SQ Lower Bounds for Learning Mixtures of Separated and Bounded Covariance Gaussians

Ilias Diakonikolas ILIAS@CS.WISC.EDU
University of Wisconsin Madison

Daniel M. Kane DAKANE@CS.UCSD.EDU
University of California, San Diego

Thanasis Pittas PITTAS@WISC.EDU
University of Wisconsin-Madison

Nikos Zarifis ZARIFIS@WISC.EDU
University of Wisconsin-Madison

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We study the complexity of learning mixtures of separated Gaussians with common unknown bounded covariance matrix. Specifically, we focus on learning Gaussian mixture models (GMMs) on $\mathbb{R}^d$ of the form $P = \sum_{i=1}^k w_i \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma} \preceq \mathbf{I}$ and $\min_{i \neq j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2 \geq k^\epsilon$ for some $\epsilon > 0$. Known learning algorithms for this family of GMMs have complexity $(dk)^{O(1/\epsilon)}$. In this work, we prove that any Statistical Query (SQ) algorithm for this problem requires compexity at least $d^{\Omega(1/\epsilon)}$. Our SQ lower bound implies a similar lower bound for low-degree polynomial tests. Our result provides evidence that known algorithms for this problem are nearly best possible.

Keywords: Gaussian mixtures, Statistical Query model, low-degree polynomial tests

1. Introduction

We study the classical problem of learning Gaussian mixture models (GMMs) in high dimensions. This problem has a long history, starting with the early work of Pearson Pearson (1894) who introduced the method of moments in this context. Over the past three decades, there has been a vast literature on learning GMMs in both statistics and theoretical computer science Dasgupta (1999); Arora and Kannan (2001); Vempala and Wang (2002); Achlioptas and McSherry (2005); Feldman et al. (2006); Kannan et al. (2008); Brubaker and Vempala (2008); Moitra and Valiant (2010); Belkin and Sinha (2010); Suresh et al. (2014); Daskalakis and Kamath (2014); Hardt and Price (2015); Diakonikolas et al. (2020a); Bakshi et al. (2020); Diakonikolas et al. (2022b); Liu and Moitra (2021); Bakshi et al. (2022). Here we focus on computational aspects of this problem with a focus on *information-computation tradeoffs* in high dimensions.

The learning setup is as follows: We have access to i.i.d. samples from an unknown k-GMM on $\mathbb{R}^d$, $P = \sum_{i=1}^k w_i \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $w_i \geq 0$ are the mixing weights satisfying $\sum_{i=1}^k w_i = 1$, $\boldsymbol{\mu}_i \in \mathbb{R}^d$ are the unknown component means and $\boldsymbol{\Sigma}_i$ are the unknown component covariances. Roughly speaking, there are two versions of the learning problem: (1) density estimation, where the goal is to compute a hypothesis distribution H that is close to P in total variation distance, and (2) parameter estimation[1], where the goal is to approximate the target parameters $w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ within

[1] A related task is that of clustering the sample based on the generating component. Once we have an accurate clustering, assuming one exists, we can individually learn the individual component parameters.

small error. While density estimation of k-GMMs on $\mathbb{R}^d$ is information-theoretically solvable with $\mathrm{poly}(d, k)$ samples, parameter estimation may require $2^{\Omega(k)}$ samples (even in one dimension) if the individual components are close to each other Moitra and Valiant (2010). On the other hand, under the standard separation assumption that the components are "nearly non-overlapping", parameter estimation can also be solved with $\mathrm{poly}(d, k)$ samples. Here we focus on families of instances satisfying appropriate separation assumptions. Even though such instances can be learned with $\mathrm{poly}(d, k)$ samples, it is by no means clear that a $\mathrm{poly}(d, k)$-*time* learning algorithm exists. In other words, we explore the relevant *information-computation tradeoffs* — inherent tradeoffs between the sample complexity and the computational complexity of learning.

A number of recent works have established information-computation tradeoffs in the context of learning GMMs. The first such result was given in Diakonikolas et al. (2017) and applied to the class of Statistical Query (SQ) algorithms[2]. Specifically, Diakonikolas et al. (2017) constructed a hard family of GMMs (henceforth informally termed as "parallel pancakes") and showed that any SQ learner for this family requires super-polynomial time. Interestingly, the class of parallel pancakes is learnable with $O(k \log d)$ samples, while any SQ learning algorithm requires $d^{\Omega(k)}$ time. It is worth noting that subsequent work Bruna et al. (2021); Gupte et al. (2022) established computational hardness for essentially the same class of instances, under widely-believed cryptographic assumptions.

In this work, we focus on a simpler and well-studied family of GMMs for which significantly faster learning algorithms are known. (We provide a detailed comparison between the family of instances we consider and the parallel pancakes construction of Diakonikolas et al. (2017) in Section 1.2.) Specifically, we consider GMMs of the form $P = \sum_{i=1}^{k} w_i \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, satisfying (a) $\min_i w_i \geq 0.9/k$, (b) $\boldsymbol{\Sigma}_i \preceq \mathbf{I}$, and (c)$\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2 \geq k^{\epsilon}$, for some $\epsilon > 0$. Condition (a) posits that the component weights are nearly uniform. (This first condition is relevant for the clustering/parameter estimation problems, as these tasks require $\Omega(1/\min_i w_i)$ samples.) Condition (b) says that each component covariance is unknown and bounded above by the identity. Finally, condition (c) requires that the component means are pairwise separated by at least k^{ϵ}, in ℓ_2-distance. Here the parameter $\epsilon > 0$ is assumed to be sufficiently large so that $k^{\epsilon} \gg \sqrt{\log k}$. This assumption is required as, even for the uniform weights and identity covariance case (i.e., when $w_i = 1/k$ and $\boldsymbol{\Sigma}_i = \mathbf{I}$ for all i), the clustering problem can be solved with $\mathrm{poly}(d, k)$ samples if and only if the pairwise mean separation is $\Delta \gg \sqrt{\log k}$ Regev and Vijayaraghavan (2017).

It is easy to see that the aforementioned family of GMMs is learnable using $\mathrm{poly}(d, k)$ samples (ignoring computational considerations). Two independent works Hopkins and Li (2018); Kothari et al. (2018) gave SoS-based learning algorithms for this family of GMMs with sample complexity $k^{O(1)} d^{O(1/\epsilon)}$ and computational complexity $(dk)^{O(1/\epsilon^2)}$. With a more careful analysis, the runtime can be further improved to $(dk)^{O(1/\epsilon)}$ Steurer and Tiegel (2021); Diakonikolas et al. (2022a). Note that for the important special case that the mean separation is $\Delta \gg \log^c(k)$, for some constant $c \geq 1/2$, these algorithms have quasi-polynomial sample and time complexities, namely $(dk)^{O(\log k)}$.

A natural question is whether the aforementioned upper bounds are inherent or can be significantly improved. Concretely, we address the following open problem:

Is there a $\mathrm{poly}(d, k)$*-time learning algorithm for separated GMMs*
with bounded covariance components and mean separation $\Delta = \mathrm{polylog}(k)$*?*

[2] Via a recent reduction Brennan et al. (2021), these SQ lower bounds imply qualitatively similar low-degree testing lower bounds.

For the special case of *spherical* components, namely when each individual Gaussian has identity covariance (i.e., $\mathbf{\Sigma}_i = \mathbf{I}$ for all i), very recent work Li and Liu (2022) made significant algorithmic progress on this question. Specifically, they gave a $\mathrm{poly}(d, k)$ time learning algorithm that succeeds as long as $\Delta \gg \log^{1/2+c}(k)$, for any constant $c > 0$. The algorithm in Li and Liu (2022) crucially leveraged the assumption that the individual components are known (and equal to the identity). On the other hand, their upper bound raised the hope that $\mathrm{poly}(d, k)$ complexity might be attainable even for unknown bounded covariance components with similar mean separation.

In this work, we provide evidence that known learning algorithms Hopkins and Li (2018); Kothari et al. (2018); Steurer and Tiegel (2021); Diakonikolas et al. (2022a) for this subclass of GMMs are qualitatively best possible. Concretely, we prove an SQ lower bound for this family of GMMs suggesting the following information-computation tradeoff: For mean separation $\Delta = k^{\epsilon}$, any (SQ) learning algorithm either requires $2^{d^{\Omega(1)}}$ time or uses at least $d^{\Omega(1/\epsilon)}$ samples. In particular, this implies that the quasi-polynomial upper bounds for mean separation of $\Delta = \mathrm{polylog}(k)$ are best possible for the class of SQ algorithms. Using known results Brennan et al. (2021), this SQ lower bound implies a qualitatively similar low-degree testing lower bound.

We also provide an interesting implication for the special case of $\epsilon = 1/2$. Specifically, we establish an SQ lower bound suggesting that any efficient SQ algorithm under separation $\Delta \ll k^{1/2}$ requires nearly *quadratically* many samples (in the dimension d). On the other hand, $O(kd)$ samples suffice without computational limitations. Recent work Diakonikolas et al. (2022b) developed an $O(dk)$-sample and computationally efficient algorithm for learning bounded covariance distributions (and, consequently, bounded covariance Gaussians) under separation $\tilde{\Omega}(k^{1/2})$. A natural open question is whether this separation bound can be significantly improved *while preserving sample near-optimality*. Perhaps surprisingly, we show that this is not possible for SQ algorithms: any efficient SQ algorithm that works for separation $Ck^{1/2}$, for a sufficiently small constant C, requires near-quadratically many samples in d. This gap suggests that the algorithm of Diakonikolas et al. (2022b) succeeds under the best possible separation within the class of computationally efficient and sample near-optimal algorithms.

1.1. Our Results

Our main result is a Statistical Query lower bound of $d^{\Omega(1/\epsilon)}$ for learning the aforementioned subclass of Gaussian mixtures with mean separation $\Delta \geq k^{\epsilon}$.

Before we formally state our contributions, we require basic background on the SQ model.

SQ Model Basics Before we state our main result, we recall the basics of the SQ model Kearns (1998); Feldman et al. (2013). Instead of drawing samples from the input distribution, SQ algorithms are only permitted query access to the distribution via the following oracle:

Definition 1 (VSTAT Oracle) *Let D be a distribution on $\mathbb{R}^d$. A statistical query is a bounded function $q : \mathbb{R}^d \to [0, 1]$. For $u > 0$, the $\mathrm{VSTAT}(u)$ oracle responds to the query q with a value v such that $|v - \mathbf{E}_{\mathbf{x}\sim D}[q(\mathbf{x})]| \leq \tau$, where $\tau = \max\{1/u, \sqrt{\mathrm{Var}_{\mathbf{x}\sim D}[q(\mathbf{x})]/u}\}$. We call τ the* tolerance *of the statistical query.*

An SQ lower bound 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

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

Main Result Our main SQ lower bound result for learning GMMs is stated informally below. A more detailed formal version is provided in Theorem 7.

Theorem 2 (Main Result, Informal) *For $d, k \in \mathbb{Z}_+$ sufficiently large and $\epsilon > 0$ such that $k^\epsilon \gg \sqrt{\log k}$, any SQ algorithm that correctly distinguishes between $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and a k-GMM on $\mathbb{R}^d$ with minimum mixing weight at least $0.99/k$, common covariance $\mathbf{\Sigma} \preceq \mathbf{I}_d$ for each component, and pairwise mean separation $\Delta \geq k^\epsilon$, either makes $2^{d^{\Omega(1)}}$ statistical queries or requires at least one query to* $\mathrm{VSTAT}(d^{\Omega(1/\epsilon)})$.

As is typically the case, our SQ lower bound applies for the hypothesis testing problem of distinguishing between the standard Gaussian and an unknown GMM in our family. Hardness for testing a fortiori implies hardness for the corresponding learning problem (see Corollary 8).

A few additional remarks are in order. First notice that our SQ lower bound applies even for the special case where the mixing weights are nearly uniform (within a factor of 2, say) and the component covariances are the same, as long as they are unknown[3]. As it will become clear from our construction, the common covariance matrix of each component has only two distinct eigenvalues: each Gaussian component behaves like a standard Gaussian in all directions that are orthogonal to a low-dimensional subspace, and along that subspace behaves like a spherical Gaussian with different variance. Finally, our lower bound applies for a large range of the parameter $\epsilon > 0$, as long as k^ϵ is at least a sufficiently large constant multiple of $\sqrt{\log k}$. Consequently, it implies that the quasi-polynomial upper bounds for separation of $\mathrm{polylog}(k)$ are best possible SQ algorithms.

The implications of our SQ lower bound to the low-degree polynomial testing model, via the result of Brennan et al. (2021), are provided in Appendix D.

Quadratic SQ Lower Bound for $\Omega(\sqrt{k})$ Separation Our second result concerns the special case where the mean separation is proportional to $k^{1/2}$, namely $Ck^{1/2}$ for a sufficiently small universal constant C (taking $C = 1/3$ suffices for our purposes). For this setting, we establish a nearly quadratic tradeoff between the sample complexity of the learning problem and the sample complexity of any efficient SQ algorithm for the problem. Specifically, we show the following:

Theorem 3 (Quadratic SQ Lower Bound, Informal) *Let $d, k \in \mathbb{Z}_+$ with d sufficiently large and $2 \leq k \ll \log d$. Any SQ algorithm that correctly distinguishes between $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and a k-GMM on $\mathbb{R}^d$ with uniform weights, common covariance $\mathbf{\Sigma} \preceq \mathbf{I}_d$ for each component, and pairwise mean separation $\Delta \geq \sqrt{k}/3$, either makes $2^{d^{\Omega(1)}}$ statistical queries or requires at least one query to* $\mathrm{VSTAT}(d^{1.99})$.

A more detailed formal version is provided in Theorem 11. The natural interpretation of the above result is as follows: any SQ algorithm for this class of instances either uses $\Omega(d^{1.99})$ many samples or requires at least $2^{d^{\Omega(1)}}$ many statistical queries (time). On the other hand, without computational constraints, $O(kd)$ samples information-theoretically suffice.

Using different techniques, Davis et al. (2021) established a low-degree testing lower bound for the $k = 2$ case with constant separation, suggesting a sample complexity tradeoff of $\tilde{\Omega}(d^2)$.

[3] Recall that known algorithms do not require these assumptions. The runtime upper bound of $(dk)^{O(1/\epsilon)}$ holds as long as the minimum weight is at least $1/\mathrm{poly}(k)$ and even if the component covariances are different.

1.2. Overview of Techniques

The best comparison to our results is the prior work of Diakonikolas et al. (2017). Both works prove SQ lower bounds for learning mixtures of separated, common covariance Gaussians. The major difference is that the Diakonikolas et al. (2017) result requires large separation relative to the *smallest* eigenvalue of the covariance (or, more accurately, relative to the quadratic form defined by the inverse covariance matrix), while our result requires large separation relative to the *largest* eigenvalue. As we will see, this seemingly small distinction leads to significant differences.

Underlying both SQ lower bound results is the hidden-direction non-Gaussian component analysis construction of Diakonikolas et al. (2017) (or, in our case, the generalization to hidden *subspaces* given in Diakonikolas et al. (2021b)). The high-level idea is that if one can find a distribution A (defined in a small number of dimensions) that matches its first t moments with the standard Gaussian, then distinguishing the standard Gaussian from a distribution D that behaves like A along a hidden subspace and is standard Gaussian in the orthogonal directions requires SQ complexity $d^{\Omega(t)}$. This generic result has been leveraged to establish SQ lower bounds for a wide range of high-dimensional statistical tasks, see, e.g., Diakonikolas et al. (2017, 2019, 2020b); Goel et al. (2020); Diakonikolas and Kane (2022); Diakonikolas et al. (2021b, 2022c, 2018, 2021a, 2020c); Chen et al. (2022). The main difficulty in each case is, of course, to construct the desired moment-matching distributions.

In our context, this means that for either result one needs to exhibit a distribution A, which is a mixture of k separated Gaussians, so that A matches many moments with the standard Gaussian. By letting A be a discrete distribution with support size k convolved with a narrow Gaussian, it suffices to find a distribution A' supported on k pairwise separated points so that A' matches t moments with a standard Gaussian.

At this point, the difference in the underlying separation assumptions becomes critical. In the parallel pancakes construction of Diakonikolas et al. (2017), one only needs the points in the support of A' to have some minimal separation so that after convolving with a very narrow Gaussian, the resulting components of A are still well separated in total variation distance. This fact allows them to use Gaussian quadrature and construct a *one-dimensional* distribution A' which matches its first $t = 2k$ moments with $\mathcal{N}(0, 1)$. This construction leads to an SQ lower bound of $d^{\Omega(k)}$. It should be noted that each unknown GMM in this old construction consists of k "skinny" Gaussians whose mean vectors all lie in the same direction. Moreover, each pair of components will have total variation distance very close to 1 and their mean vectors are separated by $\Omega(1/\sqrt{k})$.

In our setting however, we require much stronger separation assumptions. In particular, we require that the elements in the support of A' be separated by some relatively large separation Δ on the order of $k^{\epsilon} \gg \sqrt{\log(k)}$. Unfortunately, it is provably impossible to find a moment-matching construction with this kind of separation in one dimension. Intuitively, this holds because the standard Gaussian $G \sim \mathcal{N}(0, 1)$ is highly concentrated about the origin. If A' behaves similarly to G, it must also have most of its mass near the origin; but this is clearly impossible if the points of its support are pairwise separated by Δ. More rigorously, one can show that the indicator function of an interval can be reasonably well-approximated by a constant-degree polynomial with respect to the Gaussian distribution (see, e.g, Diakonikolas et al. (2010)). This implies that any distribution over $\mathbb{R}$ that matches constantly many moments with G must be relatively close to G in Kolmogorov distance, which is impossible for any discrete distribution with a widely separated support.

To circumvent this issue, we instead produce a distribution A' over $\mathbb{R}^m$, for some m on the order of Δ^2 (Proposition 9). Intuitively, this makes sense because Gaussian random points on $\mathbb{R}^m$ have pairwise separation approximately $\sqrt{m} = \Delta$; this motivates us to use points drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ to construct the support of A' (see Proposition 10, we will describe the construction in more detail in the next paragraph). Unfortunately, this choice comes with a tradeoff. As the dimension of the space of degree-t polynomials on $\mathbb{R}^m$ is approximately m^t, we will need the support of A' to be of size roughly m^t in order to have enough degrees of freedom to be able to match all of these moments. In particular, this means that the parameter k needs to be on the order of Δ^{2t}, and since we are considering separation $\Delta = k^\epsilon$, we need to choose t to be on the order of $1/\epsilon$. Thus, the resulting SQ lower bound will be on the order of $d^{\Omega(t)} = d^{\Omega(1/\epsilon)}$. Note that we cannot hope to do better, as the algorithms of Hopkins and Li (2018); Kothari et al. (2018) can be formalized as SQ algorithms with similar complexity.

It remains to explain how to construct A'. We want a distribution over a small support that matches t moments with the standard Gaussian over $\mathbb{R}^m$ and also has large pairwise separation of its support points. The simple idea behind our construction is that picking a uniformly random set of points as our support should both ensure the separation with high probability, and also produce a set that is well-representative of a Gaussian. We achieve this as follows: we pick an appropriate number of i.i.d. Gaussian random points in $\mathbb{R}^m$ and, using linear programming duality, show that with high probability there exists a moment-matching distribution supported on these points (cf. Proposition 10).

For the case $\epsilon = 1/2$ (corresponding to pairwise mean separation of $\sim \sqrt{k}$), the above analysis is suboptimal because it shows an SQ lower bound of $d^{\Omega(1/\epsilon)}$ with the constant inside the big-Ω being rather large. In order to obtain a quadratic SQ lower bound for that case, we instead provide an explicit distribution over $\mathbb{R}^m$ matching three moments with the standard Gaussian (cf. Section 5).

2. Preliminaries

We record the minimum preliminaries necessary for the main body of this paper, with the full version being provided in Appendix A.

Basic Notation We use $\mathbb{Z}_+$ for positive integers, $[n] \overset{\text{def}}{=} \{1, \ldots, n\}$ and $\|\mathbf{v}\|_2$ for the ℓ_2-norm of a vector $\mathbf{v}$. We use $\mathbf{I}_d$ to denote the $d \times d$ identity matrix. For a matrix $\mathbf{A}$, we use $\|\mathbf{A}\|_\mathrm{F}$ and $\|\mathbf{A}\|_\mathrm{op}$ to denote the Frobenius and spectral (or operator) norms respectively. We use $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For a set S, we use $\mathcal{U}(S)$ for the uniform distribution on S. We use $\phi_m(\mathbf{x})$ for the pdf of the standard Gaussian in m-dimensions $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, and $\phi(x)$ the pdf of $\mathcal{N}(0,1)$. Slightly abusing notation, we will use the same letter for a distribution and its pdf, e.g., we will denote by $P(\mathbf{x})$ the pdf of a distribution P.

Hermite Analysis We use h_k for the normalized probabilist's Hermite polynomials, which comprise a complete orthogonal basis of all functions $f : \mathbb{R} \to \mathbb{R}$ with $\mathbf{E}_{x\sim\mathcal{N}(0,1)}[f^2(x)] < \infty$. When using multi-indices $\mathbf{a} \in \mathbb{Z}^d$ as subscripts, we refer to the multivariate Hermite polynomials.

Ornstein-Uhlenbeck Operator For a $\rho \in (0,1)$, we define the *Ornstein-Uhlenbeck* (or *Gaussian noise*) operator U_ρ as the operator that maps a distribution F on $\mathbb{R}^m$ to the distribution of the random variable $\rho\mathbf{x} + \sqrt{1-\rho^2}\mathbf{z}$, where $\mathbf{x} \sim F$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ independently of $\mathbf{x}$. A standard property of the U_ρ operator is that it operates diagonally with respect to Hermite polynomials, i.e., $\mathbf{E}_{\mathbf{x}\sim U_\rho F}[h_\mathbf{a}(\mathbf{x})] = \rho^{|\mathbf{a}|}\, \mathbf{E}_{\mathbf{x}\sim F}[h_\mathbf{a}(\mathbf{x})]$, where $|\mathbf{a}| = \sum_i a_i$.

2.1. Background on the Statistical Query Model

We record the definitions from the SQ framework of Feldman et al. (2013) that we will need: For a distribution D and a family of distributions $\mathcal{D}$, we define the *decision problem over distributions* $\mathcal{B}(\mathcal{D}, D)$ as the hypothesis testing problem of distinguishing between D and a member of $\mathcal{D}$. We define the *pairwise correlation* between two distributions D_1 and D_2 as $\chi_D(D_1, D_2) = \int_{\mathbb{R}^d} D_1(\mathbf{x})D_2(\mathbf{x})/D(\mathbf{x})\,\mathrm{d}\mathbf{x}-1$. We say that a set of s distributions $\mathcal{D}=\{D_i\}_{i=1}^s$ is (γ, β)-correlated relative to a distribution D if $|\chi_D(D_i, D_j)| \leq \gamma$ for all $i \neq j$, and $|\chi_D(D_i, D_j)| \leq \beta$ for $i = j$.

Definition 4 (Statistical Query Dimension) *Let $\beta, \gamma > 0$. Consider a decision problem $\mathcal{B}(\mathcal{D}, D)$, where D is a fixed distribution and $\mathcal{D}$ is a family of distributions. Define s to be the maximum integer such that there exists a finite set of distributions $\mathcal{D}_D \subseteq \mathcal{D}$ such that $\mathcal{D}_D$ is (γ, β)-correlated relative to D and $|\mathcal{D}_D| \geq s$. The* Statistical Query dimension *with pairwise correlations (γ, β) of $\mathcal{B}$ is defined to be $\mathrm{SD}(\mathcal{B}, \gamma, \beta) := s$.*

Lemma 5 (Corollary 3.12 in Feldman et al. (2013)) *Let $\mathcal{B}(\mathcal{D}, D)$ be a decision problem. For $\gamma, \beta > 0$, let $s = \mathrm{SD}(\mathcal{B}, \gamma, \beta)$. For any $\gamma' > 0$, any SQ algorithm for $\mathcal{B}$ requires queries of tolerance at most $\sqrt{\gamma + \gamma'}$ or makes at least $s\gamma'/(\beta - \gamma)$ queries.*

Lemma 6 (Corollary 2.4 in Diakonikolas et al. (2021b)) *Let A be a distribution over $\mathbb{R}^m$ such that the first t moments of A match the corresponding moments of $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$. Let $G(\mathbf{x})=A(\mathbf{x})/\phi_m(\mathbf{x})$ be the ratio of the corresponding probability density functions. For matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{m\times d}$ such that $\mathbf{U}\mathbf{U}^\top = \mathbf{V}\mathbf{V}^\top = \mathbf{I}_m$, define $P_{A,\mathbf{U}}$ and $P_{A,\mathbf{V}}$ to be distributions over $\mathbb{R}^d$ with probability density functions $G(\mathbf{U}\mathbf{x})\phi_d(\mathbf{x})$ and $G(\mathbf{V}\mathbf{x})\phi_d(\mathbf{x})$, respectively. Then, the following holds: $|\chi_{\mathcal{N}(\mathbf{0},\mathbf{I}_m)}(P_{A,\mathbf{U}}, P_{A,\mathbf{V}})| \leq \|\mathbf{U}\mathbf{V}^\top\|_{\mathrm{op}}^{t+1}\chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}_m))$.*

Note that in the statement above, $P_{A,\mathbf{V}}$ can be rewritten in the following form:

$$P_{A,\mathbf{V}}(\mathbf{x}) = A(\mathbf{V}\mathbf{x})\frac{\phi_d(\mathbf{x})}{\phi_m(\mathbf{V}\mathbf{x})} = A(\mathbf{V}\mathbf{x})(2\pi)^{-\frac{(d-m)}{2}}e^{-\frac{1}{2}\|\mathbf{x}-\mathbf{V}^\top\mathbf{V}\mathbf{x}\|_2^2} = A(\mathbf{V}\mathbf{x})\phi_{d-m}\left(\mathrm{Proj}_{\mathcal{V}^\perp}(\mathbf{x})\right)\ , \tag{1}$$

where $\mathrm{Proj}_{\mathcal{V}^\perp}(\mathbf{x}) = \mathbf{x} - \mathbf{V}^\top\mathbf{V}\mathbf{x}$ is the projection of $\mathbf{x}$ to the subspace that is perpendicular to the subspace $\mathcal{V}$ spanned by the rows of $\mathbf{V}$. Therefore, Equation (1) demonstrates that $P_{A,\mathbf{V}}$ coincides with the distribution A in the subspace spanned by the rows of $\mathbf{V}$ and is standard Gaussian in every orthogonal direction.

3. Statistical Query Lower Bound

In this section we prove the following mored detailed version of our main result (Theorem 2).

Theorem 7 (SQ Lower Bound: Hypothesis Testing Hardness) *Let $d, k \in \mathbb{Z}_+, \epsilon > 0$ and C be a sufficiently large absolute constant. Assume $k > (C/\epsilon)^{1/\epsilon}$, $d > k^{C\epsilon}$, and $k^\epsilon > C\sqrt{\log k}$. Let the following hypothesis testing problem regarding a distribution P on $\mathbb{R}^d$:*

- *(Null Hypothesis) $P = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.*
- *(Alternative Hypothesis) P belongs to a family $\mathcal{P}$, every member of which is a mixture of Gaussians $\sum_{i=1}^k w_i\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{\Sigma})$ for some weights $w_i > 0.99/k$, mean vectors with pairwise separation $\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_2 \geq k^\epsilon$ for all distinct $i, j \in [k]$, and common covariance matrix $\mathbf{\Sigma} \preceq \mathbf{I}_d$. Moreover, $\mathrm{d}_{\mathrm{TV}}(P, \mathcal{N}(\mathbf{0}, \mathbf{I}_d)) > 0.99$ and $\mathrm{d}_{\mathrm{TV}}(P, P') > 0.99$ for all distinct $P, P' \in \mathcal{P}$.*

Any algorithm with statistical query access to P that distinguishes correctly between the two cases, does one of the following: Performs $2^{d^{\Omega(1)}}$ statistical queries, or performs at least one statistical query with tolerance $d^{-\Omega(1/\epsilon)}e^{O(k^{2\epsilon})}$.

Before moving to the proof, we state the implications of the above to the hardness of the corresponding density estimation problem.

Corollary 8 (SQ Lower Bound: Density Estimation Hardness) *Under the assumptions of Theorem 7 and the additional assumption $k^\epsilon < \sqrt{\log(d)/(C\epsilon)}$, let $\mathcal{A}$ be an SQ algorithm that given access to a mixture of Gaussians $P = \sum_{i=1}^k w_i \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ for some unknown weights $w_i > 0.99/k$, mean vectors $\boldsymbol{\mu}_i \in \mathbb{R}^d$ for $i \in [k]$ with pairwise separation $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2 \geq k^\epsilon$ and common covariance matrix $\boldsymbol{\Sigma} \preceq \mathbf{I}_d$, finds a distribution Q with $\mathrm{d}_{\mathrm{TV}}(P, Q) < 1/4$. Then $\mathcal{A}$ necessarily does one of the following: Performs $2^{d^{\Omega(1)}}$ statistical queries, or performs at least one statistical query with tolerance $\tau = d^{-\Omega(1/\epsilon)}e^{O(k^{2\epsilon})}$.*

Proof The reduction from the hypothesis testing problem of Theorem 7 to the corresponding learning problem is fairly standard, see e.g., Lemma 8.5 in Diakonikolas and Kane (2023). To check the applicability of that lemma we note that $\mathrm{d}_{\mathrm{TV}}(P, \mathcal{N}(\mathbf{0}, \mathbf{I}_d)) > 0.99 > 2(\tau+1/4)$, where the inequality uses the assumption $k^\epsilon < \sqrt{\log(d)/(C\epsilon)}$ for bounding the query tolerance τ by a constant. ■

The main ingredient towards proving Theorem 7 is Proposition 9, which establishes the existence of a low-dimensional spherical k-GMM with well separated means, that matches its first $\Omega(1/\epsilon)$ moments with the standard Gaussian. We prove this result in Section 4. In this section, we show how Theorem 7 follows from Proposition 9.

Proposition 9 *Let $\epsilon > 0$, $d, k \in \mathbb{Z}_+$, $c > 0$ be a sufficiently small constant and C be a sufficiently large constant. If $k > (C/\epsilon)^{1/\epsilon}$, $d > k^{C\epsilon}$, and $k^\epsilon > C\sqrt{\log k}$ there exists a distribution A over $\mathbb{R}^m$ with $m := k^{2\epsilon}$ that satisfies the following:*

(i) A is a mixture of k spherical Gaussians in $\mathbb{R}^m$ with variance $\delta = ck^{-2.5/m}$ and minimum mixing weight at least $0.99/k$.

(ii) A matches its first $t = \Theta(1/\epsilon)$ moments with $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$.

(iii) The means $\boldsymbol{\mu}_i, \boldsymbol{\mu}_j$ of any two distinct components have separation $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2 \geq k^\epsilon$.

(iv) For every $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{m \times d}$ with $\mathbf{U}\mathbf{U}^\top = \mathbf{V}\mathbf{V}^\top = \mathbf{I}_d$ and $\|\mathbf{U}\mathbf{V}^\top\|_{\mathrm{F}} = O(d^{-\frac{1}{10}})$, it holds $\mathrm{d}_{\mathrm{TV}}(P_{A,\mathbf{U}}, P_{A,\mathbf{V}}) > 0.99$. Moreover, for all $\mathbf{V} \in \mathbb{R}^{m \times d}$ it holds $\mathrm{d}_{\mathrm{TV}}(P_{A,\mathbf{V}}, \mathcal{N}(0, \mathbf{I}_d)) > 0.99$.

(v) $\chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}_m)) \leq \delta^{-m/2}e^{O(m)}$.

To prove Theorem 7, we create a family of distributions of the form of Equation (1) by embedding the k-GMM onto many nearly orthogonal subspaces. The resulting distributions in $\mathbb{R}^d$ will be the k-GMMs described in our main theorem's statement. We then use the properties established in Proposition 9 to argue that this family has a large SQ dimension, making it hard to learn.

Proof (Proof of Theorem 7) Recall the definition of *decision problems* (Definition 14). Let the decision problem $\mathcal{B}(\mathcal{D}, D)$ where $D = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $\mathcal{D}$ is defined to be the set of all distributions of the form $P_{A,\mathbf{V}}$ as in Equation (1). We now lower bound the SQ dimension (Definition 4) of $\mathcal{B}(\mathcal{D}, D)$. Let S be the set from the fact below.

Fact 1 (See, e.g., Lemma 17 in Diakonikolas et al. (2021b)) *Let $m, d \in \mathbb{N}$ with $m < d^{1/10}$. There exists a set S of $2^{d^{\Omega(1)}}$ matrices in $\mathbb{R}^{m\times d}$ such that every $\mathbf{U} \in S$ satisfies $\mathbf{U}\mathbf{U}^\top = \mathbf{I}_m$ and every pair $\mathbf{U}, \mathbf{V} \in S$ with $\mathbf{U} \neq \mathbf{V}$ satisfies $\|\mathbf{U}\mathbf{V}^\top\|_\mathrm{F} \leq O(d^{-1/10})$.*

Let $\mathcal{D}_D := \{P_{A,\mathbf{V}}\}_{\mathbf{V}\in S}$. Using Fact 1 and Lemma 6, we have that for any distinct $\mathbf{V}, \mathbf{U} \in S$

$$|\chi_{\mathcal{N}(\mathbf{0},\mathbf{I}_m)}(P_{A,\mathbf{U}}, P_{A,\mathbf{V}})| \leq \left\|\mathbf{U}\mathbf{V}^\top\right\|_\mathrm{op}^{t+1} \chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}_m)) \leq \Omega(d)^{-(t+1)/10}\chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}_m)) ,$$

where we used that $\|\mathbf{A}\|_\mathrm{op} \leq \|\mathbf{A}\|_\mathrm{F}$ for any matrix $\mathbf{A}$. On the other hand, when $\mathbf{V} = \mathbf{U}$, we have that $|\chi_{\mathcal{N}(\mathbf{0},\mathbf{I}_m)}(P_{A,\mathbf{U}}, P_{A,\mathbf{V}})| \leq \chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}_m))$. Thus, the family $\mathcal{D}_D$ is (γ, β)-correlated with $\gamma = \Omega(d)^{-(t+1)/10}\chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}_m))$ and $\beta = \chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}_m))$ with respect to $D = \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$. This means that $\mathrm{SD}(\mathcal{B}(\mathcal{D}, D), \gamma, \beta) \geq \exp(d^{\Omega(1)})$.

Recall that $t = \Theta(1/\epsilon)$. Applying of Lemma 5 with $\gamma' := \gamma = \Omega(d)^{-(t+1)/10}\chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}_m))$, we obtain that any SQ algorithm for $\mathcal{Z}$ requires at least $\exp(d^{\Omega(1)})d^{-O(t)} = \exp(d^{\Omega(1)})d^{-O(1/\epsilon)}$ calls to

$$\mathrm{STAT}\left(\Omega(d)^{-\Omega(1/\epsilon)}\sqrt{\chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}_m))}\right) .$$

Finally, using Proposition 9, $\chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}_m)) \leq k^{O(1)} \exp(O(m)) = k^{O(1)} \exp(O(k^{2\epsilon})) \leq \exp(O(k^{2\epsilon}))$ where we also used our assumption that k^ϵ is much bigger than $\sqrt{\log k}$. Also, the number of calls $\exp(d^{\Omega(1)})d^{-O(1/\epsilon)}$ mentioned before can be lower bounded by $\exp(d^{\Omega(1)})$ by using our assumptions that $d > k^{C\epsilon} > (C/\epsilon)^C$. ■

4. Proof of Proposition 9

In Section 4.1 we provide the basis for Proposition 9, which shows the existence of a low-dimensional *discrete* distribution using an LP-duality argument. Then, in Section 4.2 we complete the proof of Proposition 9.

4.1. Existence via LP Duality

Proposition 10 *Let a sufficiently large absolute constant C. For any $m, t \in Z_+$ with $m > Ct^2$, there exists a discrete distribution D on $\mathbb{R}^m$ such that (let $\mathrm{supp}(D)$ denote the support of D):*

(i) $|\mathrm{supp}(D)| = m^{13t}$,

(ii) *D gives mass at least $0.99/|\mathrm{supp}(D)|$ to every point in its support,*

(iii) *D matches its first t moments with $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, i.e., $\mathbf{E}_{\mathbf{x}\sim D}[p(\mathbf{x})] = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[p(\mathbf{x})]$, for every polynomial $p : \mathbb{R}^d \to \mathbb{R}$ of degree at most t,*

(iv) $0.9\sqrt{m} \leq \|\mathbf{x}\|_2 \leq 1.1\sqrt{m}$ *for all* $\mathbf{x} \in \mathrm{supp}(D)$.

(v) *for any distinct $\mathbf{x}, \mathbf{y} \in \mathrm{supp}(D)$ it holds $\|\mathbf{x} - \mathbf{y}\|_2 \geq \sqrt{m}$.*

Proof Let a set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of $N = m^{13t}$ points drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$. We will show that with non-trivial probability, taking D to be the uniform distribution over S satisfies the desired properties. The proof is based on an LP duality argument. Proving Items (ii) and (iii) is equivalent to proving that the linear program below (with unknowns $\{\mu_i\}_{i\in[N]}$) admits a solution. Let $\alpha := 0.99/N$, the desired lower bound for all weights. The LP is the following:

$$\begin{aligned} &\text{Find:} \quad \mu_1, \ldots, \mu_N \\ &\text{s.t.:} \quad \sum_{i\in[n]} \mu_i p(\mathbf{x}_i) = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p(\mathbf{x})], \qquad \text{for any at most } t\text{-degree polynomial } p \\ &\qquad\qquad \mu_i \geq \alpha, \qquad \text{for all } i \in [N] \end{aligned} \tag{2}$$

Note that the first constraint for p being the constant polynomial $p = 1$ means that the μ_i's form a valid distribution. By standard LP duality, the above is feasible unless there exists a linear combination of constraints that produces the contradiction $0 < -1$. Concretely, we start by introducing multipliers, also known as dual variables, for each constraint. For the final constraint, these will be some variables $\beta_i \geq 0$ for $i \in [N]$. Regarding the first constraint, a multiplier from $\mathbb{R}$ is assigned to every polynomial with a degree of at most t. However, since the first constraint applies to all such polynomials and the set is closed under multiplication, these dual variables can be absorbed into the polynomials and will not be explicitly written. After multiplying and summing the constraints, we obtain

$$\sum_{i\in[N]} \mu_i \left(-\beta_i + p(\mathbf{x}_i)\right) \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p(\mathbf{x})] - \alpha \sum_{i\in[N]} \beta_i \ . \tag{3}$$

To derive the dual LP, we set the coefficients of μ_i equal to zero and ask for the right-hand side of Equation (3) to be negative. This means that the primal LP (2) is feasible unless LP (4) on the left part below has a solution, where LP (4) is further equivalent to LP (5) on the right part:

$$\begin{aligned} &\text{Find:} \quad \beta_1, \ldots, \beta_N \in \mathbb{R}_+, \\ &\qquad\qquad p \text{ at most } t\text{-degree polynomial} \\ &\text{s.t.:} \quad -\beta_i + p(\mathbf{x}_i) = 0, \ \forall i \in [N] \\ &\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p(\mathbf{x})] < \alpha \sum_{i\in[N]} \beta_i \end{aligned} \tag{4}$$

$$\begin{aligned} &\text{Find:} \quad p \text{ at most } t\text{-degree polynomial} \\ &\text{s.t.:} \qquad p(\mathbf{x}_i) \geq 0, \ \forall i \in [N] \\ &\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p(\mathbf{x})] < \alpha \cdot N \cdot \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{U}(S)}[p(\mathbf{x})] \end{aligned} \tag{5}$$

For verifying the equivalence of the two LPs it suffices to note that $\sum_{i\in[N]} \beta_i = \sum_{i\in[N]} p(\mathbf{x}_i) = N\, \mathbf{E}_{\mathbf{x}\sim\mathcal{U}(S)}[p(\mathbf{x})]$. By scaling (homogeneity), we can assume in the above that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p^2(\mathbf{x})] = 1$. Recall that the points $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are samples from $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$. Since, we are proving the proposition via probabilistic argument, it remains to show that with non-trivial probability these points will be such that LP (5) is infeasible (and thus LP (2) is feasible). We prove this by contradiction: Assume that LP (5) is feasible. Let $\mathcal{U}(S)$ be the uniform distribution over S. We show that, in fact, $\mathcal{U}(S)$ approximates the first four moments of any polynomial with a degree at most t (the proof is given in Appendix B.1). Formally:

Claim 1 *Let a set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of i.i.d. samples $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$. If $N > 10m^{12t}/\eta^2$, then with probability at least 0.6 for any polynomial $p : \mathbb{R}^m \to \mathbb{R}$ of degree at most t, it holds* *(i)* $\mathbf{E}_{\mathbf{x}\sim\mathcal{U}(S)}[p(\mathbf{x})] \leq \mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p(\mathbf{x})] + \eta$,

(ii) $\mathbf{E}_{\mathbf{x}\sim\mathcal{U}(S)}[p^2(\mathbf{x})] \geq \mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p^2(\mathbf{x})] - \eta$, *and*

(iii) $\mathbf{E}_{\mathbf{x}\sim\mathcal{U}(S)}[p^4(\mathbf{x})] \leq \mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p^4(\mathbf{x})] + \eta$.

For our case, we assumed that LP (5) is feasible thus $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p(\mathbf{x})] < aN\,\mathbf{E}_{\mathbf{x}\sim\mathcal{U}(S)}[p(\mathbf{x})]$. We use Claim 1 with accuracy $\eta = 3^{-t}/200$, so the sample complexity from that claim becomes $40000 \cdot 9^t m^{12t}$. Since we assumed that $m > 6000$, the number of samples that we use is $N = m^{13t} > 40000 \cdot 9^t m^{12t}$ and thus satisfies the requirement of the claim. The claim thus yields

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{U}(S)}[p(\mathbf{x})] \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p(\mathbf{x})] + \eta < aN \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{U}(S)}[p(\mathbf{x})] + \eta \,,$$

which means that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{U}(S)}[p(\mathbf{x})] < \frac{\eta}{1-aN} \leq \frac{3^{-t}/200}{1-0.99} = \frac{3^{-t}}{2} \,. \tag{6}$$

On the other hand, for every $t \geq 1$ we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{U}(S)}[p(\mathbf{x})] \geq \frac{\mathbf{E}_{\mathbf{x}\sim\mathcal{U}(S)}[p^2(\mathbf{x})]^{3/2}}{\sqrt{\mathbf{E}_{\mathbf{x}\sim\mathcal{U}(S)}[p^4(\mathbf{x})]}} \geq \frac{(1-\eta)^{3/2}}{\sqrt{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p^4(\mathbf{x})] + \eta}} \geq \frac{0.7}{\sqrt{3^{2t} + 3^{-t}/2}} \geq \frac{3^{-t}}{2} \,, \tag{7}$$

where the penultimate inequality uses Gaussian hypercontractivity (Fact 5). Comparing Equations (6) and (7) we have obtained a contradiction.

We now show the lower bound of Item (iv). Using the concentration of the norm of a Gaussian vector (Fact 3 with $\beta = \sqrt{m}/10$), we have that

$$\mathop{\mathbf{Pr}}_{\mathbf{x}_1,\ldots,\mathbf{x}_N\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[\exists i : |\|\mathbf{x}_i\|_2 - \sqrt{m}| < 0.1\sqrt{m}] \leq 2Ne^{-m/1600} = 2m^{13t}e^{-m/1600} < 0.1 \,. \tag{8}$$

where we used that $t < \sqrt{m}/16000 < \frac{m/1600-\ln(20)}{13\ln m}$ for $m > 30000$.

Regarding Item (v), it is a standard property of the Gaussian all pairs of points are nearly-orthogonal with high probability (Fact 4 with $\alpha = 0.1$),

$$\mathop{\mathbf{Pr}}_{\mathbf{x}_1,\ldots,\mathbf{x}_N\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[\exists i \neq j : |\langle \mathbf{x}_i, \mathbf{x}_j\rangle| > m^{-0.1}] \leq N^2 e^{-m^{0.8}/5} \leq m^{26t}e^{-m^{0.8}/5} < 0.1 \,, \tag{9}$$

where the last inequality uses that $t < \sqrt{m}/16000 < \frac{m^{0.8}/5-\ln(10)}{26\ln m}$ for $m > 30000$. Conditioning on the two bad events of Equations (8) and (9) not happening, we have that for any distinct $i, j \in [N]$, it holds $\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 = \|\mathbf{x}_i\|_2^2 + \|\mathbf{x}_j\|_2^2 - 2\langle \mathbf{x}_i, \mathbf{x}_j\rangle \geq 1.62m - 2m^{-0.1} \geq m$, for $m > 2$. ■

4.2. Proof of Proposition 9

We use the following throughout the proof: Let D be the distribution from Proposition 10 with parameters $m = k^{2\epsilon}$, and $t = 1/(26\epsilon)$ (note that because of our assumption $k > (C/\epsilon)^{1/\epsilon}$ the requirement of Proposition 10 is satisfied and thus the proposition is applicable). Let $A = U_\rho(D)$, where U_ρ denotes the Ornstein-Uhlenbeck operator. We choose $\rho = \sqrt{1-\delta}$ and $\delta = ck^{-2.5/m}$ for a sufficiently small positive constant C. We prove each part of Proposition 9 separately.

Proof (Proof of Item (i)) The fact that A is a mixture of Gaussians with each component having variance δ follows immediately by the definition of A as the distribution D after Gaussian smoothing via the Ornstein-Uhlenbeck operator with parameter $\rho = \sqrt{1-\delta}$. We can also check that the number of components is k: By Proposition 10 we have that the number of components is m^{13t}. Recall that we have further selected $m = k^{2\epsilon}$. Thus, the number of components is $m^{13t} = k^{26\epsilon t}$. This is equal to k by our choice of $t = 1/(26\epsilon)$. The fact that we have mass $0.99/k$ for each Gaussian component follows from Item (ii) of Proposition 10. ■

Proof (Proof of Item (ii)) For any $\mathbf{a} \in \mathbb{N}^m$ with $|\mathbf{a}| \leq t$, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim U_\rho(D)}[h_{\mathbf{a}}(\mathbf{x})] = \rho^{|\mathbf{a}|} \mathop{\mathbf{E}}_{\mathbf{x}\sim D}[h_{\mathbf{a}}(\mathbf{x})] = \rho^{|\mathbf{a}|} \mathop{\mathbf{E}}_{\mathbf{x}\sim \mathcal{N}(\mathbf{0},\mathbf{I}_m)}[h_{\mathbf{a}}(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{x}\sim \mathcal{N}(\mathbf{0},\mathbf{I}_m)}[h_{\mathbf{a}}(\mathbf{x})] \, ,$$

where the first equality uses Fact 2, the next one uses Item (iii) of Proposition 10, and the last one is due to the property of Hermite polynomials $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_{\mathbf{a}}(\mathbf{x})] = 1$ if $|\mathbf{a}| = 0$ and zero otherwise. ■

Proof (Proof of Item (iii)) Using Item (v) of Proposition 10 combined with our choice $m = k^{2\epsilon}$ and the fact that the Ornstein-Uhlenbeck operator scales all the means by a factor of $\rho = \sqrt{1-\delta} > 1/2$, we will have that the pairwise means separation in our construction is at least $\rho k^\epsilon > k^\epsilon/2$. ■

Proof (Proof of Item 2) We first prove the claim that $\mathrm{d_{TV}}(P_{A,\mathbf{U}}, P_{A,\mathbf{V}}) > 0.99$. Since it always holds $\mathrm{d_{TV}}(P_{A,\mathbf{U}}, P_{A,\mathbf{V}}) = 1 - \int_{\mathbf{z}\in\mathbb{R}^d} \min\{P_{A,\mathbf{U}}(\mathbf{z}), P_{A,\mathbf{V}}(\mathbf{z})\}\mathrm{d}\mathbf{z}$, we will focus on upper bounding $\mathcal{I}_{\mathbf{U},\mathbf{V}} := \int_{\mathbf{z}\in\mathbb{R}^d} \min\{P_{A,\mathbf{U}}(\mathbf{z}), P_{A,\mathbf{V}}(\mathbf{z})\}\mathrm{d}\mathbf{z}$. Let $\mathbf{v}_i, \mathbf{u}_i$ be the rows of $\mathbf{V}$ and $\mathbf{U}$ respectively. Extend $\mathbf{v}_1, \ldots, \mathbf{v}_m$ to an orthonormal basis $\mathbf{v}_1, \ldots, \mathbf{v}_m, \ldots, \mathbf{v}_{2m}$ of the space spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_m, \mathbf{u}_1, \ldots, \mathbf{u}_m$. Let $\mathbf{x}$ be an orthonormal coordinate system of dimension m that is aligned with the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$ and let $\mathbf{y}$ be another coordinate system aligned with the vectors $\mathbf{v}_{m+1}, \ldots, \mathbf{v}_{2m}$. Similarly, let $\mathbf{x}'$ be the coordinate system aligned with the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_m$ and $\mathbf{y}'$ be the one for the orthogonal directions. Since, $P_{A,\mathbf{U}}, P_{A,\mathbf{V}}$ are both standard Gaussians in the subspace perpendicular to both $\mathbf{U}$ and $\mathbf{V}$, the contribution to their total variation there is zero and we are left with the integral over the two subspaces $\mathbf{U}$ and $\mathbf{V}$,

$$\mathcal{I}_{\mathbf{U},\mathbf{V}} := \int_{\mathbf{z}\in\mathbb{R}^d} \min\{P_{A,\mathbf{U}}(\mathbf{z}), P_{A,\mathbf{V}}(\mathbf{z})\}\mathrm{d}\mathbf{z} = \iint_{\mathbf{x}\in\mathbb{R}^m,\mathbf{y}\in\mathbb{R}^m} \min\{A(\mathbf{x})\phi_m(\mathbf{y}), A(\mathbf{x}')\phi_m(\mathbf{y}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \, .$$

First, note that $\phi_m(\cdot) \leq (2\pi)^{-2m} \leq 1$ pointwise thus we can focus on the factors in the integral. We can then write out what these coordinates $\mathbf{x}'$ and $\mathbf{y}'$ that appear in the integral are in terms of $\mathbf{x}, \mathbf{y}$, and then perform a change of variables from $\mathbf{x}, \mathbf{y}$ to $\mathbf{x}, \mathbf{x}'$. The steps so far are summarized in the following claim which we prove formally in Appendix B.2.

Claim 2 *Let $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{m\times d}$ matrices with $\mathbf{U}\mathbf{U}^\top = \mathbf{V}\mathbf{U}^\top = \mathbf{I}_d$ and rows $\mathbf{v}_1, \ldots, \mathbf{v}_m$ and $\mathbf{u}_1, \ldots, \mathbf{u}_m$ respectively. Let the extension $\mathbf{v}_1, \ldots, \mathbf{v}_{2m}$ of the rows of $\mathbf{V}$ to an orthonormal basis of the space spaned by $\mathbf{v}_1, \ldots, \mathbf{v}_m, \mathbf{u}_1, \ldots, \mathbf{u}_m$. Denote $\mathbf{R}_{\mathbf{V}_2} = [\mathbf{v}_{m+1} \ldots \mathbf{v}_{2m}]^\top$. Then,*

$$\mathcal{I}_{\mathbf{U},\mathbf{V}} := \int_{\mathbf{z}\in\mathbb{R}^d} \min\{P_{A,\mathbf{U}}(\mathbf{z}), P_{A,\mathbf{V}}(\mathbf{z})\}\mathrm{d}\mathbf{z} \leq \frac{1}{\det(\mathbf{U}\mathbf{R}_{\mathbf{V}_2}^\top)} \iint_{\mathbf{x}\in\mathbb{R}^m,\mathbf{x}'\in\mathbb{R}^m} \min\{A(\mathbf{x}), A(\mathbf{x}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \, .$$

We now claim that this determinant is close to one because $\mathbf{V}$ and $\mathbf{U}$ are nearly-orthogonal and thus the singular values of the matrix $\mathbf{U}\mathbf{R}_{\mathbf{V}_2}^\top$ are all close to one. We defer the proof to Appendix B.2. The requirement $d > m^C$ below holds by assumption.

Claim 3 *If $d > m^C$ for a sufficiently large absolute constant C, then* $\det(\mathbf{U}\mathbf{R}_{\mathbf{V}_2}^\top) \geq 1/2$.

We are now ready to further bound our integral $\mathcal{I}_{\mathbf{V},\mathbf{U}}$. First, by writing the distribution A as a mixture $\sum_{i\in[k]} \lambda_i A_i(\mathbf{x})$, we can break $\mathcal{I}_{\mathbf{V},\mathbf{U}}$ into contributions from every pair of components (the proof is straight forward but included in Appendix B.2 for completeness).

Claim 4 *The following bound holds:* $\mathcal{I}_{\mathbf{V},\mathbf{U}} \leq 2k \max_{i,j\in[k]} \iint_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^m} \min\{A_i(\mathbf{x}), A_j(\mathbf{x}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'$.

Recall that each component A_i of the mixture distribution A is by definition Gaussian with variance $\delta := ck^{-2.5/m}$ in all directions. Let $R := C'\sqrt{\delta m \log(1/\delta)}$ for a sufficiently large constant C' so that: $\Pr_{\mathbf{z}\sim\mathcal{N}(0,2\delta\mathbf{I}_m)}[\|\mathbf{z}\|_2 > R] \leq \delta$ (this follows by standard Gaussian norm concentration, see Claim 6 in Appendix B.2).

We can thus break the integral appearing in Claim 4 into parts based on whether $\mathbf{x}$ and $\mathbf{x}'$ fall within or outside a ball of radius R around the mean of the component (recall that R is the radius used in Equation (17)). For each individual integral we will use Claim 6 to bound the mass of the distribution outside of the ball and bound the mass inside the ball by the volume of that ball. Then by upper bounding that volume and after some algebra we can bound all terms by the following (again, the proof is deffered to Appendix):

Claim 5 $\mathcal{I}_{\mathbf{V},\mathbf{U}} \leq C^m k \delta^{0.4m}$ *for a sufficiently large absolute constant C.*

The total variation distance is thus $\mathrm{d_{TV}}(P_{A,\mathbf{U}}, P_{A,\mathbf{V}}) = 1 - \int_{z\in\mathbb{R}^d} \min\{P_{A,\mathbf{V}}(\mathbf{z}), P_{A,\mathbf{V}}(\mathbf{z})\}\mathrm{d}\mathbf{z} \geq 1 - C^m k\delta^{0.4m} \geq 0.99$, where the last step uses that $\delta = ck^{-2.5/m}$ for an appropriately small constant $c > 0$. The remaining part of the claim that $\mathrm{d_{TV}}(P_{A,\mathbf{V}}, \mathcal{N}(\mathbf{0}, \mathbf{I}_d)) > 0.99$ can be handled with similar arguments, and is deferred to Claim 7 in Appendix. ■

Proof (Proof of Item (v)) The bound is fairly standard and deferred to Claim 8 in Appendix B.2. ■

5. Beating Separation of $\Omega(\sqrt{k})$: Proof of Theorem 3

In this section, we prove the following result which is the formal version of Theorem 3.

Theorem 11 (Quadratic SQ Lower Bound for Separation $\sim k^{1/2}$) *Let $C > 0$ be a sufficiently large absolute constant. Let $d, k \in \mathbb{Z}_+$ and $c \in (0, 2/9)$ with $d > (1/c)^{C/c}$, $2 \leq k \leq (c/C)\log d$. Consider the following hypothesis testing problem regarding a distribution P on $\mathbb{R}^d$:*

- *(Null Hypothesis) $P = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.*
- *(Alternative Hypothesis) P belongs to a family $\mathcal{P}$, every member of which is a mixture of Gaussians $\sum_{i=1}^k w_i \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ with uniform weights $w_i = 1/k$, mean vectors with pairwise separation $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2 \geq \sqrt{k}/3$ for all $i \neq j \in [k]$, and common covariance matrix $\boldsymbol{\Sigma} \preceq \mathbf{I}_d$. Moreover, $\mathrm{d_{TV}}(P, \mathcal{N}(\mathbf{0}, \mathbf{I}_d)) > 0.99$ and $\mathrm{d_{TV}}(P, P') > 0.99$ for all distinct $P, P' \in \mathcal{P}$.*

Any algorithm with statistical query access to P that distinguishes correctly between the two cases, does one of the following: it performs $2^{\Omega(d^{2c})}$ statistical queries, or it uses at least one statistical query to $\mathrm{VSTAT}(\Omega(d^{2-9c}))$.

We start with a brief overview of the new ideas required for the proof.

First, it is instructive to explain why Theorem 7 and its proof do not suffice for our purposes. In particular, to use Theorem 7 in order to obtain an SQ lower bound of $2^{d^{\Omega(1)}}$ queries vs a query to $\mathrm{VSTAT}(d^2)$, we need to set the parameter ϵ (where the separation is $\Delta = k^\epsilon$) sufficiently small. This is because in that theorem, ϵ appears inside a big-Ω notation in the query tolerance and a closer examination of our proofs reveals that the hidden constant in that big-Ω is rather large (in the order of hundreds). Thus, Theorem 7 cannot yield a super-linear SQ lower bound for the $\epsilon = 1/2$ case, which corresponds to pairwise separation of $\sim \sqrt{k}$.

In more detail, the constant factor in front of ϵ in Theorem 7 is large for two reasons: (i) The number of Gaussian components in our construction (c.f. Proposition 9) was $k^{26\epsilon t}$, meaning that we had to match $t = 1/(26\epsilon)$ many moments in order to end-up with k components, and (ii) the fact about random matrices being nearly orthogonal (Fact 1) that we used was suboptimal. In particular, while the corresponding fact for vectors states that any pair of random unit vectors has inner product very close to $O(d^{-1/2})$, the generalization of that to matrices by Fact 1 stated that the pairs of random matrices $\mathbf{U}, \mathbf{V}$ have $\|\mathbf{U}\mathbf{V}^\top\|_\mathrm{F} \le O(d^{-1/10})$. The constant in the exponent is crucial here because it also appears in front of ϵ in the final SQ lower bound.

In this section, we overcome both of these issues by providing a tighter construction and analysis for the $\epsilon = 1/2$ case. In particular, we replace the existential LP-duality argument of Proposition 9 by Lemma 12 below, which provides a discrete distribution matching the first three moments with the standard Gaussian. The proof is constructive and can be found in Appendix C.

Lemma 12 (Moment Matching) *There exists a discrete distribution D on $\mathbb{R}^m$ such that: (i) D is supported on $2m$ points, (ii) D matches the first three moments with $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, and (iii) for every pair of distinct points $\mathbf{x}, \mathbf{y}$ in the support of D, it holds $\|\mathbf{x} - \mathbf{y}\|_2 \ge \sqrt{m}$.*

Moreover, we provide a tight version of Fact 1 via an improved analysis (see Appendix C for the proof).

Lemma 13 *Let C be a sufficiently large absolute constant. Let $c \in (0, 1/4)$ and $m, d \in \mathbb{N}$ with $d > (1/c)^{C/c}$ and $m < d^{c/5}/C$. There exists a set S of $2^{\Omega(d^{2c})}$ matrices in $\mathbb{R}^{m\times d}$ such that every $\mathbf{A} \in S$ satisfies $\mathbf{A}\mathbf{A}^\top = \mathbf{I}_m$ and every pair $\mathbf{A}, \mathbf{A}' \in S$ with $\mathbf{A} \neq \mathbf{A}'$ satisfies $\|\mathbf{A}'\mathbf{A}^\top\|_{\mathrm{op}} \lesssim d^{-1/2+2c}$.*

Given the above lemma, the proof of Theorem 11 follows in a similar way to Theorem 7, and is thus deferred to Appendix C.

Acknowledgments

Ilias Diakonikolas was supported by NSF Medium Award CCF-2107079, NSF Award CCF-1652862 (CAREER), and a DARPA Learning with Less Labels (LwLL) grant. Daniel Kane was supported by NSF Medium Award CCF-2107547 and NSF Award CCF-1553288 (CAREER). Thanasis Pittas was supported by NSF Medium Award CCF-2107079. Nikos Zarifis was supported in part by NSF award 2023239, NSF Medium Award CCF-2107079, and a DARPA Learning with Less Labels (LwLL) grant.

References

D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *Proceedings of the Eighteenth Annual Conference on Learning Theory (COLT)*, pages 458–469, 2005.

G. E. Andrews, R. Askey, and R. Roy. *Special Functions*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1999. doi: 10.1017/CBO9781107325937.

S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *Proceedings of the 33rd Symposium on Theory of Computing*, pages 247–257, 2001.

A. Bakshi, I. Diakonikolas, S. B. Hopkins, D. Kane, S. Karmalkar, and P. K. Kothari. Outlier-robust clustering of gaussians and other non-spherical mixtures. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020*, pages 149–159, 2020.

A. Bakshi, I. Diakonikolas, H. Jia, D.M. Kane, P. Kothari, and S. Vempala. Robustly learning mixtures of k arbitrary gaussians. In *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1234–1247, 2022. Full version available at https://arxiv.org/abs/2012.02119.

M. Belkin and K. Sinha. Polynomial learning of distribution families. In *FOCS*, pages 103–112, 2010.

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

M. Brennan, G. Bresler, S. B. Hopkins, J. Li, and T. Schramm. Statistical query algorithms and low degree tests are almost equivalent. In *Conference on Learning Theory*, pages 774–774. PMLR, 2021.

S. C. Brubaker and S. Vempala. Isotropic PCA and Affine-Invariant Clustering. In *Proc. 49th IEEE Symposium on Foundations of Computer Science*, pages 551–560, 2008.

J. Bruna, O. Regev, M. J. Song, and Y. Tang. Continuous LWE. In *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 694–707. ACM, 2021.

T. Cai, J. Fan, and T. Jiang. Distributions of angles in random packing on spheres. *Journal of Machine Learning Research*, 14(1):1837–1864, 2013.

S. Chen, J. Li, and Y. Li. Learning (very) simple generative models is hard. In *NeurIPS*, 2022.

S. Dasgupta. Learning mixtures of Gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pages 634–644, 1999.

C. Daskalakis and G. Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014*, pages 1183–1213, 2014.

D. Davis, M. Díaz, and K. Wang. Clustering a mixture of gaussians with unknown covariance. *CoRR*, abs/2110.01602, 2021. URL https://arxiv.org/abs/2110.01602.

I. Diakonikolas and D. Kane. Near-optimal statistical query hardness of learning halfspaces with massart noise. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4258–4282. PMLR, 2022. Full version available at https://arxiv.org/abs/2012.09720.

I. Diakonikolas and D. M. Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, 2023.

I. Diakonikolas, P. Gopalan, R. Jaiswal, R. Servedio, and E. Viola. Bounded independence fools halfspaces. *SIAM J. on Comput.*, 39(8):3441–3462, 2010.

I. Diakonikolas, D. M. Kane, and A. Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017*, pages 73–84, 2017. Full version at http://arxiv.org/abs/1611.03473.

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, STOC 2018*, pages 1047–1060, 2018. Full version available at https://arxiv.org/abs/1711.07211.

I. Diakonikolas, W. Kong, and A. Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019*, pages 2745–2754, 2019.

I. Diakonikolas, S. B. Hopkins, D. Kane, and S. Karmalkar. Robustly learning any clusterable mixture of gaussians. *CoRR*, abs/2005.06417, 2020a. URL https://arxiv.org/abs/2005.06417.

I. Diakonikolas, D. Kane, and N. Zarifis. Near-optimal SQ lower bounds for agnostically learning halfspaces and relus under gaussian marginals. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020b.

I. Diakonikolas, D. M. Kane, V. Kontonis, and N. Zarifis. Algorithms and SQ lower bounds for PAC learning one-hidden-layer relu networks. In *Conference on Learning Theory, COLT 2020*, volume 125 of *Proceedings of Machine Learning Research*, pages 1514–1539. PMLR, 2020c.

I. Diakonikolas, D. M. Kane, A. Pensia, T. Pittas, and A. Stewart. Statistical query lower bounds for list-decodable linear regression. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 3191–3204, 2021a.

I. Diakonikolas, D. M. Kane, T. Pittas, and N. Zarifis. The optimality of polynomial regression for agnostic learning under gaussian marginals in the sq model. In *Conference on Learning Theory*, pages 1552–1584. PMLR, 2021b.

I. Diakonikolas, D. M. Kane, S. Karmalkar, A. Pensia, and T. Pittas. List-decodable sparse mean estimation via difference-of-pairs filtering. *CoRR*, abs/2206.05245, 2022a. URL https://doi.org/10.48550/arXiv.2206.05245. Conference version in NeurIPS'22.

I. Diakonikolas, D. M. Kane, D. Kongsgaard, J. Li, and K. Tian. Clustering mixture models in almost-linear time via list-decodable mean estimation. In *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, 2022*, pages 1262–1275, 2022b. Full version available at https://arxiv.org/abs/2106.08537.

I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. Learning general halfspaces with general massart noise under the gaussian distribution. In *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 874–885, 2022c. Full version available at https://arxiv.org/abs/2108.08767.

J. Feldman, R. O'Donnell, and R. Servedio. PAC learning mixtures of Gaussians with no separation assumption. In *Proc. 19th Annual Conference on Learning Theory (COLT)*, pages 20–34, 2006.

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.

S. Goel, A. Gollakota, and A. R. Klivans. Statistical-query lower bounds via functional gradients. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

A. Gupte, N. Vafa, and V. Vaikuntanathan. Continuous LWE is as hard as LWE & applications to learning gaussian mixtures. In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022*, pages 1162–1173. IEEE, 2022.

M. Hardt and E. Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015*, pages 753–760, 2015.

S. B. Hopkins and J. Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pages 1021–1034, 2018.

R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. *SIAM J. Comput.*, 38(3):1141–1156, 2008.

M. J. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6): 983–1006, 1998.

P. K. Kothari, J. Steinhardt, and D. Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pages 1035–1046, 2018.

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, 2022.

J. Li and A. Liu. Clustering mixtures with almost optimal separation in polynomial time. In *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1248–1261. ACM, 2022.

A. Liu and A. Moitra. Settling the robust learnability of mixtures of gaussians. In *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 518–531. ACM, 2021. Full version available at https://arxiv.org/abs/2011.03622.

A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, pages 93–102, 2010.

E. Nelson. The free markoff field. *Journal of Functional Analysis*, 12(2):211–227, 1973.

R. O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. ISBN 978-1-10-703832-5.

K. Pearson. Contribution to the mathematical theory of evolution. *Phil. Trans. Roy. Soc. A*, 185: 71–110, 1894.

O. Regev and A. Vijayaraghavan. On learning mixtures of well-separated gaussians. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017*, pages 85–96, 2017.

D. Steurer and S. Tiegel. Sos degree reduction with applications to clustering and robust moment estimation. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021*, pages 374–393. SIAM, 2021.

A. T. Suresh, A. Orlitsky, J. Acharya, and A. Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1395–1403, 2014.

G. Szegö. *Orthogonal Polynomials*, volume XXIII of *American Mathematical Society Colloquium Publications*. A.M.S, Providence, 1989.

S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. In *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science*, pages 113–122, 2002.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Number 47 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge ; New York, NY, 2018. ISBN 978-1-108-41519-4. doi: 10.1017/9781108231596.

S.-A. Wegner. *Lecture Notes on High-Dimensional Data*. 2021. URL https://arxiv.org/abs/2101.05841.

Appendix A. Additional Preliminaries

A.1. Additional Notation

We use $\mathbb{Z}$ for the set of integers and $\mathbb{Z}_+$ for positive integers. For $n \in \mathbb{Z}_+$, we denote $[n] \stackrel{\text{def}}{=} \{1, \ldots, n\}$ and use $\mathcal{S}^{d-1}$ for the d-dimensional unit sphere. We use $\mathcal{S}_{d-1}(R)$ to denote the d dimensional sphere with radius R and center the origin. For a vector $\mathbf{v}$, we let $\|\mathbf{v}\|_2$ denote its ℓ_2-norm. We use $\mathbf{I}_d$ to denote the $d \times d$ identity matrix. We will drop the subscript when it is clear from the context. For a matrix $\mathbf{A}$, we use $\|\mathbf{A}\|_{\mathrm{F}}$ and $\|\mathbf{A}\|_{\mathrm{op}}$ to denote the Frobenius and spectral (or operator) norms respectively. If $\mathbf{a} = (a_1, \ldots, a_m) \in \mathbb{Z}_+^m$ is a multi-index, we denote $|\mathbf{a}| = \sum_{i=1}^m a_i$

We use $a \lesssim b$ to denote that there exists an absolute universal constant $C > 0$ (independent of the variables or parameters on which a and b depend) such that $a \leq Cb$.

We use the notation $x \sim D$ to denote that a random variable x is distributed according to the distribution D. For a random variable x, we use $\mathbf{E}[x]$ for its expectation. We use $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For a set S, we use $\mathcal{U}(S)$ to denote the uniform distribution on S and use $x \sim S$ as a shortcut for $x \sim \mathcal{U}(S)$. We denote by $\phi_m(\mathbf{x})$ the probability density function (pdf) of the standard Gaussian in m-dimensions $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, and by $\phi(x)$ the pdf of the univariate standard Gaussian $\mathcal{N}(0, 1)$. We slightly abuse notation by using the same letter for a distribution and its pdf, e.g., we will denote by $P(\mathbf{x})$ the pdf of a distribution P. We use $\mathrm{d}_{\mathrm{TV}}(P, Q)$ for the total variation distance between two distributions P, Q.

We will prefer to use capital letters for constants that are assumed to be sufficiently large and small letters for constants that need to be sufficiently small.

A.2. Hermite Analysis

Hermite polynomials form a complete orthogonal basis of the vector space $L_2(\mathbb{R}, \mathcal{N}(0, 1))$ of all functions $f : \mathbb{R} \to \mathbb{R}$ such that $\mathbf{E}_{x \sim \mathcal{N}(0,1)}[f^2(x)] < \infty$. There are two commonly used types of Hermite polynomials. The *physicist's* Hermite polynomials, denoted by H_k for $k \in \mathbb{Z}$ satisfy the following orthogonality property with respect to the weight function e^{-x^2}: for all $k, m \in \mathbb{Z}$, $\int_{\mathbb{R}} H_k(x) H_m(x) e^{-x^2} \mathrm{d}x = \sqrt{\pi} 2^k k! \mathbf{1}(k = m)$. The *probabilist's* Hermite polynomials H_{e_k} for $k \in \mathbb{Z}$ satisfy $\int_{\mathbb{R}} H_{e_k}(x) H_{e_m}(x) e^{-x^2/2} \mathrm{d}x = k! \sqrt{2\pi} \mathbf{1}(k = m)$ and are related to the physicist's polynomials through $H_{e_k}(x) = 2^{-k/2} H_k(x/\sqrt{2})$. We will mostly use the *normalized probabilist's* Hermite polynomials $h_k(x) = H_{e_k}(x)/\sqrt{k!}$, $k \in \mathbb{Z}$ for which $\int_{\mathbb{R}} h_k(x) h_m(x) e^{-x^2/2} \mathrm{d}x = \sqrt{2\pi} \mathbf{1}(k = m)$. These polynomials are the ones obtained by Gram-Schmidt orthonormalization of the basis $\{1, x, x^2, \ldots\}$ 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 $f(x) = \sum_{i \in \mathbb{Z}} a_i h_i(x)$ and we have $\lim_{n \to \infty} \mathbf{E}_{x \sim \mathcal{N}(0,1)}[(f(x) - \sum_{i=0}^n a_i h_i(x))^2] = 0$ (see, e.g., Andrews et al. (1999)). Moreover, we have the following explicit expression of $h_i(\cdot)$ (see, for example, Andrews et al. (1999); Szegö (1989)):

$$h_i(x) = \sqrt{i!} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{(-1)^j}{j!(i-2j)!} \frac{x^{i-2j}}{2^j} \ . \tag{10}$$

Extending the normalized probabilist's Hermite polynomials to higher dimensions, an orthonormal basis of $L_2(\mathbb{R}^d, \mathcal{N}(\mathbf{0}, \mathbf{I}_d))$ (with respect to the inner product $\langle f, g \rangle = \mathbf{E}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)}[f(\mathbf{x}) g(\mathbf{x})]$) can be formed by all the products of one-dimensional Hermite polynomials, i.e., $h_{\mathbf{a}}(\mathbf{x}) = \prod_{i=1}^d h_{a_i}(x_i)$, for all multi-indices $\mathbf{a} \in \mathbb{Z}^d$ (we are now slightly overloading notation by using multi-indices as subscripts). The total degree of $h_{\mathbf{a}}$ is $|\mathbf{a}| = \sum_{i=1}^d a_i$.

Ornstein-Uhlenbeck Operator For a $\rho > 0$, we define the *Gaussian noise* (or *Ornstein-Uhlenbeck*) operator U_ρ as the operator that maps a distribution F on $\mathbb{R}^m$ to the distribution of the random variable $\rho\mathbf{x} + \sqrt{1-\rho^2}\mathbf{z}$, where $\mathbf{x} \sim F$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ independently of $\mathbf{x}$. A standard property of the U_ρ operator is that it operates diagonally with respect to Hermite polynomials:

Fact 2 (see, e.g., Proposition 11.37 in O'Donnell (2014)) *For any multivariate Hermite polynomial $h_{\mathbf{a}}$, any F on $\mathbb{R}$, and $\rho \in (0,1)$, that $\mathbf{E}_{\mathbf{x}\sim U_\rho F}[h_{\mathbf{a}}(\mathbf{x})] = \rho^{|\mathbf{a}|}\,\mathbf{E}_{\mathbf{x}\sim F}[h_{\mathbf{a}}(\mathbf{x})]$, where $|\mathbf{a}| = \sum_i a_i$.*

A.3. Background on the Statistical Query Model

Definition 14 (Decision Problem over Distributions) *Let D be a fixed distribution and $\mathcal{D}$ be a distribution family. We denote by $\mathcal{B}(\mathcal{D}, D)$ the decision (or hypothesis testing) problem in which the input distribution D' is promised to satisfy either (a) $D' = D$ or (b) $D' \in \mathcal{D}$, and the goal is to distinguish between the two cases.*

Definition 15 (Pairwise Correlation) *The pairwise correlation of two distributions with probability density functions $D_1, D_2 : \mathbb{R}^d \to \mathbb{R}_+$ with respect to a distribution with density $D : \mathbb{R}^d \to \mathbb{R}_+$, where the support of D contains the supports of D_1 and D_2, is defined as $\chi_D(D_1, D_2) = \int_{\mathbb{R}^d} D_1(\mathbf{x})D_2(\mathbf{x})/D(\mathbf{x})\,\mathrm{d}\mathbf{x} - 1$.*

Definition 16 *We say that a set of s distributions $\mathcal{D} = \{D_1, \ldots, D_s\}$ is (γ, β)-correlated relative to a distribution D if $|\chi_D(D_i, D_j)| \le \gamma$ for all $i \neq j$, and $|\chi_D(D_i, D_j)| \le \beta$ for $i = j$.*

A.4. Miscallenious Facts

We require the standard concetration of the norm of Gaussian vectors (see, e.g., Theorem 3.1.1 of Vershynin (2018) or Theorem 4.7 of Wegner (2021)):

Fact 3 (Gaussian Norm Concentration) *For every $0 \le \beta \le \sigma\sqrt{d}$ we have that*

$$\mathop{\mathbf{Pr}}_{X\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_d)}[|\|\mathbf{x}\|_2 - \sigma\sqrt{d}| > \beta] \le 2\exp\left(-\frac{\beta^2}{16\sigma^2}\right) .$$

We also require the following result stating the random Gaussian vectors are nearly-orthogonal.

Fact 4 (Cai et al. (2013), also see Corollary D.3 in Diakonikolas et al. (2017)) *Let θ be the angle between two random unit vectors uniformly distributed over $\mathcal{S}^{d-1}$. Then, we have that $\mathbf{Pr}[|\cos\theta| \ge d^{-\alpha}] \le e^{-d^{1-2\alpha}/5}$, for any $0 \le \alpha \le 1/2$.*

Fact 5 (Gaussian Hypercontractivity Bogachev (1998); Nelson (1973)) *If $p : \mathbb{R}^m \to \mathbb{R}$ is a polynomial of degree at most k, for every $t \ge 2$,*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}\left[|p(\mathbf{x})|^t\right]^{\frac{1}{t}} \le (t-1)^{k/2}\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p^2(\mathbf{x})]} .$$

Fact 6 (Volume of d-Ball) *For any $R > 0$ let $\mathcal{S}_{d-1}(R) = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \le R\}$. Then,*

$$\mathrm{Vol}(\mathcal{S}_{d-1}) = O\left(\frac{1}{\sqrt{\pi d}}\left(\frac{2\pi e}{d}\right)^{d/2} R^d\right) .$$

Fact 7 *The following holds for the chi-square divergence between two univariate Gaussians:*

$$\chi^2(\mathcal{N}(\mu_1,\sigma_1^2),\mathcal{N}(\mu_2,\sigma_2^2)) = \frac{\sigma_2^2}{\sigma_1\sqrt{2\sigma_2^2-\sigma_1^2}} \exp\left(\frac{(\mu_1-\mu_2)^2}{2\sigma_2^2-\sigma_1^2}\right) - 1 .$$

Appendix B. Omitted Proofs from Section 4

B.1. Concentration of Gaussian Polynomials

We restate and prove the following:

Claim 1 *Let a set $S = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ of i.i.d. samples $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$. If $N > 10m^{12t}/\eta^2$, then with probability at least 0.6 for any polynomial $p : \mathbb{R}^m \to \mathbb{R}$ of degree at most t, it holds*

(i) $\mathbf{E}_{\mathbf{x}\sim\mathcal{U}(S)}[p(\mathbf{x})] \le \mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p(\mathbf{x})] + \eta$,

(ii) $\mathbf{E}_{\mathbf{x}\sim\mathcal{U}(S)}[p^2(\mathbf{x})] \ge \mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p^2(\mathbf{x})] - \eta$, *and*

(iii) $\mathbf{E}_{\mathbf{x}\sim\mathcal{U}(S)}[p^4(\mathbf{x})] \le \mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p^4(\mathbf{x})] + \eta$.

The proof follows by applying the lemma below for the polynomials p, p^2 and p^4 which are of degree $k = t, 2t$ and $4t$ respectively.

Lemma 17 *For any $\epsilon > 0$, if a set S of $N > 10\sigma^2 m^{3k}/\epsilon^2$ samples is drawn i.i.d. from $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, then with probability at least 0.9 we have that for all polynomials $p : \mathbb{R}^m \to \mathbb{R}$ with $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p^2(\mathbf{x})] \le \sigma^2$ and degree at most k it holds that*

$$\left|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{U}(S)}[p(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p(\mathbf{x})]\right| \le \epsilon .$$

Proof First, using Chebyshev's inequality, we have the following concentration for every normalized probabilist's Hermite polynomial:

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\mathbf{x}_1,\dots,\mathbf{x}_N\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}\left[\left|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{U}(S)}[h_{\mathbf{J}}(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[h_{\mathbf{J}}(\mathbf{x})]\right| > \frac{\epsilon}{m^k\sigma}\right] &\le \frac{\sigma^2 m^{2k}}{N\epsilon^2}\mathop{\mathbf{Var}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[h_{\mathbf{J}}(\mathbf{x})] \\ &= \frac{\sigma^2 m^{2k}}{N\epsilon^2}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[h_{\mathbf{J}}^2(\mathbf{x})] \\ &= \frac{\sigma^2 m^{2k}}{N\epsilon^2} \le \frac{0.1}{m^k} , \qquad (11)\end{aligned}$$

where the last line used that $N > 10\sigma^2 m^{3k}/\epsilon^2$. In what follows we condition on the event that $|\mathbf{E}_{\mathbf{x}\sim\mathcal{U}(S)}[h_{\mathbf{J}}(\mathbf{x})] - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[h_{\mathbf{J}}(\mathbf{x})]| \le \epsilon$ for all $\mathbf{J} \in \mathbb{N}^m : |\mathbf{J}| \le k$, which, by a union bound and Equation (11) holds with probability at least 0.9. We expand $p(\mathbf{x})$ on the basis of the normalized probabilist's Hermite polynomials $p(\mathbf{x}) = \sum_{\mathbf{J}\in\mathbb{N}^m:|\mathbf{J}|\leq k} a_{\mathbf{J}} h_{\mathbf{J}}(\mathbf{x})$, and note that $|a_{\mathbf{J}}| \leq \sigma$ for all these coefficients (because by Parseval's identity $\sum_{\mathbf{J}} a_{\mathbf{J}}^2 \leq \sigma^2$). Therefore, we conclude that

$$\begin{aligned}\left|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{U}(S)}[p(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[p(\mathbf{x})]\right| &\leq \sum_{\mathbf{J}\in\mathbb{N}^m:|\mathbf{J}|\leq k} |a_{\mathbf{J}}| \left|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{U}(S)}[h_{\mathbf{J}}(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_m)}[h_{\mathbf{J}}(\mathbf{x})]\right| \\ &\leq \sigma m^k \epsilon/(m^k\sigma) = \epsilon \;.\end{aligned}$$

■

B.2. Omitted Details from Proof of Item 2

Claim 2 *Let $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{m\times d}$ matrices with $\mathbf{U}\mathbf{U}^\top = \mathbf{V}\mathbf{U}^\top = \mathbf{I}_d$ and rows $\mathbf{v}_1, \dots, \mathbf{v}_m$ and $\mathbf{u}_1, \dots, \mathbf{u}_m$ respectively. Let the extension $\mathbf{v}_1, \dots, \mathbf{v}_{2m}$ of the rows of $\mathbf{V}$ to an orthonormal basis of the space spaned by $\mathbf{v}_1, \dots, \mathbf{v}_m, \mathbf{u}_1, \dots, \mathbf{u}_m$. Denote $\mathbf{R}_{\mathbf{V}_2} = [\mathbf{v}_{m+1} \dots \mathbf{v}_{2m}]^\top$. Then,*

$$\mathcal{I}_{\mathbf{U},\mathbf{V}} := \int_{\mathbf{z}\in\mathbb{R}^d} \min\{P_{A,\mathbf{U}}(\mathbf{z}), P_{A,\mathbf{V}}(\mathbf{z})\}\mathrm{d}\mathbf{z} \leq \frac{1}{\det(\mathbf{U}\mathbf{R}_{\mathbf{V}_2}^\top)} \iint_{\mathbf{x}\in\mathbb{R}^m, \mathbf{x}'\in\mathbb{R}^m} \min\{A(\mathbf{x}), A(\mathbf{x}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \;.$$

Proof We start with some notation. Denote by $\mathcal{V}$ the subspace spanned by $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$, and $\mathcal{U} = \mathrm{span}\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$. Extend the set $\mathbf{v}_1, \dots, \mathbf{v}_m$ to an orthonormal basis $\mathbf{v}_1, \dots, \mathbf{v}_m, \mathbf{v}_{m+1}, \dots, \mathbf{v}_{2m}$ of the vector space spanned by the vectors $\{\mathbf{v}_1, \dots, \mathbf{v}_m, \mathbf{u}_1, \dots, \mathbf{u}_m\}$. Furthermore, let the vectors $\mathbf{v}_1, \dots, \mathbf{v}_{2m} \dots, \mathbf{v}_d$ be the extension to an orthonormal basis of the entire $\mathbb{R}^d$. Let the matrices $\mathbf{R}_{\mathbf{V}_1} = [\mathbf{v}_1 \dots \mathbf{v}_m]^\top$ (note that $\mathbf{R}_{\mathbf{V}_1}$ coincides with $\mathbf{V}$ in this notation), $\mathbf{R}_{\mathbf{V}_2} = [\mathbf{v}_{m+1} \dots \mathbf{v}_{2m}]^\top$, and $\mathbf{R}_{\mathbf{V}_3} = [\mathbf{v}_{2m+1} \dots \mathbf{v}_d]^\top$. Let $\mathbf{R}_{\mathbf{V}} = [\mathbf{R}_{\mathbf{V}_1}^\top \mathbf{R}_{\mathbf{V}_2}^\top \mathbf{R}_{\mathbf{V}_3}^\top]^\top$.

We also define similar notation regarding $\mathbf{U}$. Namely, extend the set $\mathbf{u}_1, \dots, \mathbf{u}_m$ to an orthonormal basis $\mathbf{u}_1, \dots, \mathbf{u}_m, \mathbf{u}_{m+1}, \dots, \mathbf{u}_{2m}$ of the vector space spanned by the vectors $\{\mathbf{v}_1, \dots, \mathbf{v}_m, \mathbf{u}_1, \dots, \mathbf{u}_m\}$. Let $\mathbf{u}_1, \dots, \mathbf{u}_{2m} \dots, \mathbf{u}_d$ be its extension to an orthonormal basis of the entire $\mathbb{R}^d$. Let the matrices $\mathbf{R}_{\mathbf{U}_1} = [\mathbf{u}_1 \dots \mathbf{u}_m]^\top$, $\mathbf{R}_{\mathbf{U}_2} = [\mathbf{u}_{m+1} \dots \mathbf{u}_{2m}]^\top$, and $\mathbf{R}_{\mathbf{U}_3} = [\mathbf{u}_{2m+1} \dots \mathbf{u}_d]^\top$. Let $\mathbf{R}_{\mathbf{U}} = [\mathbf{R}_{\mathbf{U}_1}^\top \mathbf{R}_{\mathbf{U}_2}^\top \mathbf{R}_{\mathbf{U}_3}^\top]^\top$. Since $\mathbf{R}_{\mathbf{U}_3}$ and $\mathbf{R}_{\mathbf{V}_3}$ are meant to be orthonormal bases of the same space, we pick $\mathbf{R}_{\mathbf{U}_3} = \mathbf{R}_{\mathbf{V}_3}$.

We now focus on the our integral:

$$\mathcal{I}_{\mathbf{V},\mathbf{U}} \overset{\text{def}}{=} \int_{\mathbf{z}\in\mathbb{R}^d} \min\{P_{A,\mathbf{V}}(\mathbf{z}), P_{A,\mathbf{U}}(\mathbf{z})\}\mathrm{d}\mathbf{z} \;, \tag{12}$$

where $P_{A,\mathbf{V}}$ and $P_{A,\mathbf{U}}$ are defined as in Equation (1) (where recall that ϕ_k denotes the pdf of the k-dimensional standard Gaussian). Using that definition for $P_{A,\mathbf{V}}$ and the notation that we introduced earlier, we write

$$\begin{aligned} P_{A,\mathbf{V}}(\mathbf{z}) &= A(\mathbf{V}\mathbf{z})\phi_{d-m}\left(\mathrm{Proj}_{\mathcal{V}^\perp}(\mathbf{z})\right) \\ &= A(\mathbf{V}\mathbf{z})\phi_{d-m}\left([\mathbf{R}_{\mathbf{V}_2}^\top \mathbf{R}_{\mathbf{V}_3}^\top]^\top \mathbf{z}\right) \\ &= A(\mathbf{V}\mathbf{z})\phi_m\left(\mathbf{R}_{\mathbf{V}_2}\mathbf{z}\right)\phi_{d-2m}\left(\mathbf{R}_{\mathbf{V}_3}\mathbf{z}\right) , \end{aligned}$$

where in the last equality we separated the standard Gaussian into two components. Using a similar rewriting for $P_{A,\mathbf{U}}(\mathbf{z})$ along with $\mathbf{R}_{\mathbf{U}_3} = \mathbf{R}_{\mathbf{V}_3}$ (see first paragraphs), our integral is

$$\mathcal{I}_{\mathbf{V},\mathbf{U}} = \int_{\mathbf{z}\in\mathbb{R}^d} \min\{A(\mathbf{V}\mathbf{z})\phi_m\left(\mathbf{R}_{\mathbf{V}_2}\mathbf{z}\right)\phi_{d-2m}\left(\mathbf{R}_{\mathbf{V}_3}\mathbf{z}\right), A(\mathbf{U}\mathbf{z})\phi_m\left(\mathbf{R}_{\mathbf{U}_2}\mathbf{z}\right)\phi_{d-2m}\left(\mathbf{R}_{\mathbf{V}_3}\mathbf{z}\right)\}\mathrm{d}\mathbf{z} \;.$$

We rotate the space using the unitary matrix $\mathbf{R}_{\mathbf{V}}^{\top}$. Hence, Equation (12) becomes

$$\begin{aligned}\mathcal{I}_{\mathbf{V},\mathbf{U}} = \int_{\mathbf{z}\in\mathbb{R}^d} & \min\{A(\mathbf{V}\mathbf{R}_{\mathbf{V}}^{\top}\mathbf{z})\phi_m(\mathbf{R}_{\mathbf{V}_2}\mathbf{R}_{\mathbf{V}}^{\top}\mathbf{z})\phi_{d-2m}(\mathbf{R}_{\mathbf{V}_2}\mathbf{R}_{\mathbf{V}}^{\top}\mathbf{z}), \\ & A(\mathbf{U}\mathbf{R}_{\mathbf{V}}^{\top}\mathbf{z})\phi_m\left(\mathbf{R}_{\mathbf{U}_2}\mathbf{R}_{\mathbf{V}}^{\top}\mathbf{z}\right)\phi_{d-2m}\left(\mathbf{R}_{\mathbf{V}_3}\mathbf{R}_{\mathbf{V}}^{\top}\mathbf{z}\right)\}\mathrm{d}\mathbf{z}\ . \qquad (13)\end{aligned}$$

By definition of these matrices, it holds that $\mathbf{V}\mathbf{R}_{\mathbf{V}}^{\top} = [\mathbf{I}_{m\times m}\ \mathbf{0}_{m\times(d-m)}]$. Similarly it holds $\mathbf{R}_{\mathbf{V}_2}\mathbf{R}_{\mathbf{V}}^{\top} = [\mathbf{0}_{m\times m}\ \mathbf{I}_{m\times m}\ \mathbf{0}_{m\times(d-2m)}]$, and $\mathbf{R}_{\mathbf{V}_3}\mathbf{R}_{\mathbf{V}}^{\top} = [\mathbf{0}_{(d-2m)\times 2m}\ \mathbf{I}_{(d-2m)\ \times(d-2m)}]$. Using the notation $\mathbf{x}_{1\ldots k} = (x_1,\ldots,x_k)$ to denote the first k coordinates of a vector $\mathbf{x}\in\mathbb{R}^d$ with $d\geq k$, we have that $\mathbf{V}\mathbf{R}_{\mathbf{V}}^{\top}\mathbf{z} = \mathbf{z}_{1\ldots m}$, and similarly $\mathbf{R}_{\mathbf{V}_2}\mathbf{R}_{\mathbf{V}}^{\top}\mathbf{z} = \mathbf{z}_{m+1\ldots 2m}$, $\mathbf{R}_{\mathbf{V}_3}\mathbf{R}_{\mathbf{V}}^{\top}\mathbf{z} = \mathbf{z}_{2m+1\ldots d}$. Using that simplification and renaming $\mathbf{x} = \mathbf{z}_{1\ldots m}$, $\mathbf{y} = \mathbf{z}_{m+1\ldots 2m}$, $\mathbf{w} = \mathbf{z}_{2m+1\ldots d}$, the first part of the min operator in Equation (13) can be rewritten as $A(\mathbf{V}\mathbf{R}_{\mathbf{V}}^{\top}\mathbf{z})\phi_m(\mathbf{R}_{\mathbf{V}_2}\mathbf{R}_{\mathbf{V}}^{\top}\mathbf{z})\phi_{d-2m}(\mathbf{R}_{\mathbf{V}_2}\mathbf{R}_{\mathbf{V}}^{\top}\mathbf{z}) = A(\mathbf{x})\phi_m(\mathbf{y})\phi_{d-2m}(\mathbf{w})$. Using a similar reasoning for the second part of the min, we have that

$$\begin{aligned}\mathcal{I}_{\mathbf{V},\mathbf{U}} = \int & \min\{A(\mathbf{x})\phi_m(\mathbf{y})\phi_{d-2m}(\mathbf{w}), \\ & A(\mathbf{U}\mathbf{R}_{\mathbf{V}_1}^{\top}\mathbf{x} + \mathbf{U}\mathbf{R}_{\mathbf{V}_2}^{\top}\mathbf{y})\phi_m(\mathbf{R}_{\mathbf{U}_2}\mathbf{R}_{\mathbf{V}_1}^{\top}\mathbf{x} + \mathbf{R}_{\mathbf{U}_2}\mathbf{R}_{\mathbf{V}_2}^{\top}\mathbf{y})\phi_{d-2m}(\mathbf{w})\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{w} \\ = \int_{\mathbf{z}\in\mathbb{R}^d} & \min\{A(\mathbf{x})\phi_m(\mathbf{y}), A(\mathbf{U}\mathbf{R}_{\mathbf{V}_1}^{\top}\mathbf{x} + \mathbf{U}\mathbf{R}_{\mathbf{V}_2}^{\top}\mathbf{y})\phi_m(\mathbf{R}_{\mathbf{U}_2}\mathbf{R}_{\mathbf{V}_1}^{\top}\mathbf{x} + \mathbf{R}_{\mathbf{U}_2}\mathbf{R}_{\mathbf{V}_2}^{\top}\mathbf{y})\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\ , \qquad (14)\end{aligned}$$

where the last line takes $\phi_{d-2m}(\mathbf{w})$ as common factor and uses that its integral with respect to $\mathbf{w}$ equals to one. We now do the following change of integration variables:

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{x}'\end{bmatrix} = \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{U}\mathbf{R}_{\mathbf{V}_1}^{\top} & \mathbf{U}\mathbf{R}_{\mathbf{V}_2}^{\top}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\ .$$

The Jacobian of the inverse transformation is $1/\det(\mathbf{U}\mathbf{R}_{\mathbf{V}}{}_2^{\top})$ (where we used the fact that $\det(\mathbf{A}^{-1}) = 1/\det(\mathbf{A})$ as well as the fact that due to the identity block of the matrix the determinant ends up being only that of the bottom right block).

Performing this change of variables in Equation (14), and using the pointwise upper bound $\phi_m(\cdot)\leq(2\pi)^{-m/2}\leq 1$ we obtain

$$\mathcal{I}_{\mathbf{V},\mathbf{U}} \leq \frac{1}{\det(\mathbf{U}\mathbf{R}_{\mathbf{V}}{}_2^{\top})}\int_{\mathbf{x}\in\mathbb{R}^m}\int_{\mathbf{x}'\in\mathbb{R}^m}\min\{A(\mathbf{x}), A(\mathbf{x}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'\ . \qquad (15)$$

■

Claim 3 *If $d > m^C$ for a sufficiently large absolute constant C, then* $\det(\mathbf{U}\mathbf{R}_{\mathbf{V}}{}_2^{\top})\geq 1/2$.

Proof To prove this claim, we show that the singular values of the matrix $\mathbf{U}\mathbf{R}_{\mathbf{V}}{}_2^{\top}$ are close to 1. Recall that we have assumed that $\mathbf{U}\mathbf{U}^{\top} = \mathbf{V}\mathbf{V}^{\top} = \mathbf{I}_d$ and $\|\mathbf{U}\mathbf{V}^{\top}\|_{\mathrm{F}}\lesssim d^{-1/10}$. We have that

$$\begin{aligned}m = \|\mathbf{U}\|_{\mathrm{F}}^2 = \|\mathbf{U}\mathbf{R}_{\mathbf{V}}{}^{\top}\|_{\mathrm{F}}^2 &\leq \|\mathbf{U}\mathbf{R}_{\mathbf{V}}{}_1^{\top}\|_{\mathrm{F}}^2 + \|\mathbf{U}\mathbf{R}_{\mathbf{V}}{}_2^{\top}\|_{\mathrm{F}}^2 \\ &= \|\mathbf{U}\mathbf{V}^{\top}\|_{\mathrm{F}}^2 + \|\mathbf{U}\mathbf{R}_{\mathbf{V}}{}_2^{\top}\|_{\mathrm{F}}^2 \leq Cd^{-1/5} + \|\mathbf{U}\mathbf{R}_{\mathbf{V}}{}_2^{\top}\|_{\mathrm{F}}^2\ ,\end{aligned}$$

where C is some absolute positive constant. Hence, we have that $\|\mathbf{U}\mathbf{R}_{\mathbf{V}_2}^\top\|_{\mathrm{F}}^2 \geq m - Cd^{-1/5}$. Moreover, we also have that $\|\mathbf{U}\mathbf{R}_{\mathbf{V}_2}^\top\|_{\mathrm{op}} \leq 1$, which means that the maximum singular value of $\mathbf{U}\mathbf{R}_{\mathbf{V}_2}^\top$ is at most 1. Assume that the minimum singular value of $\mathbf{U}\mathbf{R}_{\mathbf{V}_2}^\top$ is σ_{min}. Then, we have that

$$m - 1 + \sigma_{\mathrm{min}}^2 \geq \|\mathbf{U}\mathbf{R}_{\mathbf{V}_2}^\top\|_{\mathrm{F}}^2 \geq m - Cd^{-1/5}\ .$$

Hence, $\sigma_{\mathrm{min}}^2 \geq 1 - Cd^{-1/5}$ and therefore, all the singular values of $\mathbf{U}\mathbf{R}_{\mathbf{V}_2}^\top$ are at least $(1 - Cd^{-1/5})^{1/2}$. Therefore, we have $\det(\mathbf{U}\mathbf{R}_{\mathbf{V}_2}^\top) \geq (1 - Cd^{-1/5})^{m/2} \geq 1 - C(m/2)d^{-1/5} \geq 1/2$ for $d > (Cm)^5$ (which is true by assumption). This completes the proof of Claim 3. ■

Claim 6 *If $R := C'\sqrt{\delta m \log(1/\delta)}$ for a sufficiently large absolute constant, then we have that $\Pr_{\mathbf{z}\sim\mathcal{N}(0,2\delta\mathbf{I}_m)}[\|\mathbf{z}\|_2 > R] \leq \delta$.*

Proof We have the series of inequalities

$$\begin{aligned}\Pr_{\mathbf{z}\sim\mathcal{N}(0,2\delta\mathbf{I}_m)}[\|\mathbf{z}\|_2 > R] &= \Pr_{\mathbf{z}\sim\mathcal{N}(0,2\delta\mathbf{I}_m)}[\|\mathbf{z}\|_2 > C'\sqrt{\delta m \log(1/\delta)}] \\ &\leq \Pr_{\mathbf{z}\sim\mathcal{N}(0,2\delta\mathbf{I}_m)}[\|\mathbf{z}\|_2 - \sqrt{\delta m} > (C'/2)\sqrt{\delta \log(1/\delta)}] \qquad (16) \\ &\leq 2\exp\left(-\frac{(C'/2)^2\delta\log(1/\delta)}{32\delta}\right) \leq \delta\ , \qquad (17)\end{aligned}$$

where Equation (16) uses the fact that $C'\sqrt{\delta m \log(1/\delta)} - \sqrt{\delta m} = \sqrt{\delta m}(C'\sqrt{\log(1/\delta)} - 1) \geq (C'/2)\sqrt{\delta m \log(1/\delta)} \geq (C'/2)\sqrt{\delta \log(1/\delta)}$ with the penultimate step being true because C' large enough and $\delta < 0.1$. The last step in Equation (17) uses Fact 3 with $\beta = (C'/2)\sqrt{\delta\log(1/\delta)}$. ■

Claim 4 *The following bound holds: $\mathcal{I}_{\mathbf{V},\mathbf{U}} \leq 2k \max_{i,j\in[k]} \iint_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^m} \min\{A_i(\mathbf{x}), A_j(\mathbf{x}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'$.*

Proof We have the following series of inequalities (see below for step-by-step explanations):

$$\begin{aligned}\mathcal{I}_{\mathbf{V},\mathbf{U}} &\lesssim \iint_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^m} \min\{A(\mathbf{x}), A(\mathbf{x}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \\ &= \iint_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^m} \min\left\{\sum_{i\in[k]}\lambda_i A_i(\mathbf{x}), \sum_{j\in[k]}\lambda_j A_j(\mathbf{x}')\right\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \\ &\leq \sum_{i,j\in[k]}\iint_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^m} \min\{\lambda_i A_i(\mathbf{x}), \lambda_j A_j(\mathbf{x}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \qquad (18) \\ &\leq \sum_{i,j\in[k]}\iint_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^m} \max\{\lambda_i,\lambda_j\}\min\{A_i(\mathbf{x}), A_j(\mathbf{x}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \\ &\leq \sum_{i,j\in[k]}\iint_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^m} \lambda_i\min\{A_i(\mathbf{x}), A_j(\mathbf{x}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' + \sum_{i,j\in[k]}\iint_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^m} \lambda_j\min\{A_i(\mathbf{x}), A_j(\mathbf{x}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \quad (19) \\ &= k\sum_{i,j\in[k]}\iint_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^m} (\lambda_i/k)\min\{A_i(\mathbf{x}), A_j(\mathbf{x}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' + \sum_{i,j\in[k]}\iint_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^m} (\lambda_j/k)\min\{A_i(\mathbf{x}), A_j(\mathbf{x}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'\end{aligned}$$

$$\leq 2k \max_{i,j\in[k]} \iint_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^m} \min\{A_i(\mathbf{x}), A_j(\mathbf{x}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \,, \tag{20}$$

where Equation (18) uses that $\min(a+b,c) \leq \min(a,c)+\min(b,c)$, Equation (19) uses that $\max(a,b) \leq a+b$, and for the last step, one can view the double sumation in the first term of the penultimate line as an expectation over the random choice of the indices i,j according to the distribution that selects j uniformly at random from $[k]$ and makes i equal to ℓ with probability λ_ℓ. Similar argument can be used for the second term of the penultimate line. Since the expectation is always smaller than the maximum value the last line follows. ■

Claim 5 $\mathcal{I}_{\mathbf{V},\mathbf{U}} \leq C^m k\delta^{0.4m}$ *for a sufficiently large absolute constant* C.

Proof Let $\iint_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^m} \min\{A_i(\mathbf{x}), A_j(\mathbf{x}')\} = \mathcal{I}_1+\mathcal{I}_2+\mathcal{I}_3+\mathcal{I}_4$, where

1. $\mathcal{I}_1 = \iint_{\|\mathbf{x}-\boldsymbol{\mu}_i\|_2 > R \text{ and } \|\mathbf{x}'-\boldsymbol{\mu}_j\|_2 \leq R} \min\{A_i(\mathbf{x}), A_j(\mathbf{x}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'$,
2. $\mathcal{I}_2 = \iint_{\|\mathbf{x}-\boldsymbol{\mu}_i\|_2 \leq R \text{ and } \|\mathbf{x}'-\boldsymbol{\mu}_j\|_2 > R} \min\{A_i(\mathbf{x}), A_j(\mathbf{x}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'$,
3. $\mathcal{I}_3 = \iint_{\|\mathbf{x}-\boldsymbol{\mu}_i\|_2 > R \text{ and } \|\mathbf{x}'-\boldsymbol{\mu}_j\|_2 > R} \min\{A_i(\mathbf{x}), A_j(\mathbf{x}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'$,
4. $\mathcal{I}_4 = \iint_{\|\mathbf{x}-\boldsymbol{\mu}_i\|_2 \leq R \text{ and } \|\mathbf{x}'-\boldsymbol{\mu}_j\|_2 \leq R} \min\{A_i(\mathbf{x}), A_j(\mathbf{x}')\}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'$.

We start with the first term. Recall that A_i is an m-dimensional Gaussian with mean $\boldsymbol{\mu}_i$ and variance δ in all directions. We have the following:

$$\begin{aligned}
\mathcal{I}_1 &\leq \int_{\|\mathbf{x}-\boldsymbol{\mu}_i\|_2>R} \sqrt{A_i(\mathbf{x})}\mathrm{d}\mathbf{x} \int_{\|\mathbf{x}'-\boldsymbol{\mu}_j\|_2\leq R} \sqrt{A_j(\mathbf{x}')}\mathrm{d}\mathbf{x}' && \text{(using } \min(a,b)\leq\sqrt{ab}\text{)}\\
&\leq \int_{\|\mathbf{x}-\boldsymbol{\mu}_i\|_2>R} (2\pi\delta)^{-m/4} e^{-\frac{\|\mathbf{x}-\boldsymbol{\mu}_i\|_2^2}{4\delta}} \mathrm{d}\mathbf{x} \int_{\|\mathbf{x}'-\boldsymbol{\mu}_j\|_2\leq R} (2\pi\delta)^{-m/4} e^{-\frac{\|\mathbf{x}'-\boldsymbol{\mu}_j\|_2^2}{4\delta}} \mathrm{d}\mathbf{x}'\\
&\leq (2\pi\delta)^{m/4} \int_{\|\mathbf{x}-\boldsymbol{\mu}_i\|_2>R} (2\pi\delta)^{-m/2} e^{-\frac{\|\mathbf{x}-\boldsymbol{\mu}_i\|_2^2}{4\delta}} \mathrm{d}\mathbf{x} \int_{\|\mathbf{x}'-\boldsymbol{\mu}_j\|_2\leq R} (2\pi\delta)^{-m/4} e^{-\frac{\|\mathbf{x}'-\boldsymbol{\mu}_j\|_2^2}{4\delta}} \mathrm{d}\mathbf{x}'\\
&\leq (2\pi\delta)^{m/4}\delta \cdot \delta^{-m/4}\mathrm{Vol}(\mathcal{S}_{d-1}(R)) && \text{(using Equation (17) for the first inegral)}\\
&\leq (2\pi)^{m/4}\delta \left(\frac{1}{\sqrt{\pi m}} \left(\frac{2\pi e}{m}\right)^{m/2} R^m \right) && \text{(by Fact 6)}\\
&\leq C_1^m m^{-m/2}\delta^{1+m/2} m^{m/2} (\log(1/\delta))^{m/2} && \text{(using } R = C'\sqrt{\delta m\log(1/\delta)}\text{)}\\
&\leq C_1^m \delta^{1+m/2} (\log(1/\delta))^{m/2} && (21)
\end{aligned}$$

for a sufficiently large constant C_1. The same bound can be derived for $\mathcal{I}_2$. For $\mathcal{I}_3$ we use Equation (17) for both integrals to obtain $\mathcal{I}_3 \leq (2\pi\delta)^{m/2}\delta^2$. Finally, for the last term $\mathcal{I}_4$ we have that $\mathcal{I}_4 \leq \delta^{-m/2}(\mathrm{Vol}(\mathcal{S}_{d-1}(R)))^2 \leq C_2^m \delta^{-m/2} m^{-m}\delta^m m^m (\log(1/\delta))^m \leq C_2^m \delta^{m/2}(\log(1/\delta))^m$, where the first step used $\min\{A_i(\mathbf{x}), A_j(\mathbf{x})\} \leq \delta^{-m/2}$ and that both integrals are over a ball of radius R. Putting everything together, we have shown that

$$\mathcal{I}_{\mathbf{V},\mathbf{U}} \leq C_3^m k\delta^{m/2}(\log(1/\delta))^m \leq C_4^m k\delta^{0.4m} \,. \tag{22}$$

■

Claim 7 *In the setting of Proposition 9 it holds* $\mathrm{d}_{\mathrm{TV}}(P_{A,\mathbf{V}}, \mathcal{N}(\mathbf{0}, \mathbf{I}_d)) > 0.99$.

Proof Let $\mathbf{v}_1, \dots, \mathbf{v}_m$ denote the rows of $\mathbf{V}$ and extend this set to an orthonormal basis $\mathbf{v}_1, \dots, \mathbf{v}_m, \dots, \mathbf{v}_d$ of the entire $\mathbb{R}^d$. Let $\mathbf{V}^{\perp}$ be the matrix having $\mathbf{v}_{m+1}, \dots, \mathbf{v}_d$ as rows and $\mathbf{R}$ be the matrix having $\mathbf{v}_1, \dots, \mathbf{v}_m, \dots, \mathbf{v}_d$ as rows. Using the definition from Equation (1) (and recalling that $\phi_d(\mathbf{x})$ denotes the pdf of $\mathcal{N}(0, \mathbf{I}_d)$,

$$P_{A,\mathbf{V}(\mathbf{z})} = A(\mathbf{V}\mathbf{z})\phi_{d-m}\left(\mathrm{Proj}_{\mathcal{V}^{\perp}}(\mathbf{z})\right) = A(\mathbf{V}\mathbf{z})\phi_{d-m}\left(\mathbf{V}^{\perp}\mathbf{z}\right) .$$

As before, we examine the integral $\mathcal{I} := \int_{z \in \mathbb{R}^d} \min\left\{P_{A,\mathbf{V}}(\mathbf{z}), \phi_d(\mathbf{z})\right\} \mathrm{d}\mathbf{z}$ for which we have the following:

$$\begin{aligned}
\mathcal{I} &= \int_{z \in \mathbb{R}^d} \min\left\{P_{A,\mathbf{V}}(\mathbf{z}), \phi_d(\mathbf{z})\right\} \mathrm{d}\mathbf{z} \\
&= \int_{z \in \mathbb{R}^d} \min\left\{A(\mathbf{V}\mathbf{z})\phi_{d-m}\left(\mathbf{V}^{\perp}\mathbf{z}\right), \phi_m\left(\mathrm{Proj}_{\mathcal{V}}(\mathbf{z})\right)\phi_{d-m}\left(\mathrm{Proj}_{\mathcal{V}^{\perp}}(\mathbf{z})\right)\right\} \mathrm{d}\mathbf{z} \\
&= \int_{z \in \mathbb{R}^d} \min\left\{A(\mathbf{V}\mathbf{z})\phi_{d-m}\left(\mathbf{V}^{\perp}\mathbf{z}\right), \phi_m\left(\mathbf{V}\mathbf{z}\right)\phi_{d-m}\left(\mathbf{V}^{\perp}\mathbf{z}\right)\right\} \mathrm{d}\mathbf{z} \\
&= \int_{z \in \mathbb{R}^d} \min\left\{A(\mathbf{V}\mathbf{R}^{\top}\mathbf{z})\phi_{d-m}\left(\mathbf{V}^{\perp}\mathbf{R}^{\top}\mathbf{z}\right), \phi_m\left(\mathbf{V}\mathbf{R}^{\top}\mathbf{z}\right)\phi_{d-m}\left(\mathbf{V}^{\perp}\mathbf{R}^{\top}\mathbf{z}\right)\right\} \mathrm{d}\mathbf{z} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by rotating space by } \mathbf{R}^{\top}\text{)} \\
&= \int_{z \in \mathbb{R}^d} \min\left\{A(z_1, \dots, z_m)\phi_{d-m}\left(z_{m+1}, \dots, z_d\right), \phi_m\left(z_1, \dots, z_m\right)\phi_{d-m}\left(z_{m+1}, \dots, z_d\right)\right\} \mathrm{d}\mathbf{z} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(using the definition of matrices } \mathbf{V}, \mathbf{V}^{\perp}, \mathbf{R}\text{)} \\
&= \int_{(z_1, \dots, z_m) \in \mathbb{R}^m} \min\left\{A(z_1, \dots, z_m), \phi_m\left(z_1, \dots, z_m\right)\right\} \mathrm{d}z_1 \cdots \mathrm{d}z_m \\
&= \int_{\mathbf{x} \in \mathbb{R}^m} \min\left\{A(\mathbf{x}), \phi_m\left(\mathbf{x}\right)\right\} \mathrm{d}\mathbf{x} && \text{(by renaming } \mathbf{x} = (z_1, \dots, z_m)\text{)} \\
&= \int_{\mathbf{x} \in \mathbb{R}^m} \min\left\{\sum_{i=1}^{k} \lambda_i A_i(\mathbf{x}), \phi_m\left(\mathbf{x}\right)\right\} \mathrm{d}\mathbf{x} && (A = \textstyle\sum_{i \in [k]} \lambda_i A_i) \\
&\leq \sum_{i=1}^{k} \int_{\mathbf{x} \in \mathbb{R}^m} \min\left\{\lambda_i A_i(\mathbf{x}), \phi_m\left(\mathbf{x}\right)\right\} \mathrm{d}\mathbf{x} && \text{(using } \min(a+b, c) \leq \min(a, c) + \min(b, c)\text{)} \\
&\leq k \max_{i \in [k]} \int_{\mathbf{x} \in \mathbb{R}^m} \min\left\{A_i(\mathbf{x}), \phi_m\left(\mathbf{x}\right)\right\} \mathrm{d}\mathbf{x} \, , && (23)
\end{aligned}$$

where the last step uses that $\lambda_i \leq 1$. Now, $A_i = \mathcal{N}(\boldsymbol{\mu}_i, \delta\mathbf{I}_m)$ with $\|\boldsymbol{\mu}_i\|_2 \geq 0.9\sqrt{m}$ by Item (iv) of Proposition 10 and δ is smaller than 1, thus we have that $\int_{\mathbf{x} \in \mathbb{R}^m} \min\left\{A_i(\mathbf{x}), \phi_m\left(\mathbf{x}\right)\right\} \mathrm{d}\mathbf{x} = 1 - \mathrm{d}_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}_i, \delta\mathbf{I}_m), \mathcal{N}(\mathbf{0}, \mathbf{I}_m)) \leq 1 - \mathrm{d}_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{I}_m), \mathcal{N}(\mathbf{0}, \mathbf{I}_m))$. By a rotation argument similar to what we did earlier, the contribution comes only from the error along the direction that connects the origin to the point $\boldsymbol{\mu}_i$

$$1 - \mathrm{d}_{\mathrm{TV}}\left(\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{I}_m), \mathcal{N}(\mathbf{0}, \mathbf{I}_m)\right) = 1 - \mathrm{d}_{\mathrm{TV}}\left(\mathcal{N}(\|\boldsymbol{\mu}_i\|_2, 1), \mathcal{N}(0, 1)\right) = \mathrm{erfc}\left(\frac{\|\boldsymbol{\mu}_i\|_2}{2\sqrt{2}}\right)$$

$$\leq \operatorname{erfc}\left(\sqrt{m}/4\right) \leq \frac{1}{100k} \; ,$$

where the last step requires $m > C\log(k)$, which is true since $m = k^{2\epsilon}$ and we have assumed $k^{\epsilon} > C\sqrt{\log k}$. Putting everything together and combining with the bound of Equation (23) we conclude that $\mathrm{d_{TV}}(P_{A,\mathbf{V}}, \mathcal{N}(0,\mathbf{I}_d)) = 1 - \int_{z\in\mathbb{R}^d} \min\left\{P_{A,\mathbf{V}}(\mathbf{z}), \phi_d(\mathbf{z})\right\} \mathrm{d}\mathbf{z} \geq 1 - k/(100k) = 0.99$. ■

Claim 8 *In the setting of Proposition 10, it holds* $\chi^2(A, \mathcal{N}(\mathbf{0},\mathbf{I}_m) \leq \delta^{-m/2} e^{O(m)}$.

Proof We first focus on a single component A_i, which is an isotropic Gaussian with mean $\boldsymbol{\mu}_i = (\mu_{i,1},\ldots,\mu_{i,m})$ and variance δ. Because both A_i and the standard Gaussian are product distributions in m dimensions, the integral in the definition of the $\chi^2(A_i, \mathcal{N}(\mathbf{0},\mathbf{I}_m)$ is separable and we can use Fact 7 for each coordinate. Concretely, let ϕ denote the pdf of $\mathcal{N}(0,1)$:

$$\begin{aligned}
1 + \chi^2(A_i, \mathcal{N}(0,\mathbf{I}_m)) &= \int_{\mathbf{x}\in\mathbb{R}^m} \frac{A_i^2(\mathbf{x})}{\phi(x_1)\cdots\phi(x_m)} \mathrm{d}\mathbf{x} = \prod_{j=1}^m \int_{x_j\in\mathbb{R}} \frac{\frac{1}{2\pi\delta}\exp\left(-\frac{(x_j-\mu_{i,j})^2}{\delta}\right)}{\phi(x_j)} \mathrm{d}x_j \\
&= \prod_{j=1}^m (1 + \chi^2(\mathcal{N}(\mu_{i,j},\delta), \mathcal{N}(0,1)) = \frac{1}{(\delta(2-\delta))^{m/2}} \exp\left(\frac{\|\boldsymbol{\mu}_i\|_2^2}{2-\delta}\right) \\
&\leq \delta^{-m/2} e^{1.21m} \; ,
\end{aligned}$$

where the last line uses that $\delta < 1$ and $\|\boldsymbol{\mu}_i\|_2 \leq 1.1\sqrt{m}$ by Item (iv) of Proposition 9. Denote by w_i the weights in the mixture $A = \sum_{i=1}^k w_i A_i$. Also, by using $\phi_m(\mathbf{x})$ to denote the pdf of $\mathcal{N}(\mathbf{0},\mathbf{I}_m)$ we have that

$$\begin{aligned}
1 + \chi^2(A, \mathcal{N}(\mathbf{0},\mathbf{I}_m)) &= \sum_{i=1}^k \sum_{j=1}^k w_i w_j \int_{\mathbf{x}\in\mathbb{R}^m} \frac{A_i(\mathbf{x})A_j(\mathbf{x})}{\phi_m(\mathbf{x})} \mathrm{d}\mathbf{x} \\
&\leq \sum_{i=1}^k \sum_{j=1}^k w_i w_j \sqrt{\int_{\mathbf{x}\in\mathbb{R}^m} \frac{A_i(\mathbf{x})^2}{\phi(\mathbf{x})} \mathrm{d}\mathbf{x} \int_{\mathbf{x}\in\mathbb{R}^m} \frac{A_j(\mathbf{x})^2}{\phi_m(\mathbf{x})} \mathrm{d}\mathbf{x}} \\
&= \sum_{i=1}^k \sum_{j=1}^k w_i w_j \sqrt{(1 + \chi^2(A_i, \mathcal{N}(0,\mathbf{I}_m)))\,(1 + \chi^2(A_j, \mathcal{N}(0,\mathbf{I}_m)))} \\
&\leq \delta^{-m/2} e^{1.21m} \sum_{i=1}^k \sum_{j=1}^k w_i w_j = \delta^{-m/2} e^{1.21m} \; ,
\end{aligned}$$

where the second line uses Cauchy–Schwartz inequality, and the last line uses the upper bound for $1 + \chi^2(A_i, \mathcal{N}(\mathbf{0},\mathbf{I}_m))$ that we showed in the beginning. ■

Appendix C. Omitted Proofs from Section 5

Lemma 12 (Moment Matching) *There exists a discrete distribution D on $\mathbb{R}^m$ such that: (i) D is supported on $2m$ points, (ii) D matches the first three moments with $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, and (iii) for every pair of distinct points $\mathbf{x}, \mathbf{y}$ in the support of D, it holds $\|\mathbf{x} - \mathbf{y}\|_2 \geq \sqrt{m}$.*

Proof Let $\mathbf{e}_i$ for $i \in [m]$ denote the i-th vector of the standard basis of $\mathbb{R}^m$, i.e., the vector having 1 in the i-th coordinate and zero everywhere else. Let the set of vectors $S = \{\mathbf{x}_1, \dots, \mathbf{x}_{2m}\}$ defined as $\mathbf{x}_i = \sqrt{m/2}\,\mathbf{e}_i$ for $i = 1, \dots, m$, and $\mathbf{x}_i = -\sqrt{m/2}\,\mathbf{e}_{i-m}$ for $i = m+1, \dots, 2m$.

It is easy to verify that $D = \mathcal{U}(S)$, the uniform distribution on these points, matches the first three moments with $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$: Let p be a polynomial of degree at most 3, i.e., $p(x_1, \dots, x_m) = x_1^a x_2^b x_3^c$, with $a+b+c \leq 3$ (without loss of generality, we assumed that the coordinates from $[m]$ with non-zero power are the first three). If either of a, b, c is equal to 1 or 3, then $\mathbf{E}_{\mathbf{x} \sim \mathcal{U}(S)}[p(\mathbf{x})] = 0$, because we made S symmetric about the origin. This only leaves the case $p(x_1, \dots, x_m) = x_1^2$, where we have $\mathbf{E}_{\mathbf{x} \sim \mathcal{U}(S)}[p(\mathbf{x})] = 1$, because the first coordinate is equal to $\sqrt{m/2}$ and $-\sqrt{m/2}$ only for two points in S and zero for every other one. This completes the proof. ■

Lemma 13 *Let C be a sufficiently large absolute constant. Let $c \in (0, 1/4)$ and $m, d \in \mathbb{N}$ with $d > (1/c)^{C/c}$ and $m < d^{c/5}/C$. There exists a set S of $2^{\Omega(d^{2c})}$ matrices in $\mathbb{R}^{m \times d}$ such that every $\mathbf{A} \in S$ satisfies $\mathbf{A}\mathbf{A}^\top = \mathbf{I}_m$ and every pair $\mathbf{A}, \mathbf{A}' \in S$ with $\mathbf{A} \neq \mathbf{A}'$ satisfies $\|\mathbf{A}'\mathbf{A}^\top\|_{\mathrm{op}} \lesssim d^{-1/2+2c}$.*

Proof We will use the following basic fact:

Fact 8 *For any $0 < c < 1/2$, there exists a set S' of $2^{\Omega(d^{2c})}$ unit vectors in $\mathbb{R}^d$, such that any pair $\mathbf{u}, \mathbf{v} \in S'$ with $\mathbf{u} \neq \mathbf{v}$ satisfies $|\mathbf{u}^\top \mathbf{v}| \lesssim d^{-1/2+c}$.*

Let $S' = \{\mathbf{u}_1, \dots, \mathbf{u}_{|S'|}\}$ be the set of vectors from the fact above. Let S'' be the set of matrices $\{\mathbf{B}_i\}_{i=1}^{|S'|/m}$ for where $\mathbf{B}_i$ is defined to have as rows the vectors $\mathbf{u}_j$ for $j = (i-1) \cdot m + 1, \dots i \cdot m$. Note that $|S'|/m = 2^{\Omega(d^{2c})}$ for any $d > (1/c)^{C/c}$ where C is a sufficiently large constant. Finally, let S be the set of matrices $\{\mathbf{A}_i\}_{i=1}^{|S'|/m}$ where for each $\mathbf{B}_i \in S''$ we consider the Singular Value Decomposition $\mathbf{B}_i = \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i^\top$ and we let $\mathbf{A}_i$ be the matrix obtained by replacing the diagonal matrix $\mathbf{\Sigma}_i$ with identity (i.e., changing all singular values to 1). We will show that S is the set of matrices satisfying the desideratum of Lemma 13.

In particular, we claim the following. Let C be a sufficiently large absolute constant, then:

(i) For every $i \in |S''|$, all singular values of $\mathbf{B}_i$ belong in $[1 - Cm^2 d^{-1/2+c}, 1 + Cm^2 d^{-1/2+c}]$.

(ii) For every $i \in |S''|$, it holds $\|\mathbf{A}_i - \mathbf{B}_i\|_{\mathrm{F}} \lesssim m^{2.5} d^{-1/2+c}$.

(iii) For every $i, j = 1, \dots, |S''|$, it holds $\|\mathbf{B}_i \mathbf{B}_j^\top\|_{\mathrm{op}} \lesssim m^2 d^{-1/2+c}$.

Given the above, the proof of Lemma 13 follows immediately by noting that

$$\begin{aligned}
\|\mathbf{A}_i \mathbf{A}_j^\top\|_{\mathrm{op}} &= \|(\mathbf{B}_i + \mathbf{A}_i - \mathbf{B}_i)(\mathbf{B}_j + \mathbf{A}_j - \mathbf{B}_j)^\top\|_{\mathrm{op}} \\
&\leq \|\mathbf{B}_i \mathbf{B}_j^\top\|_{\mathrm{op}} + \|\mathbf{B}_i (\mathbf{A}_j - \mathbf{B}_j)^\top\|_{\mathrm{op}} + \|(\mathbf{A}_i - \mathbf{B}_i)\mathbf{B}_j^\top\|_{\mathrm{op}} + \|(\mathbf{A}_i - \mathbf{B}_i)(\mathbf{A}_j - \mathbf{B}_j)^\top\|_{\mathrm{op}}
\end{aligned}$$

$$
\begin{aligned}
&\leq \|\mathbf{B}_i\mathbf{B}_j^\top\|_{\text{op}} + \|\mathbf{B}_i\|_{\text{op}}\|\mathbf{A}_j - \mathbf{B}_j\|_{\text{F}} + \|\mathbf{B}_j^\top\|_{\text{op}}\|\mathbf{A}_i - \mathbf{B}_i\|_{\text{F}} + \|\mathbf{A}_i - \mathbf{B}_i\|_{\text{F}}\|\mathbf{A}_j - \mathbf{B}_j\|_{\text{F}} \\
&\lesssim m^2 d^{-1/2+c} + m^3 d^{-1/2+c} + m^5 d^{-1/4+2c} \\
&\lesssim d^{-1/2+2c} \,,
\end{aligned}
$$

where the second line uses triangle inequality, the third line uses the sub-multiplicative property of the operator norm , i.e., that $\|\mathbf{U}\mathbf{V}\|_{\text{op}} \leq \|\mathbf{U}\|_{\text{op}}\|\mathbf{V}\|_{\text{op}}$ as well as the fact $\|\mathbf{V}\|_{\text{op}} \leq \|\mathbf{V}\|_{\text{F}}$, the fourth line uses our three claims (that we show later on) and the last line uses our assumption $m \ll d^{c/5}$.

We now prove the three claims. For Item (i), consider the matrix $\mathbf{B}_i\mathbf{B}_i^\top$ (which is a square $m \times m$ matrix). Using Fact 8, the sum of the absolute values of its non-diagonal entries is

$$
R = \sum_{k \neq \ell} |\mathbf{u}_{(i-1)m+k}^\top \mathbf{u}_{(i-1)m+\ell}| \lesssim m^2 d^{-1/2+c} \,.
$$

The diagonal entries of $\mathbf{B}_i\mathbf{B}_i^\top$ are all equal to one. Thus, by the Gershgorin's disc theorem **??**, every eigenvalue of $\mathbf{B}_i\mathbf{B}_i^\top$, i.e., singular value of $\mathbf{B}_i$, lies the interval $[1-R, 1+R]$.

For proving Item (ii), we note that

$$
\|\mathbf{A}_i - \mathbf{B}_i\|_{\text{F}} = \sqrt{\sum_{k=1}^{m} (\sigma_k(\mathbf{B}_i) - 1)^2} \leq \sqrt{m \cdot (R-1)^2} \lesssim m^{2.5} d^{-1/2+c} \,.
$$

Finally, regarding Item (iii), for every $i, j \in [|S''|]$ with $i \neq j$, we have that

$$
\begin{aligned}
\|\mathbf{B}_i\mathbf{B}_j^\top\|_{\text{op}} &\leq \sup_{\mathbf{z} \in \mathcal{S}^{m-1}} \mathbf{z}^\top \mathbf{B}_i \mathbf{B}_j^\top \mathbf{z} \leq \sup_{\mathbf{z} \in \mathcal{S}^{m-1}} \left\langle \sum_{k \in [m]} z_k \mathbf{u}_{(i-1)m+k}, \sum_{\ell \in [m]} z_\ell \mathbf{u}_{(j-1)m+\ell} \right\rangle \\
&\leq \sup_{\mathbf{z} \in \mathcal{S}^{m-1}} \sum_{k,\ell \in [m]} z_k z_\ell \left\langle \mathbf{u}_{(i-1)m+k}, \mathbf{u}_{(j-1)m+\ell} \right\rangle \\
&\lesssim d^{-1/2+c} \sup_{z \in \mathcal{S}^{m-1}} \sum_{k,\ell \in [m]} z_k z_\ell \lesssim m^2 d^{-1/2+c} \,,
\end{aligned}
$$

where the last line uses Fact 8. ■

Given the above lematta, we can now conclude with the proof of Theorem 11.

Proof [Proof of Theorem 11] Let C be a sufficiently large constant. Let D be the distribution from Lemma 12 with $m := k/2$ and $A = U_\rho D$ for $\delta = k^{-2.5/m}/C$, where U_ρ denotes the Ornstein-Uhlenbeck operator with parameter ρ. We choose $\rho = \sqrt{1-\delta}$.

The above means that A is a mixture of k equally weighted spherical Gaussians in $\mathbb{R}^m$, each with variance δ in every direction. By Lemma 12, the mean separation is $\rho \cdot \sqrt{k/2} = \sqrt{1 - k^{-2.5/k}/C}\sqrt{k/2} \geq \sqrt{k}/3$ for any $k \geq 2$.

The following can be shown by repeating mutatis-mutandis the same steps we followed while proving Proposition 9:

1. The first 3 moments of A match with those of $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$.
2. For every $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{m \times d}$ with $\mathbf{U}\mathbf{U}^\top = \mathbf{V}\mathbf{V}^\top = \mathbf{I}_d$ and $\|\mathbf{U}\mathbf{V}^\top\|_{\text{F}} = O(d^{-1/2+2c})$, it holds $\mathrm{d}_{\mathrm{TV}}(P_{A,\mathbf{U}}, P_{A,\mathbf{V}}) > 0.99$. Moreover, for all $\mathbf{V} \in \mathbb{R}^{m \times d}$ it holds $\mathrm{d}_{\mathrm{TV}}(P_{A,\mathbf{V}}, \mathcal{N}(0, \mathbf{I}_d)) > 0.99$.

3. $\chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}_m) \leq e^{O(k)}$.

Now by also following the same steps as in the proof of Theorem 7, but replacing Fact 1 by Lemma 13, we obtain that every SQ algorithm for solving our hypothesis testing problem, either needs $2^{\Omega(d^{2c})}$ queries or at least one query to

$$\mathrm{VSTAT}(\Omega(d^{2-8c})/\chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}_m))) \,.$$

We note that $\Omega(d^{2-8c})/\chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}_m)) \geq \Omega(d^{2-8c})e^{-O(k)} \geq \Omega(d^{2-9c})$, where the last inequality uses our assumption $k < (c/C)\log d$. Also note that Lemma 13 was indeed applicable, since its requirement $m < d^{c/5}/C$ is satisfied because we have $m := k/2 < 0.5(c/C)\log d < d^{c/5}/C$, where the first inequality is one of our assumptions and the second follows by our other assumption $d > (1/c)^{C/c}$. This completes the proof of Theorem 11. ■

Appendix D. Lower Bounds for Low-Degree Polynomial Tests

Problem 18 *Let a distribution A on $\mathbb{R}^m$. For a matrix $\mathbf{V} \in \mathbb{R}^{m \times d}$, we let $P_{A,\mathbf{V}}$ be the distribution as in Equation (1), i.e., the distribution that coincides with A on the subspace spanned by the rows of $\mathbf{V}$ and is standard Gaussian in the orthogonal subspace. Let S be the set of nearly orthogonal vectors from Fact 1. Let $\mathcal{S} = \{P_{A,v}\}_{u \in S}$. We define the simple hypothesis testing problem where the null hypothesis is $\mathcal{N}(\mathbf{0}, I_d)$ and the alternative hypothesis is $P_{A,\mathbf{V}}$ for some $\mathbf{V}$ uniformly selected from S.*

We now describe the model in more detail. We will consider tests that are thresholded polynomials of low-degree, i.e., output H_1 if the value of the polynomial exceeds a threshold and H_0 otherwise. We need the following notation and definitions. 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 $f : \mathcal{X} \to \mathbb{R}$, $g : \mathcal{X} \to R$ 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 $f(x_1, \ldots, x_n) : \mathbb{R}^{n \times d} \to \mathbb{R}$ has sample-wise degree (r, ℓ) if each monomial uses at most ℓ different samples from $x_1, \ldots, x_n$ and uses degree at most r for each of them. Let $\mathcal{C}_{r,\ell}$ be linear space of all polynomials of sample-wise degree (r, ℓ) with respect to the inner product defined above. For a function $f : \mathbb{R}^{n \times d} \to \mathbb{R}$, we use $f^{\leq r, \ell}$ to be the orthogonal projection onto $\mathcal{C}_{r,\ell}$ with respect to the inner product $\langle \cdot, \cdot \rangle_{D_0^{\otimes n}}$. Finally, for the null distribution D_0 and a distribution P, define the likelihood ratio $\overline{P}^{\otimes n}(x) := P^{\otimes n}(x)/D_0^{\otimes n}(x)$.

Definition 19 (n-sample τ-distinguisher) *For the hypothesis testing problem between D_0 (null distribution) and D_1 (alternate distribution) over $\mathcal{X}$, we say that a function $p : \mathcal{X}^n \to \mathbb{R}$ is an n-sample τ-distinguisher if $|\mathbf{E}_{X \sim D_0^{\otimes n}}[p(X)] - \mathbf{E}_{X \sim D_1^{\otimes n}}[p(X)]| \geq \tau\sqrt{\mathbf{Var}_{X \sim D_0^{\otimes n}}[p(X)]}$. We call τ the* advantage *of the polynomial p.*

Note that if a function p has advantage τ, then the Chebyshev's inequality implies that one can furnish a test $p' : \mathcal{X}^n \to \{D_0, D_1\}$ by thresholding p such that the probability of error under the null distribution is at most $O(1/\tau^2)$. We will think of the advantage τ as the proxy for the inverse of the probability of error (see Theorem 4.3 in Kunisky et al. (2022) for a formalization of this intuition under certain assumptions) and we will show that the advantage of all polynomials up to a certain degree is $O(1)$. It can be shown that for hypothesis testing problems of the form of Problem 18, the best possible advantage among all polynomials in $\mathcal{C}_{r,\ell}$ is captured by the low-degree likelihood ratio (see, e.g., Brennan et al. (2021); Kunisky et al. (2022)):

$$\left\| \mathop{\mathbf{E}}_{v\sim\mathcal{U}(S)} \left[\left(\overline{P}_{A,\mathbf{V}}^{\otimes n} \right)^{\leq r,\ell} \right] - 1 \right\|_{D_0^{\otimes n}} ,$$

where in our case $D_0 = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.

To show that the low-degree likelihood ratio is small, we use the result from Brennan et al. (2021) stating that a lower bound for the SQ dimension translates to an upper bound for the low-degree likelihood ratio. Therefore, given that we have already established in previous section that $\mathrm{SD}(\mathcal{B}(\{P_{A,\mathbf{V}}\}_{\mathbf{V}\in S}, \mathcal{N}(\mathbf{0}, \mathbf{I}_d)), \gamma, \beta) = 2^{d^c}$ for $\gamma = \Omega(d)^{(t+1)/10}\chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}_d))$ and $\beta = \chi^2(A, \mathcal{N}(0, 1))$, we one can obtain the corollary:

Theorem 20 *Let a sufficiently small positive constant c. Let the hypothesis testing problem of Problem 18 the distribution A matches the first t moments with $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$. For any $d \in \mathbb{Z}_+$ with $d = t^{\Omega(1/c)}$, any $n \leq \Omega(d)^{(t+1)/10}/\chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}_m))$ and any even integer $\ell < d^c$, we have that*

$$\left\| \mathop{\mathbf{E}}_{v\sim\mathcal{U}(S)} \left[\left(\overline{P}_{A,\mathbf{V}}^{\otimes n} \right)^{\leq \infty,\ell} \right] - 1 \right\|_{D_0^{\otimes n}} \leq 1 \, .$$

The interpretation of this result is that unless the number of samples used n is greater than $\Omega(d)^{(t+1)/10}/\chi^2(A, \mathcal{N}(\mathbf{0}, \mathbf{I}_m))$, any polynomial of degree roughly up to d^c fails to be a good test (note that any polynomial of degree ℓ has sample-wise degree at most (ℓ, ℓ)).