal{M}_{A, \mathbf{v}}}[h(\mathbf{y}^{(1:n)})] - \mathbf{E}_{\mathbf{x}^{(1:n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[h(\mathbf{x}^{(1:n)})]|$. We will show this via the hybrid argument. In particular, let $\mathbf{x}^{(1:n)}$ be i.i.d. samples from the standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{y}^{(1:n)}$ be i.i.d. samples from the hidden direction distribution $\mathcal{M}_{A, \mathbf{v}}$. In the i -th replacement step of the hybrid argument, we compare the expected values of $h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \mathbf{y}^{(i+1)}, \dots, \mathbf{y}^{(n)})$ and $h(\mathbf{x}^{(1)}, \dots, \mathbf{y}^{(i)}, \mathbf{y}^{(i+1)}, \dots, \mathbf{y}^{(n)})$. We show that the difference between the expected values is on the order of $d^{-c_g m/2}$.

Proposition 3.5 (Replacement Step). *For any $c \in (0, 1/4)$, $d, m, k \in \mathbb{Z}_+$ such that m is even, and $d > \max(m^{C/c}, k^{C/c})$ for some sufficiently large constant C ,*

if $p : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ is a degree- k polynomial, $\mathbf{v} \in \mathbb{R}^d$ is a unit vector satisfying Equation (12) with $\varepsilon = 0.05$, A is a one-dimensional distribution that matches the first m moments with $\mathcal{N}(0, 1)$, and $h : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ is the mollified PTF from Equation (23), then the following holds: For every $i \in [n]$

$$\left| \mathbf{E} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \mathbf{y}^{(i+1)}, \dots, \mathbf{y}^{(n)}) \right] - \mathbf{E} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \mathbf{y}^{(i+1)}, \dots, \mathbf{y}^{(n)}) \right] \right| \leq d^{-c_g m/2 + cm}, \quad (31)$$

where c_g is the parameter used in Equation (23), $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{M}_{A, \mathbf{v}}$, and $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Our main result follows immediately from Lemma 3.4 and Proposition 3.5.

Proof of Theorem 1.6. Let c^* be the constant parameter in the statement of Theorem 1.6. Fix $\varepsilon = 0.05$, $c = c^*/2$, and $c_g = 2(1/4 - c^*/2)$. By Lemma 3.1, we have that a randomly chosen \mathbf{v} satisfies Equation (12) with high probability 0.99. We will condition on such a \mathbf{v} in the rest of the proof. First, combining Proposition 3.5 (the proposition is applicable because of our assumptions $d > \max(k^{C^*/c^*}, m^{C^*/c^*})$ in the statement of Theorem 1.6 and $c = c^*/2$) for each position $i \in [n]$ with the triangle inequality yields that

$$\left| \mathbf{E}_{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \sim \mathcal{M}_{A, \mathbf{v}}} \left[h(\mathbf{y}^{(1:n)}) \right] - \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[h(\mathbf{x}^{(1:n)}) \right] \right| \leq nd^{-c_g m/2 + cm}.$$

Since we have $c_g = 2(1/4 - c^*/2)$ and $c = c^*/2$, the right hand side can be further bounded from above by $nd^{-(1/4 - c^*)m}$, which is at most 0.05 by our assumption that $n \ll d^{(1/4 - c^*)m}$. We can then apply the sandwiching lemma (Lemma 3.4) with $\delta = 0.05$.⁷ This yields that

$$\left| \mathbf{E}_{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \sim \mathcal{M}_{A, \mathbf{v}}} \left[\text{sign}(p(\mathbf{y}^{(1:n)})) \right] - \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}} \left[\text{sign}(p(\mathbf{x}^{(1:n)})) \right] \right| \leq \delta + \varepsilon = 0.1. \quad (32)$$

So far we have shown that Equation (32) holds with probability 0.99. From this, we can complete the proof as follows. First, denote by \mathcal{E} the event in Equation (32). We have the following by Jensen's inequality:

$$\left| \mathbf{E}_{\substack{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1}) \\ \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{M}_{A, \mathbf{v}}}} \left[\text{sign}(p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})) \right] - \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\text{sign}(p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})) \right] \right| \quad (33)$$

$$\leq \mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} \left[\left| \mathbf{E}_{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \sim \mathcal{M}_{A, \mathbf{v}}} \left[\text{sign}(p(\mathbf{y}^{(1:n)})) \right] - \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\text{sign}(p(\mathbf{x}^{(1:n)})) \right] \right| \right]. \quad (34)$$

Now let the random variable

$$Z = \left| \mathbf{E}_{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \sim \mathcal{M}_{A, \mathbf{v}}} \left[\text{sign}(p(\mathbf{y}^{(1:n)})) \right] - \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\text{sign}(p(\mathbf{x}^{(1:n)})) \right] \right|$$

for brevity. We can further bound the RHS in Equation (34) as follows:

$$\begin{aligned} \mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} [Z] &= \mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} [Z \mathbf{1}(\mathcal{E})] + \mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} [Z \mathbf{1}(\mathcal{E}^c)] \\ &\leq 0.1 + \mathbf{E}_{\mathbf{v} \sim \mathcal{U}(\mathcal{S}^{d-1})} [\mathbf{1}(\mathcal{E}^c)] \\ &\leq 0.1 + 0.01 = 1.01, \end{aligned}$$

⁷The lemma is applicable since the assumption $k < d^{c^*/C^*}$ from Theorem 1.6 implies that $(k/0.05) < d^{(1/4 - c_g/2)/4}$ as long as $c_g = 2(1/4 - c^*/2)$ and C^* is sufficiently large.

where we used that under the event \mathcal{E} we have $Z \leq 0.1$, we also used that Z is always at most 1 since it is a difference of PTFs that take values in $\{0, 1\}$ and we used that \mathcal{E} happens with probability at least 0.99 over the choice of \mathbf{v} .

This concludes the proof of [Theorem 1.6](#). \square

3.3 Proof of [Proposition 3.5](#)

In the rest of this section, we focus on establishing a single replacement step by proving [Proposition 3.5](#). This subsection is organized as follows. In [Section 3.3.1](#), we introduce a truncation procedure that reduces the problem to the case where the non-Gaussian component A has bounded support, which will be helpful later on, and argue that the truncation will not significantly change the expectation of the mollified PTF. In [Section 3.3.2](#), we show that if the derivative of the polynomial is small at any point in the truncated interval, it is small throughout the entire interval. In [Section 3.3.3](#), we perform a Taylor expansion of the mollified PTF, derive expressions for the derivatives in the higher-order terms, and upper bound them using the results from [Section 3.3.2](#). Finally, [Section 3.3.4](#) concludes the proof of [Proposition 3.5](#) by carefully combining the results from the previous sections.

3.3.1 Domain Truncation

Let A be a distribution on \mathbb{R} which matches the first m moments with $\mathcal{N}(0, 1)$. Note that the hidden direction distribution $\mathcal{M}_{A, \mathbf{v}}$ from [Definition 1.4](#) can be viewed of as the sum $\bar{\mathbf{y}} + \xi \mathbf{v}$, where $\bar{\mathbf{y}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I} - \mathbf{v}\mathbf{v}^\top)$ and $\xi \sim A$. Hence, [Equation \(31\)](#) can be alternatively written as

$$\left| \mathbf{E} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)}) \right] - \mathbf{E} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \bar{\mathbf{y}} + z \mathbf{v}, \dots, \mathbf{y}^{(n)}) \right] \right| \leq d^{-c_g m/2 + cm}, \quad (35)$$

where $\xi \sim A$, $z \sim \mathcal{N}(0, 1)$, $\bar{\mathbf{y}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I} - \mathbf{v}\mathbf{v}^\top)$, $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)} \sim \mathcal{M}_{A, \mathbf{v}}$, and $\mathbf{y}^{(i+1)}, \dots, \mathbf{y}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

To show [Equation \(35\)](#), we will condition on fixed values for all random variables except from ξ and z and show that

$$\left| \mathbf{E}_{\xi \sim A} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)}) \right] - \mathbf{E}_{z \sim \mathcal{N}(0, 1)} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \bar{\mathbf{y}} + z \mathbf{v}, \dots, \mathbf{y}^{(n)}) \right] \right| \leq d^{-c_g m/2 + cm}. \quad (36)$$

[Equation \(35\)](#) will then follow from [Equation \(36\)](#) via an averaging argument. It turns out that the arguments of the rest of the subsection (in particular the part that analyzes the Taylor expansion in [Section 3.3.3](#)) will be easier if ξ and z are bounded random variables. Fortunately, the two are both concentrated around 0, and we can therefore truncate them to the interval $[-d^{-c_{\text{trunc}}}, d^{c_{\text{trunc}}}]$ for some constant c_{trunc} . In particular, we argue that after the truncation we still have an approximate moment matching condition, and the mass of the non-Gaussian component A outside the truncated interval cannot be too large.

Lemma 3.6 (Domain Truncation). *For any $d, m \in \mathbb{Z}_+$, where m is even and $d \gg m$, and every $c_{\text{trunc}} > 0$, if A is a distribution on \mathbb{R} which matches the first m moments with $\mathcal{N}(0, 1)$, then for all $t \in [m-1]$ it holds that*

$$\left| \mathbf{E}_{x \sim A} [x^t \mid x \in [-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]] - \mathbf{E}_{y \sim \mathcal{N}(0, 1)} [y^t \mid y \in [-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]] \right| \leq m^{O(t)} d^{-c_{\text{trunc}}(m-t)}. \quad (37)$$

Moreover, it holds

$$\mathbf{Pr}_{x \sim A} [x \notin [-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]] \leq m^m d^{-c_{\text{trunc}} m}. \quad (38)$$

Proof of Lemma 3.6. We start with Equation (37). Let us first denote $I = [-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]$ and $p := \Pr_{x \sim A}[x \notin I]$. We can write

$$\mathbf{E}_{x \sim A}[x^t | x \in I] = \frac{\mathbf{E}_{x \sim A}[x^t \mathbb{1}(x \in I)]}{1 - p} = \frac{\mathbf{E}_{x \sim A}[x^t]}{1 - p} - \frac{\mathbf{E}_{x \sim A}[x^t \mathbb{1}(x \notin I)]}{1 - p}.$$

Rearranging, we have that

$$\left| \mathbf{E}_{x \sim A}[x^t | x \in I] - \mathbf{E}_{x \sim A}[x^t] \right| \leq p \left| \mathbf{E}_{x \sim A}[x^t | x \in I] \right| + \left| \mathbf{E}_{x \sim A}[x^t \mathbb{1}(x \notin I)] \right|. \quad (39)$$

We can upper bound p using the higher-order Chebyshev's inequality and the moment matching property of A :

$$p := \Pr_{x \sim A}[|x| > d^{c_{\text{trunc}}}] \leq \frac{\mathbf{E}_{x \sim A}[|x|^m]}{d^{c_{\text{trunc}} m}} = \frac{\mathbf{E}_{x \sim \mathcal{N}(0,1)}[|x|^m]}{d^{c_{\text{trunc}} m}} \leq \frac{m^{m/2}}{d^{c_{\text{trunc}} m}} \quad (40)$$

Using this bound on p , the first term in the RHS of Equation (39) is

$$\begin{aligned} p \left| \mathbf{E}_{x \sim A}[x^t | x \in I] \right| &\leq p \cdot \mathbf{E}_{x \sim A}[|x|^t | x \in I] \leq \frac{p}{1 - p} \mathbf{E}_{x \sim A}[|x|^t] \\ &= \frac{p}{1 - p} \mathbf{E}_{x \sim \mathcal{N}(0,1)}[|x|^t] \leq m^{O(m)} d^{-c_{\text{trunc}} m}, \end{aligned} \quad (41)$$

where we used the definition of conditional expectation, the moment matching property of A and that $d \gg m$. The second term in the RHS of Equation (39) can be upper bounded using Holder's inequality as follows:

$$\begin{aligned} \left| \mathbf{E}_{x \sim A}[x^t \mathbb{1}(x \notin I)] \right| &\leq \left(\mathbf{E}_{x \sim A}[x^m] \right)^{t/m} p^{1-t/m} \\ &= \left(\mathbf{E}_{x \sim \mathcal{N}(0,1)}[x^m] \right)^{t/m} p^{1-t/m} \\ &\leq m^{O(t)} d^{-c_{\text{trunc}}(m-t)}. \end{aligned} \quad (42)$$

Combining Equations (41) and (42) then gives that $\left| \mathbf{E}_{x \sim A}[x^t | x \in I] - \mathbf{E}_{x \sim A}[x^t] \right| \leq m^{O(t)} d^{-c_{\text{trunc}}(m-t)}$. Repeating the same steps (using $\mathcal{N}(0, 1)$ in place of A), it can be shown that

$$\left| \mathbf{E}_{y \sim \mathcal{N}(0,1)}[y^t | y \in I] - \mathbf{E}_{y \sim \mathcal{N}(0,1)}[y^t] \right| \leq m^{O(t)} d^{-c_{\text{trunc}}(m-t)}$$

as well. The rest of the proof of Equation (37) then follows from the triangle inequality and the fact that $\mathbf{E}_{y \sim \mathcal{N}(0,1)}[y^t] = \mathbf{E}_{x \sim A}[x^t]$. Finally, we note that Equation (38) has already been shown in Equation (40). This completes the proof of Lemma 3.6. \square

In the rest of the subsection, we focus in bounding the difference in expected values under the truncated distributions. In particular, our goal is to show that

$$\left| \mathbf{E}_{\xi \sim \bar{A}} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)}) \right] - \mathbf{E}_{z \sim \bar{\mathcal{N}}(0,1)} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \bar{\mathbf{y}} + z \mathbf{v}, \dots, \mathbf{y}^{(n)}) \right] \right| \leq d^{-c_g m + c/2}, \quad (43)$$

where \bar{A} and $\bar{\mathcal{N}}(0, 1)$ correspond to the distributions A and $\mathcal{N}(0, 1)$ conditioned on the domain $[-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]$ respectively.

3.3.2 Controlling Derivatives of Nearby Points

Recall that the mollifier g (Equation (22)) gives zero value to points where the derivatives of the polynomial p fail to decay at the desired rate. Fix some points $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i-1)}, \bar{\mathbf{y}}, \mathbf{y}^{(i+1)}, \dots, \mathbf{y}^{(n)} \in \mathbb{R}^d$. If it happens to be the case that for all $\xi \in [-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]$ there exists some $t \in [k]$ such that

$$\left\| p^{[t]} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 \geq 3d^{-c_g t} p^2 \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right),$$

it follows by the definition of the mollifier that the mollified PTF will be zero over the entire truncated domain, i.e., for all $\xi \in [-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]$ we will have

$$\mathbf{E}_{\xi \sim \bar{A}} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)}) \right] = \mathbf{E}_{z \sim \mathcal{N}(0,1)} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \bar{\mathbf{y}} + z \mathbf{v}, \dots, \mathbf{y}^{(n)}) \right] = 0.$$

Consequently, Equation (43) will hold trivially.

Hence, it suffices to consider the complementary case: there exists some $\xi^* \in [-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]$ such that the desired derivative decay holds for all $t \in [k]$. We formalize this complementary condition in the definition of a *well-behaved* point set below.

Definition 3.7 (Well-Behaved Point Set). *Let $p : \mathbb{R}^{d \times n} \rightarrow \mathbb{R}$ be a polynomial on n points of \mathbb{R}^d , $I \subseteq \mathbb{R}$ be an interval, and $c_g > 0$ be a parameter. Let $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i-1)}, \bar{\mathbf{y}}, \mathbf{y}^{(i+1)}, \dots, \mathbf{y}^{(n)} \in \mathbb{R}^d$. We say that these points form a well-behaved point set at position i (with respect to p , I and c_g) if there exists some $\xi^* \in I$ such that*

$$\forall t \in [k] \quad \left\| p^{[t]} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 \leq 3d^{-c_g t} \left| p \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right|^2.$$

Throughout this section, we will always use $I = [-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]$ for the interval, and c_g will be the same parameter as in the definition of the mollified PTF (Equation (22)). A key technical lemma we will prove is that if we condition on a well-behaved point set, the derivatives must also be “approximately” well-behaved for all ξ in the truncated domain.

Lemma 3.8 (Derivative Decay of Nearby Points). *Let $p : \mathbb{R}^{d \times n} \rightarrow \mathbb{R}$ be a polynomial on n points of \mathbb{R}^d , and $c_g, c_{\text{trunc}} > 0$ be parameters satisfying $c_g/2 - c_{\text{trunc}} \geq \Omega(1)$. Let $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i-1)}, \bar{\mathbf{y}}, \mathbf{y}^{(i+1)}, \dots, \mathbf{y}^{(n)} \in \mathbb{R}^d$ be a well-behaved point set at position i with respect to p , interval $I = [-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]$, and c_g . Then it holds that*

$$\forall t \in [k] \quad \left\| p^{[t]} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 \leq O(1) d^{-c_g t} p^2 \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \quad (44)$$

for all $\xi \in I$.

Proof. By Definition 3.7, there exists some point ξ^* such that

$$\forall k' \in [k] : \left\| p^{[k']} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 \leq 3d^{-c_g t} p^2 \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right). \quad (45)$$

Fix an arbitrary $\xi \in [-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]$. We claim that the following holds.

Claim 3.9. *Consider the setting and notation of Lemma 3.8. For every $k' \in \{0, 1, \dots, k\}$, the following holds:*

$$\begin{aligned} & \left| \left\| p^{[k']} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 - \left\| p^{[k']} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 \right| \\ & \leq O \left(d^{-c_g k' + (c_g/2 - c_{\text{trunc}})} \right) p^2 \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right). \end{aligned} \quad (46)$$

We first show how to obtain [Equation \(44\)](#) using [Claim 3.9](#). Applying [Equation \(46\)](#) with $k' = 0$ yields

$$\begin{aligned} & \left| p^2 \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) - p^2 \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right| \\ & \leq O \left(d^{-(c_g/2 - c_{\text{trunc}})} \right) p^2 \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right). \end{aligned}$$

Since we assume that $(c_g/2 - c_{\text{trunc}}) \geq \Omega(1)$, it then follows that

$$p^2 \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \leq (1 + o(1)) p^2 \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right). \quad (47)$$

Applying [Equation \(46\)](#) for $k' \in [k]$ gives

$$\begin{aligned} & \left\| p^{[k']} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 \\ & \leq \left\| p^{[k']} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 + O \left(d^{-c_g(k'+0.01)} \right) p^2 \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \\ & \leq \left(d^{-c_g k'} + O \left(d^{-c_g k' + (c_g/2 - c_{\text{trunc}})} \right) \right) p^2 \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \quad (\text{using } \text{Equation (45)}) \\ & \leq (1 + o(1)) d^{-c_g k'} p^2 \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right). \quad (\text{using } \text{Equation (47)}) \end{aligned}$$

This proves [Equation \(44\)](#) and completes the proof of [Equation \(44\)](#). \square

It remains to show [Claim 3.9](#).

Proof of Claim 3.9. We will use the Taylor expansion of $\|p^{[k']}(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)})\|_F^2$ (which is a degree at most $2(k - k')$ polynomial) in the variable ξ centered at ξ^* . In particular, Taylor's theorem gives that

$$\begin{aligned} & \left\| p^{[k']} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 \\ & = \left\| p^{[k']} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 \\ & + \sum_{t=1}^{2(k-k')} \left(\frac{\partial^t}{\partial \xi^t} \left\| p^{[k']} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 \Big|_{\xi=\xi^*} \right) \frac{(\xi - \xi^*)^t}{t!}. \end{aligned} \quad (48)$$

We claim that the derivative $\frac{\partial^t}{\partial \xi^t} \left\| p^{[k']} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 \Big|_{\xi=\xi^*}$ satisfies the following bound:

$$\begin{aligned} & \frac{\partial^t}{\partial \xi^t} \left\| p^{[k']} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 \Big|_{\xi=\xi^*} \\ & \leq O(1) 2^t d^{-c_g(t/2+k')} p^2 \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right). \end{aligned} \quad (49)$$

Combining [Equation \(49\)](#) and $|\xi - \xi^*| \leq d^{c_{\text{trunc}}}$ then gives that

$$\begin{aligned} & \left| \frac{\partial^t}{\partial \xi^t} \left\| p^{[k']} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 \Big|_{\xi=\xi^*} (\xi - \xi^*)^t \right| \leq O(1) 2^t d^{-c_g(t/2+k')+t c_{\text{trunc}}} \\ & = O(1) 2^t d^{-c_g k' - (c_g/2 - c_{\text{trunc}}) t}. \end{aligned}$$

Combining the above with [Equation \(48\)](#), and the fact that $\sum_t 2^t/t! = O(1)$ then gives that

$$\begin{aligned} & \left| \left\| p^{[k']} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 - \left\| p^{[k']} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 \right| \\ & \leq O \left(d^{-c_g k' - (c_g/2 - c_{\text{trunc}})t} \right) p^2 \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right). \end{aligned}$$

This concludes the proof of [Equation \(46\)](#). It remains to show [Equation \(49\)](#). Using the product rule, the derivative $\frac{\partial^t}{\partial \xi^t} \left\| p^{[k']} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 \Big|_{\xi=\xi^*}$ is a sum of at most 2^t terms of the following form:

$$\left\langle p^{[\beta]} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \mathbf{e}(i)^{\otimes \beta'}, p^{[\gamma]} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \mathbf{e}(i)^{\otimes \gamma'} \right\rangle,$$

where $\beta, \beta', \gamma, \gamma'$ are natural numbers such that $\beta - \beta' = \gamma - \gamma'$, and $\beta + \gamma = 2k' + t$. Combining the above observation with the Cauchy–Schwarz inequality then gives the bound

$$\begin{aligned} & \left| \frac{\partial^t}{\partial \xi^t} \left\| p^{[k']} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 \Big|_{\xi=\xi^*} \right| \\ & \leq 2^t \max_{\beta, \gamma: \beta+\gamma=2k'+t} \left\| p^{[\beta]} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F \left\| p^{[\gamma]} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F \\ & \leq O(1) 2^t d^{-c_g(k'+t/2)} p^2(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)}), \end{aligned} \tag{Equation (45)}$$

where the last line used that the set of points $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i-1)}, \bar{\mathbf{y}}, \mathbf{y}^{(i+1)}, \dots, \mathbf{y}^{(n)}$ is well-behaved ([Definition 3.7](#)). This concludes the proof of [Claim 3.9](#). \square

3.3.3 Taylor Expansion of the Mollified PTF and Bounds for the Higher-Order Terms

As explained in the technical overview of [Section 1.2](#), the goal is to prove [Equation \(36\)](#) by performing a Taylor expansion of h . We then use the moment-matching property of the hidden direction distribution to bound the contribution of the low-order terms to the difference of expectations on the LHS of [Equation \(36\)](#) and leverage the derivative decay property of the mollifier to show that the contribution from the high-order error term is also small. The main result of this subsection formalizes the second argument that bounds the derivative appearing in the Taylor error term in [Lemma 3.10](#). In particular, we consider the degree- m expansion of h along the direction of \mathbf{v} at its i -th coordinate around some point ξ^* that will be specified later:

$$\begin{aligned} h(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)}) &= \sum_{t=0}^{m-1} ((\xi - \xi^*)^t / t!) D_{i,\mathbf{v}}^t h \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \\ &\quad + ((\xi - \xi^*)^m / m!) D_{i,\mathbf{v}}^m h \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \hat{\xi} \mathbf{v}, \dots, \mathbf{y}^{(n)} \right), \end{aligned} \tag{50}$$

where $\hat{\xi}$ is some point between ξ and ξ^* which also depends on $\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}}, \dots, \mathbf{y}^{(n)}$. Recall that A has its first m -moments matched with $\mathcal{N}(0, 1)$. So the expected values of the first m terms in [Equation \(50\)](#) are identical under $\xi \sim A$ and $\xi \sim \mathcal{N}(0, 1)$. The rest of the section will focus on how we control the magnitude of the last term $((\xi - \xi^*)^m / m!) D_{i,\mathbf{v}}^m h(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \hat{\xi} \mathbf{v}, \dots, \mathbf{y}^{(n)})$.

We now proceed to control the magnitudes of the derivatives $D_{i,\mathbf{v}}^m h \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \hat{\xi} \mathbf{v}, \dots, \mathbf{y}^{(n)} \right)$.

Lemma 3.10 (Mollified PTF Derivative Decay). *Let $p : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}$ be a degree k polynomial, \mathbf{v} be a unit vector satisfying [Equation \(12\)](#) with $\varepsilon = 0.05$, and h be the mollified PTF defined in [Equation \(23\)](#). Let c_g be*

the constant that appears in the definition of h , and $c_{trunc} > 0$ be a parameter satisfying $c_g/2 - c_{trunc} \geq \Omega(1)$. Let $\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}}, \dots, \mathbf{y}^{(n)}$ be a well-behaved point set (cf. [Definition 3.7](#)) at position i , with respect to $p, [-d^{c_{trunc}}, d^{c_{trunc}}]$, and c_g . Then for all $t \in \mathbb{Z}_+$ and all $\xi \in [-d^{c_{trunc}}, d^{c_{trunc}}]$ it holds that

$$D_{i,\mathbf{v}}^t h(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)}) \leq (k+t)^{O(t)} d^{-c_g t/2}.$$

The first observation is that computing $D_{i,\mathbf{v}}^t h(\cdot)$ essentially boils down to computing the derivatives of the mollifier $D_{i,\mathbf{v}}^t g(\cdot)$.

Claim 3.11. *For all $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \in \mathbb{R}^d$, we have that*

$$D_{i,\mathbf{v}}^t h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) = \text{sign}(p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})) D_{i,\mathbf{v}}^t g(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}).$$

Proof. The claim follows from the product rule and the fact that the derivatives of $\text{sign}(p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}))$ are 0 almost everywhere. \square

Writing down the exact expression of the derivatives of the mollifier i.e., $D_{i,\mathbf{v}}^t g(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$, is quite tedious. However, we show that the derivative is the sum of at most $(2k)^t$ many terms of a specific functional form. For presentation purposes, similarly to earlier sections, we abbreviate $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ by $\mathbf{x}^{(1:n)}$.

Lemma 3.12 (Unfolded Derivatives of the Mollifier). *Let $i \in [n], t \in [k]$, and \mathbf{v} be some unit vector, and p, ρ, g defined as in [Equation \(21\)](#). Then $D_{i,\mathbf{v}}^t g(\mathbf{x}^{(1:n)})$ is a sum of at most $T := k^{O(t)}$ terms where the j -th term is of the form:*

$$\begin{aligned} \Lambda_{j,t} := \pm d^{(c_g/2)\kappa_j^{(t)}} & \left(\prod_{(\alpha, \alpha') \in \mathcal{A}_j^{(t)}} \rho^{(\alpha)} \left(\frac{d^{c_g \alpha'} \|p^{[\alpha']}(\mathbf{x}^{(1:n)})\|_F^2}{p^2(\mathbf{x}^{(1:n)})} \right) \right) \\ & \left(\prod_{(\beta, \beta', \gamma, \gamma') \in \mathcal{B}_j^{(t)}} \frac{\langle p^{[\beta]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \beta'}, p^{[\gamma]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \gamma'} \rangle}{p^2(\mathbf{x}^{(1:n)})} \right), \end{aligned} \quad (51)$$

where $\kappa_j^{(t)} \in \mathbb{Z}_+$, $\mathcal{A}_j^{(t)}$ is a multiset made up of elements from $(\{0\} \cup [k])^2$, $\mathcal{B}_j^{(t)}$ is a multiset made up of elements from $(\{0\} \cup [k])^4$ ⁸, and $\mathbf{e}(i)$ is the i -th standard basis vector. Moreover, $\mathcal{A}_j^{(t)}, \mathcal{B}_j^{(t)}$ satisfy

- **Maximum ρ derivative degree:** $\sum_{\alpha \in \mathcal{A}_j^{(t)}} \alpha \leq t$ for all $\mathcal{A}_j^{(t)}$.
- **Cardinality bound:** $|\mathcal{A}_{j,t}| + |\mathcal{B}_{j,t}| \leq k + t$.
- **Degree growth:** $-\kappa_j^{(t)} + \sum_{(\beta, \beta', \gamma, \gamma') \in \mathcal{B}_j^{(t)}} (\beta + \gamma) \geq t$ for every $j \in [T]$.

Proof. We will show this by induction on t . The base case is when $t = 0$. This corresponds to the case where we just have one term, where $\kappa_1^{(0)} = 0$, $\mathcal{A}_1^{(0)} = \{(i, i) \mid i \in [k]\}$, and $\mathcal{B}_1^{(0)} = \emptyset$. The properties are then immediate.

⁸Of course, for the expression to be well defined, we will need $\beta - \beta' = \gamma - \gamma'$ as the tensor dimensions will not match up otherwise.

We proceed to show the inductive step. For convenience, define

$$\begin{aligned}\Lambda_{j,t}^{-(\alpha,\alpha')} &:= \Lambda_{j,t} \left(\rho^{(\alpha)} \left(\frac{d^{c_g} \alpha' \|p^{[\alpha']}(\mathbf{x}^{(1:n)})\|_F^2}{p^2(\mathbf{x}^{(1:n)})} \right) \right)^{-1}, \\ \Lambda_{j,t}^{-(\beta,\beta',\gamma,\gamma')} &:= \Lambda_{j,t} \left(\frac{\langle p^{[\beta]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \beta'}, p^{[\gamma]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \gamma'} \rangle}{p^2(\mathbf{x}^{(1:n)})} \right)^{-1}.\end{aligned}$$

By the inductive hypothesis, we have that $D_{i,\mathbf{v}}^{t+1} g(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) = \sum_j D_{i,\mathbf{v}} \Lambda_{j,t}$. So it remains to compute $D_{i,\mathbf{v}} \Lambda_{j,t}$. In particular, by the product rule, we have that

$$\begin{aligned}D_{i,\mathbf{v}} \Lambda_{j,t} &:= \sum_{(\alpha,\alpha') \in \mathcal{A}_{j,t}} \Lambda_{j,t}^{-(\alpha,\alpha')} D_{i,\mathbf{v}} \rho^{(\alpha)} \left(\frac{d^{c_g} \alpha' \|p^{[\alpha']}(\mathbf{x}^{(1:n)})\|_F^2}{p^2(\mathbf{x}^{(1:n)})} \right) \\ &\quad + \sum_{(\beta,\beta',\gamma,\gamma') \in \mathcal{B}_{j,t}} \Lambda_{j,t}^{-(\beta,\beta',\gamma,\gamma')} D_{i,\mathbf{v}} \frac{\langle p^{[\beta]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \beta'}, p^{[\gamma]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \gamma'} \rangle}{p^2(\mathbf{x}^{(1:n)})}.\end{aligned}$$

We first analyze the term $D_{i,\mathbf{v}} \left(\langle p^{[\beta]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \beta'}, p^{[\gamma]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \gamma'} \rangle \right) p^2(\mathbf{x}^{(1:n)})$.

$$\begin{aligned}&D_{i,\mathbf{v}} \frac{\langle p^{[\beta]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \beta'}, p^{[\gamma]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \gamma'} \rangle}{p^2(\mathbf{x}^{(1:n)})} \\ &= \frac{D_{i,\mathbf{v}} \langle p^{[\beta]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \beta'}, p^{[\gamma]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \gamma'} \rangle}{p^2(\mathbf{x}^{(1:n)})} \\ &\quad - \frac{\langle p^{[\beta]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \beta'}, p^{[\gamma]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \gamma'} \rangle D_{i,\mathbf{v}} p^2(\mathbf{x}^{(1:n)})}{p^4(\mathbf{x}^{(1:n)})} \tag{Quotient rule} \\ &= \frac{\langle p^{[\beta+1]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \beta'+1}, p^{[\gamma]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \gamma'} \rangle}{p^2(\mathbf{x}^{(1:n)})} + \frac{\langle p^{[\beta]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \beta'}, p^{[\gamma+1]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \gamma'+1} \rangle}{p^2(\mathbf{x}^{(1:n)})} \\ &\quad - \frac{\langle p^{[\beta]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \beta'}, p^{[\gamma]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \gamma'} \rangle}{p^2(\mathbf{x}^{(1:n)})} \frac{2(p^{[1]}(\mathbf{x}^{(1:n)}) \mathbf{e}^{(i)}) p(\mathbf{x}^{(1:n)})}{p^2(\mathbf{x}^{(1:n)})}. \tag{Product rule}\end{aligned}$$

One can check that each term in the summation above is still of the desired form. Moreover, we have that:

- The total derivative degree on β is $\beta + \gamma + 1$ while the power κ_j in the leading constant stays unchanged, ensuring the desired potential growth.
- Since the last line contains 3 terms, it follows that the number of terms in

$$\sum_{(\beta,\beta',\gamma,\gamma') \in \mathcal{B}_{j,t}} \Lambda_{j,t}^{-(\beta,\beta',\gamma,\gamma')} D_{i,\mathbf{v}} \frac{\langle p^{[\beta]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \beta'}, p^{[\gamma]}(\mathbf{x}^{(1:n)}) (\mathbf{e}^{(i)})^{\otimes \gamma'} \rangle}{p^2(\mathbf{x}^{(1:n)})}$$

is $3|\mathcal{B}_{j,t}|$.

- The set cardinality bound increases by 1 due to the third term.

Next, we analyze the term $D_{i,\mathbf{v}}\rho^{(\alpha)} \left(d^{c_g}\alpha' \|p^{[\alpha']}(\mathbf{x}^{(1:n)})\|_F^2 / p^2(\mathbf{x}^{(1:n)}) \right)$.

$$\begin{aligned}
& D_{i,\mathbf{v}}\rho^{(\alpha)} \left(\frac{d^{c_g}\alpha' \|p^{[\alpha']}(\mathbf{x}^{(1:n)})\|_F^2}{p^2(\mathbf{x}^{(1:n)})} \right) \\
&= \rho^{(\alpha+1)} \left(\frac{d^{c_g}\alpha' \|p^{[\alpha']}(\mathbf{x}^{(1:n)})\|_F^2}{p^2(\mathbf{x}^{(1:n)})} \right) D_{i,\mathbf{v}} \frac{d^{c_g}\alpha' \|p^{[\alpha']}(\mathbf{x}^{(1:n)})\|_F^2}{p^2(\mathbf{x}^{(1:n)})} && \text{(Chain rule)} \\
&= d^{c_g}\alpha' \rho^{(\alpha+1)} \left(\frac{d^{c_g}\alpha' \|p^{[\alpha']}(\mathbf{x}^{(1:n)})\|_F^2}{p^2(\mathbf{x}^{(1:n)})} \right) \left(\frac{D_{i,\mathbf{v}}\|p^{[\alpha']}(\mathbf{x}^{(1:n)})\|_F^2}{p^2(\mathbf{x}^{(1:n)})} - \frac{\|p^{[\alpha']}(\mathbf{x}^{(1:n)})\|_F^2 D_{i,\mathbf{v}} p^2(\mathbf{x}^{(1:n)})}{p^4(\mathbf{x}^{(1:n)})} \right) \\
&\quad \text{(Quotient rule)}
\end{aligned} \tag{52}$$

$$\begin{aligned}
&= d^{c_g}\alpha' \rho^{(\alpha+1)} \left(\frac{d^{(c_g/2)2\alpha'} \|p^{[\alpha']}(\mathbf{x}^{(1:n)})\|_F^2}{p^2(\mathbf{x}^{(1:n)})} \right) \left(\frac{2 \langle p^{[\alpha'+1]}(\mathbf{x}^{(1:n)}) \mathbf{e}^{(i)}, p^{[\alpha']}(\mathbf{x}^{(1:n)}) \rangle}{p^2(\mathbf{x}^{(1:n)})} \right) \\
&\quad - \frac{\langle p^{[\alpha']}(\mathbf{x}^{(1:n)}) \mathbf{e}^{(i)}, p^{[\alpha']}(\mathbf{x}^{(1:n)}) \rangle \langle p^{[1]}(\mathbf{x}^{(1:n)}) \mathbf{e}^{(i)}, p(\mathbf{x}^{(1:n)}) \rangle}{p^2(\mathbf{x}^{(1:n)})}. && \text{(Product Rule)} \\
\end{aligned} \tag{53}$$

One can check that each term in the summation above is still of the desired form. Moreover, we have that

- The total derivative degree is $2\alpha' + 1$ while the power κ_j in the leading constant increases by $2\alpha'$, ensuring the desired potential growth.
- Since the above equation has 2 additive terms, it follows that

$$\sum_{(\alpha,\alpha') \in \mathcal{A}_{j,t}} \Lambda_{j,t}^{-(\alpha,\alpha')} D_{i,\mathbf{v}}\rho^{(\alpha)} \left(\frac{d^{c_g}\alpha' \|p^{[\alpha']}(\mathbf{x}^{(1:n)})\|_F^2}{p^2(\mathbf{x}^{(1:n)})} \right)$$

contains at most $2|\mathcal{A}_{j,t}|$ terms.

- The set cardinality bound increases by 1 due to the last term.
- The maximum derivative degree of ρ increases by 1.

The total number of terms in $D_{i,\mathbf{v}}^{t+1}g(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$ is at most $(O(k))^t (3|\mathcal{B}_{j,t}| + 2|\mathcal{A}_{j,t}|) \leq (O(k))^{t+1}$. This concludes the inductive step as well as the proof of [Lemma 3.12](#). \square

We are now ready to conclude the proof of [Lemma 3.10](#).

Proof of Lemma 3.10. For notational convenience, we define $\mathbf{x}^{(i)} = \bar{\mathbf{y}} + \xi \mathbf{v}$. Recall that g and h are the mollifier and the mollified PTF respectively, defined in [Equations \(22\)](#) and [\(23\)](#). By [Claim 3.11](#), it holds

$$\left| D_{i,\mathbf{v}}^t h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) \right| \leq \left| D_{i,\mathbf{v}}^t g(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) \right|.$$

By [Lemma 3.12](#), $D_{i,\mathbf{v}}^t g(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$ is the sum of at most $k^{O(t)}$ terms $\Lambda_{j,t}$ of the form given in [Equation \(51\)](#). We now claim that each term in the form of [Equation \(51\)](#) is at most $t^{O(t)} d^{-c_g t/2}$. It follows immediately that

$$\begin{aligned}
\left| D_{i,\mathbf{v}}^t g(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) \right| &\leq t^{O(t)} k^{O(t)} d^{-c_g t/2} \\
&\leq (t+k)^{O(t)} d^{-c_g t/2}
\end{aligned}$$

for all $t \in \mathbb{Z}_+$.

It remains to show that the quantity $\Lambda_{j,t}$ in [Equation \(51\)](#) is at most $t^{O(t)}d^{-c_g t/2}$. Recall that $\|\rho^{(\alpha)}\|_\infty$ is at most $\alpha^{O(\alpha)}$ by [Equation \(21\)](#). Combining this with the maximum degree property of [Lemma 3.12](#) then gives that

$$\prod_{(\alpha, \alpha') \in \mathcal{A}_j^{(t)}} \rho^{(\alpha)} \left(\frac{d^{c_g \alpha'} \|p^{[\alpha']}(\mathbf{x}^{(1:n)})\|_F^2}{p^2(\mathbf{x}^{(1:n)})} \right) \leq t^{O(t)}. \quad (54)$$

Next, applying the Cauchy–Schwarz inequality gives that

$$\frac{\langle p^{[\beta]}(\mathbf{x}^{(1:n)}) (e^{(i)})^{\otimes \beta'}, p^{[\gamma]}(\mathbf{x}^{(1:n)}) (e^{(i)})^{\otimes \gamma'} \rangle}{p^2(\mathbf{x}^{(1:n)})} \leq \frac{\|p^{[\beta]}(\mathbf{x}^{(1:n)})\|_F}{|p(\mathbf{x}^{(1:n)})|} \frac{\|p^{[\gamma]}(\mathbf{x}^{(1:n)})\|_F}{|p(\mathbf{x}^{(1:n)})|}. \quad (55)$$

Recall that we define $\mathbf{x}^{(i)} := \bar{\mathbf{y}} + \xi \mathbf{v}$. By the assumption of the lemma, $\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}}, \dots, \mathbf{x}^{(n)}$ form a well-behaved point set with respect to $p, [-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]$, and c_g (cf. [Definition 3.7](#)), and $\xi \in [-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]$. Therefore, [Lemma 3.8](#) is applicable. This gives that

$$\frac{\|p^{[\beta]}(\mathbf{x}^{(1:n)})\|_F}{|p(\mathbf{x}^{(1:n)})|} \frac{\|p^{[\gamma]}(\mathbf{x}^{(1:n)})\|_F}{|p(\mathbf{x}^{(1:n)})|} \leq O(d^{-c_g(\beta+\gamma)/2}). \quad (56)$$

We will use [Equation \(56\)](#) to bound the remaining part of $\Lambda_{j,t}$ from [Equation \(51\)](#); all the factors in the RHS of [Equation \(51\)](#) excluding the $\prod_{(\alpha, \alpha') \in \mathcal{A}_j^{(t)}} \rho^{(\alpha)} \left(d^{c_g \alpha'} \|p^{[\alpha']}(\mathbf{x}^{(1:n)})\|_F^2 / p^2(\mathbf{x}^{(1:n)}) \right)$ that have already been bounded earlier. Combining [Equation \(56\)](#) with [Equation \(55\)](#) gives the following bound for the term

$$d^{(c_g/2)\kappa_j^{(t)}} \prod_{(\beta, \beta', \gamma, \gamma') \in \mathcal{B}_{j,t}} \frac{\langle p^{[\beta]}(\mathbf{x}^{(1:n)}) (e^{(i)})^{\otimes \beta'}, p^{[\gamma]}(\mathbf{x}^{(1:n)}) (e^{(i)})^{\otimes \gamma'} \rangle}{p^2(\mathbf{x}^{(1:n)})} \\ \leq O\left(d^{-(c_g/2)(\sum_{(\beta, \beta', \gamma, \gamma') \in \mathcal{B}_{j,t}} \beta + \gamma - \kappa_j^{(t)})}\right).$$

By the degree growth property of [Lemma 3.12](#), it follows that

$$d^{(c_g/2)\kappa_j^{(t)}} \prod_{(\beta, \beta', \gamma, \gamma') \in \mathcal{B}_{j,t}} \frac{\langle p^{[\beta]}(\mathbf{x}^{(1:n)}) (e^{(i)})^{\otimes \beta'}, p^{[\gamma]}(\mathbf{x}^{(1:n)}) (e^{(i)})^{\otimes \gamma'} \rangle}{p^2(\mathbf{x}^{(1:n)})} \leq d^{-c_g t/2}. \quad (57)$$

Combining [Equation \(57\)](#) and [Equation \(54\)](#) then gives that the expression in [Equation \(51\)](#) is at most $t^{O(t)}d^{-c_g t/2}$. This concludes the proof of [Lemma 3.10](#). \square

3.3.4 Putting Everything Together: Proof of [Proposition 3.5](#)

We are now ready to conclude the proof of [Proposition 3.5](#), restated below for convenience:

Proposition 3.5 (Replacement Step). *For any $c \in (0, 1/4)$, $d, m, k \in \mathbb{Z}_+$ such that m is even, and $d > \max(m^{C/c}, k^{C/c})$ for some sufficiently large constant C ,*

if $p : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ is a degree- k polynomial, $\mathbf{v} \in \mathbb{R}^d$ is a unit vector satisfying [Equation \(12\)](#) with $\varepsilon = 0.05$, A is a one-dimensional distribution that matches the first m moments with $\mathcal{N}(0, 1)$, and $h : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ is the mollified PTF from [Equation \(23\)](#), then the following holds: For every $i \in [n]$

$$\left| \mathbf{E} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \mathbf{y}^{(i+1)}, \dots, \mathbf{y}^{(n)}) \right] - \mathbf{E} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{y}^{(i)}, \mathbf{y}^{(i+1)}, \dots, \mathbf{y}^{(n)}) \right] \right| \leq d^{-c_g m/2 + cm}, \quad (31)$$

where c_g is the parameter used in [Equation \(23\)](#), $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{M}_{A, \mathbf{v}}$, and $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Proof. First, using an averaging argument, it suffices to show the following for an arbitrary set of points $S_i := \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i-1)}, \bar{\mathbf{y}}, \mathbf{y}^{(i+1)}, \dots, \mathbf{y}^{(n)}\}$:

$$\begin{aligned} & \left| \mathbf{E}_{\xi \sim A} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)}) \right] - \mathbf{E}_{z \sim \bar{\mathcal{N}}(0,1)} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \bar{\mathbf{y}} + z \mathbf{v}, \dots, \mathbf{y}^{(n)}) \right] \right| \\ & \leq d^{-c_g m/2 + cm}. \end{aligned} \quad (58)$$

Throughout the proof, we will fix $c_{\text{trunc}} = c_g/2 - c/2$, and set the truncated domain to be $I = [-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]$.

It then follows from [Lemma 3.6](#) that

$$\max \left(\mathbf{E}_{\xi \sim A} [\xi \in I], \mathbf{E}_{\xi \sim \bar{\mathcal{N}}(0,1)} [\xi \in I] \right) \leq m^m d^{-c_{\text{trunc}} m} \leq d^{-c_g m/2 + cm},$$

where the last inequality follows from the assumption that $m < d^{c/C}$

Since the mollified PTF is constructed to be bounded from above by 1, showing [Equation \(58\)](#) can be reduced to showing

$$\begin{aligned} & \left| \mathbf{E}_{\xi \sim \bar{A}} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)}) \right] - \mathbf{E}_{z \sim \bar{\mathcal{N}}(0,1)} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \bar{\mathbf{y}} + z \mathbf{v}, \dots, \mathbf{y}^{(n)}) \right] \right| \\ & \leq d^{-c_{\text{trunc}} m + cm/2}, \end{aligned} \quad (59)$$

where \bar{A} and $\bar{\mathcal{N}}(0,1)$ are the truncated versions of the distributions that condition on ξ and z being inside the interval $[-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]$.

We will show [Equation \(59\)](#) by considering two cases. The first is when S_i is not a well-behaved point set ([Definition 3.7](#)) at position i . That is, we have that for every $\xi \in [-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]$, there exists some t such that

$$\left\| p^{[t]} \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right\|_F^2 > 3d^{-c_g t} \left(p(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{y}, \dots, \mathbf{y}^{(m)}) \right)^2.$$

By the definition of the mollified PTF (cf. [Equation \(22\)](#)), we immediately have that in this case $h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)}) = 0$ for all $\xi \in [-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]$. Thus, [Equation \(59\)](#) follows trivially in this case as the left hand side is zero.

Now consider the complementary case where S_i is a well-behaved point set at position i . By [Lemma 3.10](#), we have that

$$D_{i,\mathbf{v}}^t h \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \leq (k+t)^{O(t)} d^{-c_g t/2} \quad (60)$$

for all $t \in [m]$ and $\xi \in [-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]$. In this case, we will show [Equation \(59\)](#) by rewriting h in terms of its Taylor expansion, and then bounding the differences between the Taylor terms. Consider the degree- m Taylor expansion of h around 0:

$$\begin{aligned} h(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)}) &= \sum_{t=0}^{m-1} (\xi^t / t!) D_{i,\mathbf{v}}^t h \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \\ &\quad + (\xi^m / m!) D_{i,\mathbf{v}}^m h \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \widehat{\xi} \mathbf{v}, \dots, \mathbf{y}^{(n)} \right), \end{aligned} \quad (61)$$

where $\widehat{\xi}$ is some point between ξ and ξ^* which also depends on $\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}}, \dots, \mathbf{y}^{(n)}$ and ξ^* lies in $[-d^{c_{\text{trunc}}}, d^{c_{\text{trunc}}}]$. For convenience, we write $\Delta(a) := \mathbf{E}_{\xi \sim \bar{A}}[\xi^a] - \mathbf{E}_{\xi \sim \bar{\mathcal{N}}(0,1)}[\xi^a]$. For $t \in [m-1]$, we have that

$$\left| D_{i,\mathbf{v}}^t h \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \left(\mathbf{E}_{\xi \sim \bar{A}}[\xi^t] - \mathbf{E}_{\xi \sim \bar{\mathcal{N}}(0,1)}[\xi^t] \right) \right| \quad (62)$$

$$= \left| D_{i,\mathbf{v}}^t h \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \Delta(a) \right| \quad (63)$$

$$\begin{aligned} &\leq (m+t)^{O(t)} d^{-c_g t/2} d^{-c_{\text{trunc}}(m-t)} && (\text{by Equation (60), and Lemma 3.6}) \\ &\leq (m+t)^{O(t)} d^{-c_{\text{trunc}} m}. && (\text{using } c_{\text{trunc}} = c_g/2 - 0.01) \end{aligned}$$

In particular, this implies that

$$\left| \mathbf{E}_{\xi \sim \bar{A}} \left[\sum_{t=1}^{m-1} (\xi^t / t!) D_{i,\mathbf{v}}^t h \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right] - \right| \quad (64)$$

$$\left| \mathbf{E}_{\xi \sim \bar{\mathcal{N}}(0,1)} \left[\sum_{t=1}^{m-1} (\xi^t / t!) D_{i,\mathbf{v}}^t h \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \xi^* \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \right] \right| \quad (65)$$

$$\leq m m^{O(m)} d^{-c_{\text{trunc}} m} = m^{O(m)} d^{-c_{\text{trunc}} m}, \quad (66)$$

where we used the triangle inequality. For the last Taylor remainder term, applying Equation (60) again gives that

$$(\xi^m / m!) D_{i,\mathbf{v}}^m h \left(\mathbf{x}^{(1)}, \dots, \bar{\mathbf{y}} + \widehat{\xi} \mathbf{v}, \dots, \mathbf{y}^{(n)} \right) \leq (\xi^m / m!) (k+m)^{O(m)} d^{-c_g m/2},$$

for all ξ . In particular, this implies that the expected value of the Taylor remainder term under the distribution of $\xi \sim \bar{A}$ is at most

$$\begin{aligned} &(k+m)^{O(m)} d^{-c_g m/2} \mathbf{E}_{\xi \sim \bar{A}} [\xi^m / m!] \\ &\leq (k+m)^{O(m)} d^{-c_g m/2} \mathbf{E}_{\xi \sim A} [\xi^m / m!] && (\text{since the mass of } A \text{ within the truncated interval is } 1 - o(1)) \\ &= (k+m)^{O(m)} d^{-c_g m/2} \mathbf{E}_{\xi \sim \bar{\mathcal{N}}(0,1)} [(\xi^m / m!)] && (\text{since we assume the degree-}m \text{ moment of } A \text{ match with } \mathcal{N}(0,1)) \\ &\leq (k+m)^{O(m)} d^{-c_g m/2} && (\text{by the Gaussian moment bound}) \end{aligned}$$

The same bound can be established for the Taylor remainder term under the distribution of $\xi \sim \bar{\mathcal{N}}(0,1)$ as well. It then follows from the triangle inequality that the difference between the expected value of the Taylor remainder term under $\bar{\mathcal{N}}(0,1)$ and \bar{A} is at most $(k+m)^{O(m)} d^{-c_g m/2}$. Combining this with Equations (61) and (64) then shows that

$$\begin{aligned} &\left| \mathbf{E}_{\xi \sim \bar{A}} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \bar{\mathbf{y}} + \xi \mathbf{v}, \dots, \mathbf{y}^{(n)}) \right] - \mathbf{E}_{\xi \sim \bar{\mathcal{N}}(0,1)} \left[h(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \bar{\mathbf{y}} + z \mathbf{v}, \dots, \mathbf{y}^{(n)}) \right] \right| \\ &\leq (k+m)^{O(m)} d^{-c_{\text{trunc}} m} \leq d^{-c_{\text{trunc}} m + cm/2}, \end{aligned}$$

where the last inequality follows from the assumption that $d > \max(k^{C/c}, m^{C/c})$. This concludes the proof of Equation (59), as well as Proposition 3.5. \square

References

- [ABX08] B. Applebaum, B. Barak, and D. Xiao. On basing lower-bounds for learning on worst-case assumptions. In *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008*, pages 211–220. IEEE Computer Society, 2008.
- [BBH⁺21] M. Brennan, G. Bresler, S. B. Hopkins, J. Li, and T. Schramm. Statistical query algorithms and low-degree tests are almost equivalent. In *Proceedings of The 34th Conference on Learning Theory, COLT*, 2021.
- [BHK⁺16] B. Barak, S. B. Hopkins, J. A. Kelner, P. Kothari, A. Moitra, and A. Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. *FOCS*, 2016.
- [BKS⁺06] G. Blanchard, M. Kawanabe, M. Sugiyama, V. Spokoiny, and K. B. Müller. In search of non-gaussian components of a high-dimensional distribution. *Journal of Machine Learning Research*, 7(9):247–282, 2006.
- [BKW19] A. S. Bandeira, D. Kunisky, and A. S. Wein. Computational hardness of certifying bounds on constrained pca problems. *arXiv preprint arXiv:1902.07324*, 2019.
- [BS16] B. Barak and D. Steurer. Proofs, beliefs, and algorithms through the lens of sum-of-squares. 2016.
- [Cho61] C. K. Chow. On the characterization of threshold functions. In *Proceedings of the Symposium on Switching Circuit Theory and Logical Design (FOCS)*, pages 34–38, 1961.
- [CW01] A. Carbery and J. Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in R^n . *Mathematical Research Letters*, 8(3):233–248, 2001.
- [DDFS14] A. De, I. Diakonikolas, V. Feldman, and R. A. Servedio. Nearly optimal solutions for the chow parameters problem and low-weight approximation of halfspaces. *J. ACM*, 61(2):11:1–11:36, 2014.
- [Der65] M. Dertouzos. *Threshold Logic: A Synthesis Approach*. MIT Press, Cambridge, MA, 1965.
- [DGJ⁺09] I. Diakonikolas, P. Gopalan, R. Jaiswal, R. Servedio, and E. Viola. Bounded independence fools halfspaces. In *Proc. 50th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 171–180, 2009.
- [DGJ⁺10] I. Diakonikolas, P. Gopalan, R. Jaiswal, R. Servedio, and E. Viola. Bounded independence fools halfspaces. *SIAM Journal on Computing*, 39(8):3441–3462, 2010.
- [DH24] R. Dudeja and D. Hsu. Statistical-computational trade-offs in tensor pca and related problems via communication complexity. *The Annals of Statistics*, 52(1):131–156, 2024.
- [DHK⁺10] I. Diakonikolas, P. Harsha, A. Klivans, R. Meka, P. Raghavendra, R. A. Servedio, and L. Y. Tan. Bounding the average sensitivity and noise sensitivity of polynomial threshold functions. In *STOC*, pages 533–542, 2010.
- [DJNS13] E. Diederichs, A. Juditsky, A. Nemirovski, and V. Spokoiny. Sparse non gaussian component analysis by semidefinite programming. *Machine learning*, 91:211–238, 2013.

- [DK19] I. Diakonikolas and D. M. Kane. Degree- d chow parameters robustly determine degree- d ptfs (and algorithmic applications). In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pages 804–815. ACM, 2019.
- [DK23] I. Diakonikolas and D. M. Kane. *Algorithmic high-dimensional robust statistics*. Cambridge university press, 2023.
- [DKK⁺22] I. Diakonikolas, D. M. Kane, S. Karmalkar, A. Pensia, and T. Pittas. Robust sparse mean estimation via sum of squares. In *Conference on Learning Theory*, pages 4703–4763. PMLR, 2022.
- [DKK⁺24] I. Diakonikolas, D. M. Kane, V. Kontonis, S. Liu, and N. Zarifis. Super non-singular decompositions of polynomials and their application to robustly learning low-degree ptfs. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024, Vancouver, BC, Canada, June 24-28, 2024*, pages 152–159. ACM, 2024.
- [DKN10] I. Diakonikolas, D. M. Kane, and J. Nelson. Bounded independence fools degree-2 threshold functions. In *FOCS*, pages 11–20, 2010.
- [DKP⁺21] I. Diakonikolas, D. M. Kane, A. Pensia, T. Pittas, and A. Stewart. Statistical query lower bounds for list-decodable linear regression. *Advances in Neural Information Processing Systems*, 34:3191–3204, 2021.
- [DKPP24] I. Diakonikolas, S. Karmalkar, S. Pang, and A. Potechin. Sum-of-squares lower bounds for non-gaussian component analysis. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 949–958. IEEE, 2024.
- [DKRS23] I. Diakonikolas, D. Kane, L. Ren, and Y. Sun. Sq lower bounds for non-gaussian component analysis with weaker assumptions. *Advances in Neural Information Processing Systems*, 36:4199–4212, 2023.
- [DKS17] I. Diakonikolas, D. M. Kane, and A. Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84, 2017.
- [DKS18a] I. Diakonikolas, D. M. Kane, and A. Stewart. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pages 1061–1073, 2018.
- [DKS18b] I. Diakonikolas, D. M. Kane, and A. Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060, 2018.
- [DKWB24] Y. Ding, D. Kunisky, A. S. Wein, and A. S. Bandeira. Subexponential-time algorithms for sparse pca. *Foundations of Computational Mathematics*, 24(3):865–914, 2024.
- [DRST14] I. Diakonikolas, P. Raghavendra, R. A. Servedio, and L. Y. Tan. Average sensitivity and noise sensitivity of polynomial threshold functions. *SIAM J. Comput.*, 43(1):231–253, 2014.
- [DS14] A. De and R. A. Servedio. Efficient deterministic approximate counting for low-degree polynomial threshold functions. In *Symposium on Theory of Computing, STOC 2014*, pages 832–841, 2014.

- [FGR⁺13] V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proceedings of STOC’13*, pages 655–664, 2013. Full version in Journal of the ACM, 2017.
- [GS19] N. Goyal and A. Shetty. Non-gaussian component analysis using entropy methods. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 840–851, 2019.
- [Hop18] S. Hopkins. *Statistical inference and the sum of squares method*. Cornell University, 2018.
- [Hop24] S. Hopkins. Personal communication, 2024.
- [HS17] S. B. Hopkins and D. Steurer. Efficient bayesian estimation from few samples: community detection and related problems. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 379–390. IEEE, 2017.
- [HSW24] S. Hopkins, T. Schramm, and A. Wein. Low-degree polynomial methods in average-case complexity. AIM and NSF Workshop, Pasadena, California, December 2024. Available at <https://aimath.org/workshops/upcoming/lowdegreecomplexity/>.
- [Kan11a] D. M. Kane. k-independent gaussians fool polynomial threshold functions. In *IEEE Conference on Computational Complexity*, pages 252–261, 2011.
- [Kan11b] D. M. Kane. A small prg for polynomial threshold functions of gaussians. In *FOCS*, pages 257–266, 2011.
- [Kan12] D. M. Kane. A pseudorandom generator for polynomial threshold functions of gaussian with subpolynomial seed length. *CoRR*, abs/1210.1280, 2012.
- [Kan14] D. M. Kane. The correct exponent for the gotsman-linial conjecture. *Computational Complexity*, 23(2):151–175, 2014.
- [Kea93] M. J. Kearns. Efficient noise-tolerant learning from statistical queries. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing*, pages 392–401, 1993.
- [Kea98] M. J. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.
- [KM22] Z. Kelley and R. Meka. Random restrictions and prgs for ptfs in gaussian space. In *Proceedings of the 37th Computational Complexity Conference*, pages 1–24, 2022.
- [KWB19] D. Kunisky, A. S. Wein, and A. S. Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. In *ISAAC Congress (International Society for Analysis, its Applications and Computation)*, pages 1–50. Springer, 2019.
- [MP88] M. Minsky and S. Papert. *Perceptrons: an introduction to computational geometry (expanded edition)*. MIT Press, Cambridge, MA, 1988.
- [MTT61] S. Muroga, I. Toda, and S. Takasu. Theory of majority switching elements. *J. Franklin Institute*, 271:376–418, 1961.
- [MW21] C. Mao and A. S. Wein. Optimal spectral recovery of a planted vector in a subspace. *arXiv preprint arXiv:2105.15081*, 2021.

- [MZ10] R. Meka and D. Zuckerman. Pseudorandom generators for polynomial threshold functions. In *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, pages 427–436, 2010.
- [O'D14] R. O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.
- [OST20] R. O'Donnell, R. A. Servedio, and L. Tan. Fooling gaussian ptfs via local hyperconcentration. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1170–1183, 2020.
- [Ros58] F. Rosenblatt. The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–407, 1958.
- [SKBM08] M. Sugiyama, M. Kawanabe, G. Blanchard, and K. B. Muller. Approximating the best linear unbiased estimator of non-gaussian signals with gaussian noise. *IEICE transactions on information and systems*, 91(5):1577–1580, 2008.
- [SNS16] H. Sasaki, G. Niu, and M. Sugiyama. Non-gaussian component analysis with log-density gradient estimation. In *Artificial Intelligence and Statistics*, pages 1177–1185. PMLR, 2016.
- [SW22] T. Schramm and A. S. Wein. Computational barriers to estimation from low-degree polynomials. *The Annals of Statistics*, 50(3):1833–1858, 2022.
- [Sze67] G. Szegö. *Orthogonal Polynomials*. Number τ. 23 in American Mathematical Society colloquium publications. American Mathematical Society, 1967.
- [VX11] S. S. Vempala and Y. Xiao. Structure from local optima: Learning subspace juntas via higher order pca. *arXiv preprint arXiv:1108.3329*, 2011.
- [Wei] A. S. Wein. Computational Complexity of Statistics: New Insights from Low-Degree Polynomials.
- [Wei24] A. S. Wein. Personal communication, 2024.
- [ZHT06] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

Appendix

A Additional Preliminaries

A.1 Basics of Hermite Polynomials

Hermite polynomials form a complete orthogonal basis of the vector space $L^2(\mathbb{R}, \mathcal{N}(0, 1))$ of all functions $f : \mathbb{R} \rightarrow \mathbb{R}$ such that $\mathbf{E}_{x \sim \mathcal{N}(0, 1)}[f^2(x)] < \infty$. We will use the *normalized probabilist's* Hermite polynomials, which have unit norm and are pairwise orthogonal with respect to the Gaussian measure, i.e., $\int_{\mathbb{R}} h_k(x) h_m(x) e^{-x^2/2} dx = \sqrt{2\pi} \mathbf{1}(k = m)$. These polynomials are the ones obtained by Gram-Schmidt orthonormalization of the basis $\{1, x, x^2, \dots\}$ with respect to the inner product $\langle f, g \rangle_{\mathcal{N}(0, 1)} := \mathbf{E}_{x \sim \mathcal{N}(0, 1)}[f(x)g(x)]$. Every function $f \in L^2(\mathbb{R}, \mathcal{N}(0, 1))$ can be uniquely written as $f(x) = \sum_{i=0}^{\infty} a_i h_i(x)$ and we have $\lim_{n \rightarrow \infty} \mathbf{E}_{x \sim \mathcal{N}(0, 1)}[(f(x) - \sum_{i=0}^n a_i h_i(x))^2] = 0$. We have the following closed form formula (see, e.g., [Sze67]):

$$h_n(x) = \sqrt{n!} \sum_{j=0}^{\lfloor n/2 \rfloor} \frac{(-1)^j}{j!(n-2j)!2^j} x^{n-2j}. \quad (67)$$

To extend the basis to d -dimensions, we use a multi-indices $\mathbf{J} \in \mathbb{N}^d$ to define the d -variate normalized Hermite polynomial. For $\mathbf{J} = (v_1, \dots, v_d)$ we define $h_{\mathbf{J}}(\mathbf{x}) = \prod_{i=1}^d h_{v_i}(\mathbf{x}_i)$. The total degree of $h_{\mathbf{J}}$ is $|\mathbf{J}| = \sum_{v_i \in \mathbf{J}} v_i$. Given a function $f \in L^2(\mathbb{R}^d, \mathcal{N}(\mathbf{0}, \mathbf{I}))$ we compute its Hermite coefficients as $\widehat{f}(\mathbf{J}) = \mathbf{E}_{\mathbf{x} \sim \mathcal{N}_n}[f(\mathbf{x})h_{\mathbf{J}}(\mathbf{x})]$ and express it uniquely as $\sum_{\mathbf{J} \in \mathbb{N}^n} \widehat{f}(\mathbf{J}) h_{\mathbf{J}}(\mathbf{x})$. For more details on the Gaussian space and Hermite Analysis (especially from the theoretical computer science perspective), we refer the reader to [OD14]. Most of the facts about Hermite polynomials that we use in this work are well known properties and can be found, for example, in [Sze67].

We denote by $f^{[k]}(x)$ the degree k part of the Hermite expansion of f , $f^{[k]}(\mathbf{x}) = \sum_{|\mathbf{J}|=k} \widehat{f}(\mathbf{J}) \cdot h_{\mathbf{J}}(\mathbf{x})$. We say that a polynomial q is harmonic of degree k if it is a linear combination of degree k Hermite polynomials, that is q can be written as

$$q(\mathbf{x}) = q^{[k]}(\mathbf{x}) = \sum_{\mathbf{J}: |\mathbf{J}|=k} c_{\mathbf{J}} h_{\mathbf{J}}(\mathbf{x}).$$

We will use the following fact, stating that odd degree Hermite polynomials are small around the origin.

Claim A.1 (Upper Bound on Hermite Polynomial around the Origin). *Let $\delta \in (0, 1/2)$ be such that $\delta < k^{-C}$ for some sufficiently large constant C , and $h_{\mathbf{a}}$ be a degree- k multivariate Hermite polynomial. We then have that $h_{\mathbf{a}}(\delta \mathbf{1}) < 1$.*

Proof. Consider a univariate Hermite polynomial $h_k : \mathbb{R} \mapsto \mathbb{R}$. By the explicit formula of $h_k(\delta)$ in Equation (67), it follows that the polynomial is dominated by its constant term (when k is even) or its linear term (when k is odd) when δ is a sufficiently small polynomial in its degree k . It is not hard to verify that the coefficient of the constant term or linear term is smaller than 1. It then follows that $h_k(\delta) < 1$ when δ is a sufficiently small polynomial in k . Since the multivariate Hermite polynomials are just products of many univariate Hermite polynomials, it follows that $h_{\mathbf{a}}(\delta \mathbf{1}) < 1$. \square

Another useful property of Hermite polynomials is that there exists a nice recurrence relationship between the polynomial itself and its derivative.

Fact A.2. *We have that $\frac{d}{dx} h_k(x) = k h_{k-1}(x)$.*

The fact implies the following bound on the Lipschitzness of the multivariate Hermite polynomials are around the origin.

Claim A.3 (Lipchitz continuity of Hermite Polynomials around the Origin). *Let $k \in \mathbb{Z}_+$, and $\delta \in (0, 1/2)$ be at most a sufficiently small polynomial in k , and $h_{\mathbf{a}} : \mathbb{R}^n \mapsto \mathbb{R}$ be a multivariate degree- k Hermite polynomial. Then it holds that*

$$|h_{\mathbf{a}}(\delta \mathbf{1}_n) - h_{\mathbf{a}}(\mathbf{0}_n)| \leq \sqrt{k^3 n} \delta.$$

Proof. For notational convenience, we define $h_{\mathbf{b}}(\mathbf{x})$ to be the constant 0 function when \mathbf{b} contains any negative entries.

Under this notation, [Fact A.2](#) then implies that the gradient of $h_{\mathbf{a}}$ is simply

$$\nabla h_{\mathbf{a}}(\mathbf{x}) = (\mathbf{a}_1 h_{\mathbf{a}-\mathbf{e}(1)}(\mathbf{x}), \dots, \mathbf{a}_n h_{\mathbf{a}-\mathbf{e}(n)}(\mathbf{x}))^\top,$$

where $e(i)$ is the multi-index having zeroes everywhere except from the i -th position, where it has 1. Combining the above with [Claim A.1](#) then gives that

$$\begin{aligned} \|\nabla h_{\mathbf{a}}(\delta \mathbf{1})\|_2^2 &\leq \sum_{i=1}^n k^2 h_{\mathbf{a}-\mathbf{e}(i)}^2(\delta \mathbf{1}) \\ &= \sum_{i: \mathbf{a}_i > 0} k^2 h_{\mathbf{a}-\mathbf{e}(i)}(\delta \mathbf{1})^2 \\ &\leq k^3, \end{aligned}$$

where the last inequality follows from [Claim A.1](#) and the fact that there can be at most k positive entries in \mathbf{a} . We therefore have that

$$|h_{\mathbf{a}}(\delta \mathbf{1}_n) - h_{\mathbf{a}}(\mathbf{0}_n)| \leq \|\nabla h_{\mathbf{a}}(\delta \mathbf{1})\|_2 \|\delta \mathbf{1}\|_2 \leq \sqrt{k^3 n} \delta.$$

This concludes the proof of [Claim A.3](#). □

A.2 Other Facts

Fact A.4 (Isserlis's Theorem). *Let $(x_1, \dots, x_k) \sim \mathcal{N}(0, \Sigma)$. Then,*

$$\mathbf{E}[x_1 \cdots x_k] = \sum_{p \in P_k^2} \prod_{\{i,j\} \in p} \mathbf{E}[x_i x_j],$$

where P_k^2 is the set of all matchings of $\{1, \dots, k\}$.

Lemma A.5. *There exists a smooth function $\rho : \mathbb{R} \mapsto [0, 1]$ satisfying that (1) $\rho(x) = 1$ if $|x| < 1$, (2) $\rho(x) = 0$ if $|x| \geq 3$, and $\|\rho^{(t)}(x)\|_\infty \leq t^{O(t)}$.*

Proof. Within the context of this proof, we call a function smooth if its t -th order derivative is bounded from above by $t^{O(t)}$. First, consider the function $f : \mathbb{R} \mapsto \mathbb{R}$ defined as

$$\rho_0(x) := \begin{cases} 0 & \text{if } x \geq 0 \\ \exp(-1/x^2) & \text{otherwise.} \end{cases}$$

Then we have ρ_0 is a smooth function, and $\rho_0(x) = 0$ for all $x \geq 0$. Next, define $\rho_1(x) := \rho_0(-1 + x)\rho_0(-1 - x)$. ρ_1 is still a smooth function, and we have $\rho_1(x) = 0$ if $|x| \geq 1$, and $\rho_1(x) > 0$ if $|x| < 1$. We can define a probability distribution supported on $[-1, 1]$ whose probability density function is exactly proportional to ρ_1 . Denote by ρ_2 the cumulative density function of this probability distribution. ρ_2 remains a smooth function. Furthermore, we have that $\rho_2(x) = 0$ if $x < -1$, and $\rho_2(x) = 1$ if $x > 1$. Finally, define $\rho(x) = \rho_2(2 + x)\rho_2(2 - x)$. ρ is still a smooth function, and it satisfies that $\rho(x) = 1$ if $|x| < 1$, and $\rho(x) = 0$ if $|x| > 3$. This concludes the proof of [Lemma A.5](#). □

B Separation between PTF Tests and LDP Tests

In this section, we show that having no γ -advantageous polynomials in the sense of [Definition 1.2](#) for a testing problem does not necessarily rule out the existence of a good PTF test in the sense of [Definition 1.3](#). In this section, we will work with hypothesis testing where the family of distributions for the alternative hypothesis consists of only one distribution. We restate the simplified version of γ -advantageous for simple hypothesis testing below:

Definition B.1 (γ -advantageous polynomial). *Let $\gamma > 0$, $p : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ be a degree- k , n -sample polynomial, and D_\emptyset be a distribution in \mathbb{R}^d , \mathcal{D}_{alt} be a distribution family in \mathbb{R}^d and μ be the uniform distribution over \mathcal{D}_{alt} . We say that p is a degree- k , n -sample, γ -advantageous polynomial with respect to $D_\emptyset, \mathcal{D}_{\text{alt}}$ if:*

$$\begin{aligned} & \left| \mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim D_\emptyset} [p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})] - \mathbf{E}_{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \sim D_{\text{alt}}} [p(\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)})] \right| \\ & > \gamma \max \left(\sqrt{\mathbf{Var}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim D_\emptyset} [p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})]}, \sqrt{\mathbf{Var}_{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \sim D_{\text{alt}}} [p(\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)})]} \right). \end{aligned} \quad (68)$$

We show that even in one-dimension there exist distributions $D_\emptyset, D_{\text{alt}}$ for which there is no low-degree γ -advantageous polynomial, but the hypothesis problem can be easily solved with a polynomial threshold test of degree $k = 1$ using $n = 1$ sample. The definition of the testing problem, and the statement of the claim is given below.

Definition B.2 (δ -Gap Threshold Test under ε -Gaussian Noise). *Let $\delta, \varepsilon \in (0, 1)$. We consider the hypothesis testing problem of distinguishing between the two distributions D_\emptyset and D_{alt} in \mathbb{R} defined as follows: (1) $D_\emptyset := (1 - \varepsilon)p_0 + \varepsilon\mathcal{N}(0, 1)$, where p_0 is a point mass on 0, and (2) $D_{\text{alt}} := (1 - \varepsilon)p_\delta + \varepsilon\mathcal{N}(0, 1)$, where p_δ is a point mass on δ .*

Theorem B.3. *Let $\varepsilon, \gamma, \delta \in (0, 1/2)$, and $n, k \in \mathbb{Z}_+$. Assume that $\sqrt{k^n \varepsilon^{-n} k^3 n} \delta \leq \gamma$. Then there is no degree- k , n -sample, γ -advantageous polynomial with respect to $D_\emptyset, D_{\text{alt}}$ from [Definition B.2](#). However, the 1-sample linear threshold function $\mathbb{1}\{x > \delta/2\}$ distinguishes between $D_\emptyset, D_{\text{alt}}$ with probability at least $1 - \varepsilon$.*

Proof. The fact that the linear threshold function test $\mathbb{1}\{x > \delta/2\}$ distinguishes between $D_\emptyset, D_{\text{alt}}$ with probability $1 - \varepsilon$ is immediate by the definition of the problem.

We will use the notation $D_\emptyset^{\otimes n}$ and $D_{\text{alt}}^{\otimes n}$ to denote the product distribution of n samples under the two hypotheses. Now, consider a degree- k polynomial $p : \mathbb{R}^n \mapsto \mathbb{R}$. Since [Equation \(68\)](#) is invariant to shifting and scaling of the polynomial, we can assume without loss of generality that $\mathbf{Var}_{\mathbf{x} \sim (D_\emptyset)^{\otimes n}} [p(\mathbf{x})] = 1$ and $\mathbf{E}_{\mathbf{x} \sim (D_\emptyset)^{\otimes n}} [p(\mathbf{x})] = 0$.

Under this assumption, note that we can bound from above the L_2 -norm of p under the standard Gaussian distribution as

$$\begin{aligned} \mathbf{E}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [p^2(\mathbf{x})] & \leq \varepsilon^{-n} \mathbf{E}_{\mathbf{x} \sim (D_\emptyset)^{\otimes n}} [p^2(\mathbf{x})] \\ & = \varepsilon^{-n} \mathbf{Var}_{\mathbf{x} \sim (D_\emptyset)^{\otimes n}} [p(\mathbf{x})] = \varepsilon^{-n}, \end{aligned}$$

where the first inequality follows from the fact that \mathbf{x} will be sampled from the n -dimensional standard Gaussian distribution with probability ε^n by the definition of D_\emptyset , and the last two equalities follow from our assumptions on $\mathbf{Var}_{\mathbf{x} \sim (D_\emptyset)^{\otimes n}} [p(\mathbf{x})]$ and $\mathbf{E}_{\mathbf{x} \sim (D_\emptyset)^{\otimes n}} [p(\mathbf{x})]$.

It then suffices for us to show that

$$\left| \mathbf{E}_{\mathbf{x} \sim (D_\emptyset)^{\otimes n}} [p(\mathbf{x})] - \mathbf{E}_{\mathbf{y} \sim (D_{\text{alt}})^{\otimes n}} [p(\mathbf{y})] \right| \leq \gamma, \quad (69)$$

for any polynomial p satisfying $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [p^2(\mathbf{x})] \leq \varepsilon^{-n}$. We first prove a structural claim.

Claim B.4. *Let $m \leq n$, and $q : \mathbb{R}^m \mapsto \mathbb{R}$ be an arbitrary degree- k polynomial such that $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [q^2(\mathbf{x})] \leq \varepsilon^{-n}$. Then it holds that*

$$|q(\delta \mathbf{1}_m) - q(\mathbf{0}_m)| \leq \gamma.$$

Proof. Assume that $q(\mathbf{x})$ admits the Hermite decomposition $q(\mathbf{x}) = \sum_{\mathbf{a} \in \mathbb{N}^n : |\mathbf{a}| \leq k} c_{\mathbf{a}} h_{\mathbf{a}}(\mathbf{x})$ (see [Section A.1](#) for definitions and notation regarding Hermite polynomials). By our assumption that $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [q^2(\mathbf{x})] \leq \varepsilon^{-n}$, the coefficients in the Hermite decomposition should satisfy that $\sum_{\mathbf{a} \in \mathbb{N}^n : |\mathbf{a}| \leq k} c_{\mathbf{a}}^2 \leq \varepsilon^{-n}$. Using [Claim A.3](#) and the assumption that δ is a sufficiently small polynomial in k , we have that $h_{\mathbf{a}}(\delta \mathbf{1}_m) - h_{\mathbf{a}}(\mathbf{0}_m) \leq \sqrt{k^3 m} \delta$ for all $\mathbf{a} \in \mathbb{N}^n : |\mathbf{a}| \leq k$. It follows that

$$\begin{aligned} |q(\delta \mathbf{1}_m) - p(\delta \mathbf{0}_m)| &\leq \sum_{\mathbf{a} \in \mathbb{N}^n : |\mathbf{a}| \leq k} |c_{\mathbf{a}}| |h_{\mathbf{a}}(\delta \mathbf{1}_m) - h_{\mathbf{a}}(\mathbf{0}_m)| \\ &\leq \sum_{\mathbf{a} \in \mathbb{N}^n : |\mathbf{a}| \leq k} |c_{\mathbf{a}}| \sqrt{k^3 m} \delta \\ &\leq \sqrt{k^n \varepsilon^{-n} k^3 m} \delta \leq \gamma \end{aligned}$$

where the first inequality is by the triangle inequality, the second inequality is by the bound on $h_{\mathbf{a}}(\delta \mathbf{1}_m) - h_{\mathbf{a}}(\mathbf{0}_m)$, the third inequality is by Cauchy's inequality and the bound $\sum_{\mathbf{a} \in \mathbb{N}^n : |\mathbf{a}| \leq k} c_{\mathbf{a}}^2 \leq \varepsilon^{-n}$, and the last inequality is by our assumption on $k, n, \delta, \gamma, \varepsilon$. This concludes the proof of [Claim B.4](#). \square

Given two vectors $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{z} \in \mathbb{R}^{n-m}$, and a subset of indices $S \subseteq [n]$ with cardinality $|S| = m$, we define $\mathbf{x} := \mathbf{y} \cup_S \mathbf{z}$ as the vector that has $\mathbf{x}_i = \mathbf{y}_i$ if $i \in S$ and $\mathbf{x}_i = \mathbf{z}_i$ otherwise. Given an arbitrary subset of indices $S \subseteq [n]$, we can then define the function $q_S : \mathbb{R}^{n-|S|} \mapsto \mathbb{R}$ as

$$q_S(\mathbf{y}) := \mathbf{E}_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{|S|})} [p(\mathbf{z} \cup_S \mathbf{y})].$$

It is not hard to see that q_S is a degree at most k polynomial in \mathbf{y} satisfying $\mathbf{E}_{\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n-|S|})} [q_S^2(\mathbf{y})] = \mathbf{E}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)} [p^2(\mathbf{x})]$, which is at most ε^{-n} .

Note that we can decompose the difference in [Equation \(69\)](#) by conditioning on different subsets of samples that are sampled from the Gaussian distribution. In particular, we have that

$$\begin{aligned} \left| \mathbf{E}_{\mathbf{x} \sim (D_\emptyset)^{\otimes n}} [p(\mathbf{x})] - \mathbf{E}_{\mathbf{y} \sim (D_{\text{alt}})^{\otimes n}} [p(\mathbf{y})] \right| &\leq \max_{S \subseteq [n]} \left| \mathbf{E}_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{|S|})} [p(\mathbf{z} \cup_S (\delta \mathbf{1}_{n-|S|})) - p(\mathbf{z} \cup_S \mathbf{0}_{n-|S|})] \right| \\ &= \max_{S \subseteq [n]} |q_S(\delta \mathbf{1}_{n-|S|}) - q_S(\mathbf{0}_{n-|S|})| \leq \gamma, \end{aligned}$$

where the equality is by the definition of q_S , and the last inequality follows from [Claim B.4](#). This shows [Equation \(69\)](#), and concludes the proof of [Theorem B.3](#). \square

C (Near-)Optimality of the Sample Lower Bound in [Theorem 1.6](#)

Theorem C.1. *For any $d, m \in \mathbb{Z}_+$, there exists a univariate distribution A on \mathbb{R} that matches m moments with $\mathcal{N}(0, 1)$ so that for $n \gg (C d m \log d)^{m/4}$, where C is a sufficiently large constant, and $k > 4 \log n$, there exists a degree- k polynomial $p : \mathbb{R}^{n \times d} \mapsto \{0, 1\}$ that successfully distinguishes $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathcal{M}_{A, \mathbf{v}}$ for any unit vector $\mathbf{v} \in \mathbb{R}^d$ $\text{sign}(p(\cdot))$ with constant probability: $\mathbf{E}_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \sim \mathcal{N}(0, 1)} [\text{sign}(p(\mathbf{x}^{1:n}))] < 1/10$ but $\mathbf{E}_{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)} \sim \mathcal{M}_{A, \mathbf{v}}} [\text{sign}(p(\mathbf{y}^{1:n}))] > 9/10$ for all unit vector $\mathbf{v} \in \mathbb{R}^d$.*

The basic idea is to construct a distribution A that takes values slightly larger than $d^{1/4}$ with probability almost $d^{-m/4}$ but has its first m moments matched with $\mathcal{N}(0, 1)$. If so, any sample \mathbf{y} from $\mathcal{M}_{A, \mathbf{v}}$ that witnesses the extreme values of A will have an unusually large norm compared to the Gaussian case. However, since the distribution of $\|\mathbf{x}\|_2^2$ over the standard Gaussian has its variance being approximately $d^{1/2}$, this leads to a detectable discrepancy. This forms the basis of our algorithm for distinguishing $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathcal{M}_{A, \mathbf{v}}$: the algorithm draws approximately $d^{m/4}$ many samples and simply looks for a sample with an abnormally large ℓ_2 norm. After that, we show that it is not difficult to turn this algorithm into an actual PTF test with a small loss in efficiency.

To begin with, the following proposition constructs a moment-matching distribution A that has a non-trivial amount of mass on some extreme value R .

Proposition C.2. *For any positive integer m , and real number $R > 0$, there exists a distribution A on \mathbb{R} that matches m moments with the standard Gaussian and satisfies that $\Pr[A = R] \geq R^{-m}/\text{poly}(m)$.*

Proof. Let ε be R^{-m} divided by a sufficiently large polynomial in m . We define A to be a probability distribution with the following probability density function:

$$A(x)dx = G(x)dx + (p(x)\mathbb{1}\{|x| < 1\}dx) + \varepsilon \delta_{x=R}$$

where $G : \mathbb{R} \mapsto \mathbb{R}_+$ is the probability density function of the standard Gaussian, $\delta_{x=R}$ is a point mass at $x = R$, and $p(x)$ is some degree- m polynomial we will specify later. It is clear that $\Pr_{x \sim A}[x = R] = \varepsilon$. We just need to show that there is a polynomial p which ensures that (1) $A(x)$ is non-negative, (2) $\int_{\mathbb{R}} A(x)dx = 1$, and (3) A match enough moments with the standard Gaussian $\mathcal{N}(0, 1)$:

$$\int_{\mathbb{R}} G(x)x^t dt = \int_{\mathbb{R}} A(x)x^t dt$$

for all integers $0 \leq t \leq m$. In particular, we need some polynomial p such that

$$\int_{-1}^1 p(x)x^t dt = \varepsilon R^t$$

for all such $0 \leq t \leq m$, and $p(x) + G(x) \geq 0$ for all $|x| \leq 1$. Such a polynomial can be constructed with standard techniques based on linear programming (see e.g. exercise 8.3 of [DK23]). This concludes the proof of [Proposition C.2](#). \square

We are now ready to prove [Theorem C.1](#).

Proof of Theorem C.1. Let A be the distribution given by [Proposition C.2](#) with $R = Cd^{1/4}$, where C is some sufficiently large constant multiple of $\log^{1/4}(n)$. Let $t = \lfloor \log n \rfloor$. Define the polynomial $p : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ as

$$p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) = \sum_{i=1}^n \left[\left\| \mathbf{x}^{(i)} \right\|_2^2 - d \right]^{2t}. \quad (70)$$

If the $\mathbf{x}^{(i)}$'s are independent Gaussians, we have that $\|\mathbf{x}^{(i)}\|_2^2 - 1$ is a degree-2 polynomial with L^2 norm $O(\sqrt{d})$. Therefore, by Gaussian hypercontractivity, we have that $\mathbf{E} \left[\left(\|\mathbf{x}^{(i)}\|_2^2 - 1 \right)^{2t} \right] \leq O(t)^t d^t$. In particular, this implies that

$$\mathbf{E}[p(\mathbf{x}^{(1:n)})] \leq O(t)^t d^t n.$$

Note that p is a non-negative polynomial. Applying Markov's inequality therefore gives that

$$\mathbf{Pr} \left[p(\mathbf{x}^{(1:n)}) < (C't)^t d^t \right] \geq 9/10,$$

where C' is some sufficiently large constant.

On the other hand, suppose that \mathbf{x} is drawn from $\mathcal{M}_{A,\mathbf{v}}$ conditioned on $\mathbf{v}^\top \mathbf{x} = R$ (which happens with probability at least $R^{-m}/\text{poly}(m)$). Then we immediately have that $\|\mathbf{x}\|_2^2 = R^2 + \|\mathbf{x}^\perp\|_2^2$, where \mathbf{x}^\perp is the part of \mathbf{x} that is orthogonal to \mathbf{v} , which is distributed like $\mathcal{N}(\mathbf{0}, \mathbf{I} - \mathbf{v}\mathbf{v}^\top)$. Therefore, $\mathbf{E} \left[\|\mathbf{x}\|_2^2 - d \mid \mathbf{v}^\top \mathbf{x} = R \right] = R^2 - 1$ and

$$\mathbf{Var} \left[\|\mathbf{x}\|_2^2 - d \mid \mathbf{v}^\top \mathbf{x} = R \right] = \mathbf{Var} \left[\|\mathbf{x}^\perp\|_2^2 \right] = O(d).$$

Therefore, by Chebyshev's inequality, we have

$$\mathbf{Pr} \left[\|\mathbf{x}\|_2^2 - d > R^2/2 \mid \mathbf{v}^\top \mathbf{x} = R \right] > 1/2,$$

which further implies that

$$\mathbf{Pr} \left[\|\mathbf{x}\|_2^2 - d > R^2/2 \right] > (1/2)R^{-m}/\text{poly}(m) > 10/n.$$

Thus, if $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ are drawn independently from $\mathcal{M}_{A,\mathbf{v}}$, the probability that there exists some $i \in [n]$ such that $\|\mathbf{x}^{(i)}\|_2^2 - d > R^2/2$ is at least $9/10$. However, if this happens, we immediately have that

$$p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) \geq \left(\|\mathbf{x}^{(i)}\|_2^2 - d \right)^{2t} \geq R^{4t}/2^{2t} = C^{4t}d^t/2^{2t} \geq T.$$

This shows a separation between the two cases, and completes the proof of [Theorem C.1](#). \square

D Comparison with Information-Computation Gaps for NGCA from Prior Work

D.1 Bound on Low-Degree Likelihood Ratio

The following result follows by combining [\[BBH⁺21\]](#), (which shows that a lower bound on the statistical query dimension of a hypothesis testing problem implies an upper bound on the norm of the low-degree likelihood ratio) and the statistical query dimension bound from [\[DKS17\]](#). The details of this combination can be found in [\[DKP⁺21\]](#) (Corollary 6.4, treating y as a fixed value).

Additional notation For a distribution D over \mathcal{X} , we use $D^{\otimes n}$ to denote the joint distribution of n i.i.d. samples from D . For two functions $f : \mathcal{X} \rightarrow \mathbb{R}$, $g : \mathcal{X} \rightarrow \mathbb{R}$ and a distribution D , we use $\langle f, g \rangle_D$ to denote the inner product $\mathbf{E}_{X \sim D}[f(X)g(X)]$. We use $\|f\|_D$ to denote $\sqrt{\langle f, f \rangle_D}$. We say that a polynomial $f : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}$ has sample-wise degree (r, ℓ) if each monomial uses at most ℓ different samples from $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ and uses degree at most r for each of them. Let $\mathcal{C}_{r,\ell}$ be the linear space of all polynomials of sample-wise degree (r, ℓ) with respect to the inner product defined above. For a function $f : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}$, we use $f^{\leq r, \ell}$ to be the orthogonal projection onto $\mathcal{C}_{r,\ell}$ with respect to the inner product $\langle \cdot, \cdot \rangle_{\mathcal{N}(\mathbf{0}, \mathbf{I})^{\otimes n}}$. Finally, for the null distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and a distribution $\mathcal{M}_{A,\mathbf{v}}$ from [Problem 1.5](#), define the likelihood ratio $\overline{\mathcal{M}}_{A,\mathbf{v}}^{\otimes n}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$ to be the ratio of the pdf of $\overline{\mathcal{M}}_{A,\mathbf{v}}^{\otimes n}$ on the point $(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$ divided by the pdf of $\mathcal{N}(\mathbf{0}, \mathbf{I})$ evaluated on the point $(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$. The χ^2 -distance between two distributions D and R on \mathcal{X} is defined as $\chi^2(D, R) := \int_{x \in \mathcal{X}} D^2(x)/R(x)dx - 1$.

Theorem D.1. *For any $c \in (0, 1/2)$ the following holds. There exists a subset S of the d -dimensional unit sphere for which the following hold. Let a sufficiently small positive constant c . Let $\mathcal{M}_{A,\mathbf{v}}$ denote the distribution from [Problem 1.5](#) and assume that A matches the first m moments with $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and the vector \mathbf{v} is drawn from the uniform distribution over S . For any $d \in \mathbb{Z}_+$ with $d = m^{\Omega(1/c)}$, any $n \leq \Omega(d)^{(m+1)(1/2-c)} / \chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}))$ and any even integer $\ell < d^{c/4}$, we have that*

$$\left\| \mathbf{E}_{v \sim \mathcal{U}(S)} \left[\left(\overline{\mathcal{M}}_{A,\mathbf{v}}^{\otimes n} \right)^{\leq \infty, \ell} \right] - 1 \right\|_{\mathcal{N}(\mathbf{0}, \mathbf{I})^{\otimes n}} \leq 1. \quad (71)$$

The quantity in the right hand side of [Equation \(71\)](#) is the norm of the low-degree likelihood ratio. This is an equivalent rewriting of the best possible advantage β in [Definition 1.2](#) with $D_\emptyset = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $D_{\text{alt}} = \frac{1}{|S|} \sum_{v \in S} \mathcal{M}_{A,\mathbf{v}}$. In particular the variant of that definition where the right hand side in [Equation \(68\)](#) only scales with the standard deviation under the null distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ instead of the maximum. Comparing with [Theorem 1.6](#), the three conditions of [Theorem 1.6](#) appear in some form in [Theorem D.1](#): The sample complexity condition now is $n \leq \Omega(d)^{(m+1)/(1/2-c)} / \chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}))$, which has a better constant $1/2$ in the exponent. Surprisingly, we show in [Theorem C.1](#) that the exponent from [Theorem 1.6](#) is essentially best possible, implying that PTF tests are inherently slightly more powerful than LDPs even for the NGCA problem. The condition $k < d^{\Omega(1)}$ of [Theorem 1.6](#) corresponds to the part of the statement in [Theorem D.1](#) restring $\ell < d^{c/4}$. Since the low-degree likelihood ratio uses sample-wise degree (∞, ℓ) this means that the total degree of the resulting polynomial is restricted to be at most $\ell < d^{c/4}$. Finally, the condition $d > m^{\Omega(1)}$ appears in both [Theorem 1.6](#) and [Theorem D.1](#) with different constants.

D.2 Hardness in the Statistical Query Model

In this section we restate the result from [\[DKS17\]](#) regarding the information-computation gap for NGCA within the Statistical Query model. Before we restate the theorem, we recall the basics of the SQ model [\[Kea98, FGR⁺13\]](#). Instead of drawing samples from the input distribution, SQ algorithms are only permitted query access to the distribution via the following oracle:

Definition D.2 (STAT Oracle). *Let D be a distribution in \mathbb{R}^d . A statistical query is a bounded function $f : \mathbb{R}^d \rightarrow [-1, 1]$. For $\tau > 0$, the $\text{STAT}(\tau)$ oracle responds to the query f with a value v such that $|v - \mathbf{E}_{\mathbf{x} \sim D}[f(\mathbf{x})]| \leq \tau$. We call τ the tolerance of the statistical query.*

An information-computation gap in this model for a learning problem Π is typically of the following form: any SQ algorithm for Π must either make a large number of queries q or at least one query with small tolerance τ . When simulating a statistical query in the standard PAC model (by averaging i.i.d. samples to approximate expectations), the number of samples needed for a τ -accurate query can be as high as $\Omega(1/\tau)$.

Thus, we can intuitively interpret an SQ lower bound as a tradeoff between runtime of $\Omega(q)$ or a sample complexity of $\Omega(1/\tau^2)$.

The statement for NGCA is the following.

Theorem D.3 (See, e.g., Proposition 8.14 in [DK23]). *For any constant $c \in (0, 1/2)$ and any $m, d \in \mathbb{Z}_+$ with $d \geq ((m+1) \log d)^{2/c}$ the following hold. If A is a distribution on \mathbb{R} that matches the first m moments with $\mathcal{N}(0, 1)$ and $\mathcal{M}_{A,\mathbf{v}}$ denotes the distribution from Definition 1.4, then any SQ algorithm for distinguishing between $\mathcal{M}_{A,\mathbf{v}}$ and $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (when \mathbf{v} is unknown to the algorithm) requires either $2^{\Omega(d^c)}$ many SQ queries or at least one query to STAT with accuracy $\tau \leq 2d^{-(m+1)(1/4-c/2)}\sqrt{\chi^2(A, \mathcal{N}(0, 1))}$.*

The interpretation of the above is a trade-off between exponential runtime and sample complexity at least $d^{(m+1)(1/2-c)/\chi^2(A, \mathcal{N}(0, 1))}$. As shown in [DKRS23], the dependence on $\chi^2(A, \mathcal{N}(0, 1))$ was merely an artifact of the original analysis and can be removed, at the cost of a larger constant in the exponent in the sample complexity. As one can see Theorem D.3 uses the same three assumptions as Theorem 1.6 up to differences in the constant and polylog factors. In the particular, the assumption $d \geq ((m+1) \log d)^{2/c}$ in Theorem D.3 ensures that $2^{dc/2} \geq d^{(m+1)(c-1/2)}$, i.e., both the runtime and the sample complexity are at least $d^{(m+1)(c-1/2)}$.

E Applications to Learning Mixture Models and Robust Statistics

In this section, we show that Theorem 1.6 implies strong lower bounds against PTF tests for a range of problems in machine learning theory and robust statistics.

Note that since PTF produces binary output, all the lower bounds will be for the testing version of the corresponding statistical estimation problems.

Similar to the approach taken in [DKPP24], we first define two meta testing problems that can be instantiated to model the testing version of various statistical estimation problems under the total-variation corruption model and the Hubert Contamination model.

Definition E.1 (TV-Corruption Model). *Let \mathcal{D} be a set of distributions. We define $\mathcal{B}_{TV}(\tau, \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \mathcal{D})$ to be the following hypothesis testing problem: Given n i.i.d. samples $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\} \subseteq \mathbb{R}^d$ drawn from one of the following two distributions, the goal is to determine which one generated the samples: (a) $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$; and (b) D' such that $d_{TV}(D', D) \leq \tau$ for D drawn uniformly at random from \mathcal{D} .*

Definition E.2 (Huber Contamination Model). *Let \mathcal{D} be a family of distributions. We define $\mathcal{B}_{huber}(\tau, \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \mathcal{D})$ to be the following hypothesis testing problem: Given n i.i.d. samples $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\} \subseteq \mathbb{R}^d$ drawn from one of the following two distributions, determine which one generated the samples: (a) $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$; (b) D' , which is $(1-\tau)D + \tau B$, where D is drawn uniformly at random from \mathcal{D} and B is an arbitrary distribution possibly dependent on D .*

We start with the testing version of robust mean estimation of isotropic Gaussian distribution.

Problem E.3 (Hypothesis-Testing-Robust-Mean-Estimation with Identity Covariance). *Let $\tau > 0$ and $B = O(\log^{1/2}(1/\tau))$ be a parameter. The problem is $\mathcal{B}_{TV}(\tau, \mathcal{N}(0, \mathbf{I}_d), \mathcal{D})$ (Definition E.1), where every $D \in \mathcal{D}$ is of the form $\mathcal{N}(\boldsymbol{\mu}_D, \mathbf{I}_d)$ and $\|\boldsymbol{\mu}_D\| \geq \Omega(\tau \log(1/\tau)^{1/2})/B^2$.*

It is shown in [DKS17] that the above testing problem can be reduced to NGCA whose non-Gaussian component A matches $m = B$ many moments. The lower bound against PTF tests then follows.

Corollary E.4. *Let $h : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ be a degree- k PTF. If h solves the hypothesis testing problem in Problem E.3, then we must either have $n \geq d^{B(1-c^*)/4}$ or $k \geq d^{\Omega(1)}$, where c^* is some small constant.*

The second application is on robust mean estimation of distributions with bounded m -moments. The testing version of the problem is as follows (see Section 6 from [DKK⁺22] for the justification).

Problem E.5 (Hypothesis-Testing-Robust-Mean-Estimation with Bounded m -th Moments). *Let m be a positive integer and $\tau \in (0, 1)$. Hypothesis-Testing-RME-Bounded- m -Moments is the problem $\mathcal{B}_{\text{huber}}(\tau, \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \mathcal{D})$ (**Definition E.2**) where each $D \in \mathcal{D}$ satisfies the following: (i) the mean vector μ satisfies $\|\mu\| \geq \Omega(\frac{1}{m}\tau^{1-1/m})$; (ii) D has subgaussian tails of order m , i.e., for all $\mathbf{v} \in \mathbb{R}^d$ and $1 \leq i \leq m$, $\mathbf{E}_{\mathbf{x} \sim D}[|\mathbf{v}^\top (\mathbf{x} - \mu)|^i]^{1/i} \leq O(\sqrt{i})$.*

It is shown in [DKK⁺22] that the problem can be reduced to NGCA whose non-Gaussian component A matches m many moments with the standard Gaussian. We therefore obtain the following lower bound against PTF tests.

Corollary E.6. *Let $h : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ be a degree- k PTF. If h solves the hypothesis testing problem in Problem E.5, then we must either have $n \geq d^{m(1-c^*)/4}$ or $k \geq d^{\Omega(1)}$, where c^* is some small constant.*

The third application is on Gaussian list-decodable mean estimation. The testing version is given below. See [DKS18b] for a thorough walkthrough of this problem, and the reduction between its learning version and the testing version.

Problem E.7 (Hypothesis-Testing-List-Decodable-Mean-Estimation). *Given $\tau \in (0, \frac{1}{2})$ and positive integer $m \geq 2$, the hypothesis-testing-LDME is the problem $\mathcal{B}_{\text{huber}}(1-\tau, \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \mathcal{D})$ (**Definition E.2**), where every $D \in \mathcal{D}$ has the form $\mathcal{N}(\mu_D, \mathbf{I}_d)$ for some $\mu_D \in \mathbb{R}^d$ whose ℓ_2 -norm is at least $\Omega((m\tau)^{-1/m})$.*

It is shown in [DKS18b] that this problem can be reduced to NGCA whose non-Gaussian component matches m moments with the standard Gaussian. We hence obtain the following lower bound against PTF tests.

Corollary E.8. *Let $h : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ be a degree- k PTF. If h solves the hypothesis testing problem in Problem E.7, then we must either have $n \geq d^{m(1-c^*)/4}$ or $k \geq d^{\Omega(1)}$, where c^* is some small constant.*

The last application is on learning mixtures of k Gaussians. It is shown in [DKS17] that this learning problem can be reduced from the following testing problem.

Problem E.9 (Hypothesis-Testing- m -GMM). *Let $0 < \gamma < 1$. Hypothesis-Testing- m -GMM is the problem $\mathcal{B}_{\text{huber}}(0, \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \mathcal{D})$ (**Definition E.2**), where every $D \in \mathcal{D}$ is a mixture of m Gaussians such that each pair of the Gaussians are $1 - \gamma$ apart in total variation and $d_{\text{TV}}(D, \mathcal{N}(\mathbf{0}, \mathbf{I}_d)) \geq \frac{1}{2}$.*

It has been shown in the same work that the testing problem can be further reduced from NGCA whose non-Gaussian component matches $2m - 1$ moments with $\mathcal{N}(0, 1)$. We therefore obtain the following lower bound.

Corollary E.10. *Let $h : \mathbb{R}^{n \times d} \mapsto \mathbb{R}$ be a degree- k PTF. If h solves the hypothesis testing problem in Problem E.9, then we must either have $n \geq d^{m(1-c^*)/2}$ or $k \geq d^{\Omega(1)}$, where c^* is some small constant.*