Electronic Colloquium on Computational Complexity, Report No. 177 (2016)

Statistical Query Lower Bounds for Robust Estimation of High-dimensional Gaussians and Gaussian Mixtures

Ilias Diakonikolas
University of Southern California
diakonik@usc.edu

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

Alistair Stewart
University of Southern California
alistais@usc.edu

November 10, 2016

Abstract

We prove the first *Statistical Query lower bounds* for two fundamental high-dimensional learning problems involving Gaussian distributions: (1) learning Gaussian mixture models (GMMs), and (2) robust (agnostic) learning of a single unknown mean Gaussian. In particular, we show a *super-polynomial gap* between the (information-theoretic) sample complexity and the complexity of *any* Statistical Query algorithm for these problems. Statistical Query (SQ) algorithms are a class of algorithms that are only allowed to query expectations of functions of the distribution rather than directly access samples. This class of algorithms is quite broad: with the sole exception of Gaussian elimination over finite fields, all known algorithmic approaches in machine learning can be implemented in this model.

Our SQ lower bound for Problem (1) is qualitatively matched by known learning algorithms for GMMs (all of which can be implemented as SQ algorithms). At a conceptual level, this result implies that – as far as SQ algorithms are concerned – the computational complexity of learning GMMs is inherently exponential *in the dimension of the latent space* – even though there is no such information-theoretic barrier. Our lower bound for Problem (2) implies that the accuracy of the robust learning algorithm in [DKK+16b] is essentially best possible among all polynomial-time SQ algorithms. On the positive side, we give a new SQ learning algorithm for this problem with optimal accuracy whose running time nearly matches our lower bound. Both our SQ lower bounds are attained via a unified moment-matching technique that may be useful in other contexts. Our SQ learning algorithm for Problem (2) relies on a filtering technique that removes outliers based on higher-order tensors.

Our lower bound technique also has implications for related inference problems, specifically for the problem of robust *testing* of an unknown mean Gaussian. Here we show an information-theoretic lower bound which separates the sample complexity of the robust testing problem from its non-robust variant. This result is surprising because such a separation does not exist for the corresponding learning problem.

ISSN 1433-8092

1 Introduction

1.1 Background and Overview For the unsupervised estimation problems considered here, the input is a probability distribution which is accessed via a sampling oracle, i.e., an oracle that provides i.i.d. samples from the underlying distribution. Statistical Query (SQ) algorithms are a restricted class of algorithms that are only allowed to query expectations of functions of the distribution rather than directly access samples. This class of algorithms is quite broad: with the sole exception of Gaussian elimination over finite fields, all known algorithmic approaches in machine learning can be implemented in this model. These include spectral techniques, moment methods, local search (e.g., Expectation Maximization), and many others (see, e.g., [FGR$^+$13] for a detailed discussion).

A number of techniques have been developed in information theory and statistics to characterize the sample complexity of inference tasks. These involve both techniques for proving sample complexity upper bounds (e.g., VC dimension, metric entropy) and information-theoretic lower bounds (e.g., Fano and Le Cam methods). On the other hand, computational lower bounds have been much more scarce in the unsupervised setting. Perhaps surprisingly, it is possible to prove *unconditional* lower bounds on the complexity of *any* SQ algorithm that solves a given learning problem. Given the ubiquity and generality of SQ algorithms, an SQ lower bound provides strong evidence of the problem's computational intractability.

In this paper, we prove the first *Statistical Query lower bounds* for two fundamental learning tasks involving high-dimensional Gaussian distributions: (1) learning Gaussian mixture models (GMMs), and (2) robust (agnostic) learning of a single Gaussian. These problems are ubiquitous in applications across the data sciences and have been intensely investigated by different communities of researchers for several decades. Our lower bounds show a *super-polynomial gap* between the (information-theoretic) sample complexity and the complexity of *any* Statistical Query algorithm for these problems, and are essentially tight. Specifically, our lower bound for Problem (1) is qualitatively matched by known learning algorithms for GMMs (all of which can be implemented as SQ algorithms). For Problem (2), we give a new SQ algorithm whose running time nearly matches our lower bound. Both our SQ lower bounds are attained via a unified moment-matching technique that may be useful in other contexts. Our robust SQ learning algorithm relies on a filtering technique that removes outliers based on higher-order tensors.

Our lower bound technique also have implications for related inference problems, specifically for the problem of robustly testing an unknown mean Gaussian. Here we show an information-theoretic lower bound which separates the sample complexity of the robust testing problem from its non-robust variant. This result is surprising because such a separation does not exist for the corresponding learning problem.

Before we discuss our contributions in detail, we provide the necessary background for the Statistical Query model and the unsupervised estimation problems that we study.

Statistical Query Algorithms. A Statistical Query (SQ) algorithm relies on an oracle that given any bounded function on a single domain element provides an estimate of the expectation of the function on a random sample from the input distribution. This computational model was introduced by Kearns [Kea98] in the context of supervised learning as a natural restriction of the PAC model [Val84]. Subsequently, the SQ model has been extensively studied in a plethora of contexts (see, e.g., [Fel16b] and references therein).

A recent line work [FGR$^+$13, FPV15, FGV15, Fel16a] developed a framework for SQ algorithms in the context of search problems over distributions – encompassing the distribution estimation problems we study in this work. It turns out that one can prove unconditional lower bounds on the complexity of SQ algorithms via the notion of *Statistical Query dimension*. This complexity measure was introduced in [BFJ$^+$94] for PAC learning of Boolean functions, and was recently generalized to the unsupervised setting [FGR$^+$13, Fel16a]. A lower bound on the SQ dimension of a learning problem provides an unconditional lower bound on the sample and computational complexity of any SQ algorithm for the problem.

Learning Gaussian Mixture Models. A mixture model is a convex combination of distributions of known type. The most commonly studied case is a Gaussian mixture model (GMM). An *n-dimensional k-GMM* is a distribution in $\mathbb{R}^n$ that is composed of k unknown Gaussian components, i.e., $F = \sum_{i=1}^{k} w_i N(\mu_i, \Sigma_i)$, where the weights w_i, mean vectors μ_i, and covariance matrices Σ_i are unknown. The problem of learning a GMM from samples has received tremendous attention in statistics, and, more recently, in TCS. A long line of work initiated by Dasgupta [Das99, AK01, VW02, AM05, KSV08, BV08] provides computationally efficient algorithms for recovering the parameters of a GMM under separability assumptions. More recently, efficient parameter estimation algorithms have been obtained [MV10, BS10, HP15] under minimal information-theoretic separation assumptions. The related problems of density estimation and proper learning have also been extensively studied [FOS06, SOAJ14, DK14, MV10, HP15]. In density estimation (resp. proper learning), the goal is to output some hypothesis (resp. GMM) that is close to the unknown mixture in total variation distance.

The sample complexity of density estimation (and proper learning) for n-dimensional k-GMMs, up to variation distance ϵ, is easily seen to be $\mathrm{poly}(n, k, 1/\epsilon)$ – without separation assumptions (see Appendix A). For the problem of parameter estimation, the situation is somewhat more subtle: In full generality, the sample complexity of parameter estimation in n dimensions is of the form $\mathrm{poly}(n) \cdot (1/\gamma)^{\Omega(k)}$, where the parameter $\gamma > 0$ quantifies the separation between the components. In fact, even for the 1-dimensional setting, one can construct [MV10, HP15] information-theoretic sample complexity lower bounds of the form $(1/\gamma)^{\Omega(k)}$. On the other hand, such instances are pathological. If the components are nearly non-overlapping, i.e., when the total variation distance between any pair of components is very close to 1, the aforementioned information-theoretic impediment does not apply. Specifically, for nearly non-overlapping instances, the sample complexity of parameter estimation can be shown to be $\mathrm{poly}(n, k)$ as well (see Appendix B).

In summary, the sample complexity of both variants of the learning problem is of the form $\mathrm{poly}(n) f(k)$ for an appropriate function of k. On the other hand, the sample complexity and (hence) running time of all known algorithms for either variant of the learning problem scales as $n^{g(k)}$, for an appropriate function $g(k) \geq k$. This running time upper bound is super-polynomial in the sample complexity of the problem for super-constant values of k, and it is tight for these algorithms even for the case that the underlying GMM has almost non-overlapping components. This raises the following natural question:

Question 1.1. *Is there a $\mathrm{poly}(n, k)$-time density estimation algorithm for n-dimensional k-GMMs? Is there a $\mathrm{poly}(n, k)$-time parameter learning algorithm for nearly non-overlapping n-dimensional k-GMMs?*

Robust Learning of a Gaussian. In the preceding paragraphs, we were working under the assumption that the unknown distribution generating the samples is *exactly* a mixture of Gaussians. The more general setting of *robust* (or agnostic) learning – when our assumption about the model is only *approximately* true – turns out to be significantly more challenging computationally. Specifically, until recently, even the basic setting of robustly learning an unknown mean Gaussian with identity covariance matrix was poorly understood. Without corruptions, this problem is straightforward: The empirical mean gives a sample-optimal and computationally efficient estimator. Unfortunately, the empirical estimate is very brittle and fails in the presence of corruptions.

The standard definition of agnostically learning a Gaussian (see, e.g., Definition 2.1 in [DKK+16b] and references therein) is the following: Instead of drawing samples from a perfect Gaussian, we have access to a distribution D that is promised to be *close* to an unknown Gaussian G, specifically ϵ-close, in total variation distance. This is the only assumption about the distribution D, which may otherwise be arbitrary: the ϵ-fraction of "errors" can be adversarially selected. The goal of an agnostic learning algorithm is to output a hypothesis distribution H that is as close as possible to G (or, equivalently, D) in variation distance. If H is additionally required to be a Gaussian, the agnostic learning algorithm is called proper. Note that the

minimum variation distance, $d_{\mathrm{TV}}(H, G)$, information-theoretically achievable under these assumptions is $O(\epsilon)$, and we would like to obtain a polynomial-time robust algorithm with this performance guarantee.

Agnostically learning a single high-dimensional Gaussian is arguably *the* prototypical problem in robust statistics [Hub64, HRRS86, HR09]. Early work in this field [Tuk75, DG92] studied the sample complexity of the problem. Specifically, for the case of an unknown mean and known covariance Gaussian, the Tukey median [Tuk75] guarantees $O(\epsilon)$-error with $O(n/\epsilon^2)$ samples (see, e.g., [CGR15] for a simple proof). Since $\Omega(n/\epsilon^2)$ samples are information-theoretically necessary – even in the non-robust setting – the robustness requirement does not change the sample complexity of this learning problem.

The computational complexity of agnostically learning a Gaussian is less understood. Until recently, all known polynomial time estimators could only guarantee error of $\Theta(\epsilon\sqrt{n})$. Two recent works [DKK+16b, LRV16] made a first step in designing robust polynomial-time estimators for this problem. The results of [DKK+16b] apply in the standard agnostic model described above; [LRV16] works in a weaker model, where the noisy distribution D is of the form $(1-\epsilon)G + \epsilon N$, where N is an unknown distribution. For the problem of robustly estimating an unknown mean Gaussian $N(\mu, I)$, [LRV16] obtains an error guarantee of $O(\epsilon\sqrt{\log n})$, while [DKK+16b] obtains error $O(\epsilon\sqrt{\log(1/\epsilon)})$, independent of the dimension[1].

A natural and important open problem, put forth by these works [DKK+16b, LRV16], is the following:

Question 1.2. *Is there a* $\mathrm{poly}(n/\epsilon)$*- time agnostic learning algorithm, with error* $O(\epsilon)$*, for an* n*-dimensional Gaussian?*

High-dimensional Hypothesis Testing. So far, we have discussed the problem of learning an unknown distribution that is promised to belong (exactly or approximately) in a given family (Gaussians, mixtures of Gaussians). A related inference problem is that of *hypothesis testing* [NP33, LR05]: Given samples from a distribution in a given family, we want to distinguish between a null hypothesis and an alternative hypothesis. Starting with [GR00, BFR+00], this broad question has been extensively investigated in TCS with a focus on discrete probability distributions. A natural way to solve a distribution testing problem is to learn the distribution in question to good accuracy, and then check if the corresponding hypothesis is close to one satisfying the null hypothesis. This testing-via-learning approach is typically suboptimal and the main goal in this area is to obtain testers with sub-learning sample complexity.

In this paper, we study two natural hypothesis testing analogues of the high-dimensional learning problems discussed in the previous paragraphs. Specifically, we study the sample complexity of (i) *robustly* testing an unknown mean Gaussian, and (ii) testing a GMM.

To motivate (i), we consider arguably the most basic high-dimensional testing task: Given samples from a Gaussian $N(\mu, I)$, where $\mu \in \mathbb{R}^n$ is unknown, distinguish between the case that $\mu = \mathbf{0}$ versus $\|\mu\|_2 \geq \epsilon$. (The latter condition is equivalent, up to constant factors, to $d_{\mathrm{TV}}(N(\mu, I), N(0, I)) \geq \epsilon$.) The classical test for this task is Hotelling's T-squared statistic [Hot31], which is unfortunately not defined when the sample size is smaller than the dimension [ZB96]. More recently, testers that succeed in the sub-linear regime have been developed [SD08] (also see [ZB96, CQ10]). In Appendix C, we give a simple and natural tester for this problem that uses $O(\sqrt{n}/\epsilon^2)$ samples, and show that this sample bound is information-theoretically optimal, up to constant factors.

Now suppose that our Gaussianity assumption about the unknown distribution is only *approximately* satisfied. Formally, we are given samples from a distribution F on $\mathbb{R}^n$ which is promised to be either (a) a zero mean Gaussian $N(0, I)$, or (b) $\epsilon/100$-close in total variation distance to $N(\mu, I)$, for an unknown $\mu \in \mathbb{R}^n$ with $\|\mu\|_2 \geq \epsilon$. The *robust* hypothesis testing problem is to distinguish, with high constant probability,

[1]The algorithm of [LRV16] can be extended to work in the standard agnostic model, at the expense of an increased error guarantee of $O(\epsilon\sqrt{\log n \log(1/\epsilon)})$.

between these two cases. Note that condition (b) implies that $d_{\mathrm{TV}}(D, N(0, I)) = \Omega(\epsilon)$, and therefore the two cases are distinguishable. [2]

Robust hypothesis testing is of fundamental importance and has been extensively studied in robust statistics [HR09, HRRS86, Wil97]. Perhaps surprisingly, it is poorly understood in the most basic settings, even information-theoretically. Specifically, the sample complexity of our aforementioned robust mean testing problem has remained open. It is easy to see that the natural tester of Appendix C fails in the robust setting. On the other hand, the testing-via-learning approach implies a sample upper bound of $O(n/\epsilon^2)$ for our robust testing problem – by using, e.g., the Tukey median. This discussion raises the following natural question:

Question 1.3. *Is there an* information-theoretic *gap between robust testing and non-robust testing? What is the sample complexity of* robustly *testing the mean of a high-dimensional Gaussian?*

We conclude with our hypothesis testing problem regarding GMMs: Given samples from a distribution F on $\mathbb{R}^n$, we want to distinguish between the case that $F = N(0, I)$, or F is a k-mixture $\sum_i w_i N(\mu_i, \Sigma_i)$. This is a natural high-dimensional testing problem that we believe merits investigation in its own right. The natural open question here is where there exists a tester for this problem with sub-learning sample complexity.

1.2 Our Results The main contribution of this paper is a general technique to prove lower bounds for a number of high-dimensional estimation problems involving Gaussian distributions. We use analytic and probabilistic ideas to construct families of hard instances for the high-dimensional estimation problems described in Section 1.1. Using our technique, we prove strong Statistical Query (SQ) lower bounds that answer Questions 1.1 and 1.2 in the negative for the broad class of SQ algorithms[3]. As an additional important application of our technique, we obtain information-theoretic lower bounds on the sample complexity of the corresponding testing problems. (We note that our testing lower bounds apply to *all* algorithms.) Specifically, we answer Question 1.3 in the affirmative, by showing that the robustness requirement makes the Gaussian testing problem information-theoretically harder.

In the body of this section, we state our results and elaborate on their implications and the connections between them. Our first main result is a lower bound of $n^{\Omega(k)}$ on the complexity of any SQ algorithm that learns an arbitrary n-dimensional k-GMM to constant accuracy (see Theorem 4.1 for the formal statement):

Theorem 1.1 (SQ Lower Bound for Learning GMMs)**.** *Any SQ algorithm that learns an arbitrary n-dimensional k-GMM to constant accuracy, for all $n \geq k^8$, requires $2^{n^{\Omega(1)}} \geq n^{\Omega(k)}$ queries to an SQ oracle of precision $n^{-O(k)}$.*

Theorem 1.1 establishes a *super-polynomial gap* between the information-theoretic sample complexity of learning GMMs and the complexity of *any* SQ learning algorithm for this problem. It is worth noting that our hard instance is a family of high-dimensional GMMs whose components are *almost non-overlapping*. Specifically, for each GMM $F = \sum_{i=1}^{k} w_i N(\mu_i, \Sigma_i)$ in the family, the total variation distance between any pair of Gaussian components can be made as large as $1 - 1/\mathrm{poly}(n, k)$. More specifically, for our family of hard instances, the sample complexity of both density and parameter learning is $O(k \cdot \log n)$, while any SQ algorithm takes at least $n^{\Omega(k)}$ time.

At a conceptual level, Theorem 1.1 implies that – as far as SQ algorithms are concerned – the computational complexity of learning high-dimensional GMMs is inherently exponential *in the dimension of the latent space* – even though there is no such information-theoretic barrier in general. Our SQ lower bound

[2] Robust testing should not be confused with tolerant testing, where the completeness is relaxed. In our context, tolerant testing corresponds to distinguishing between $d_{\mathrm{TV}}(D, N(0, I)) \leq \epsilon/2$ versus $d_{\mathrm{TV}}(D, N(0, I)) \geq \epsilon$, where $D = N(\mu, I)$, and is easily seen to be solvable with $O(\sqrt{n}/\epsilon^2)$ samples as well.

[3] We remind the reader that all known algorithmic techniques for learning problems, with the exception of Gaussian elimination over finite fields, are SQ or can be made SQ.

identifies a common barrier of all known algorithmic approaches for this learning problem, and provides a rigorous explanation as to why algorithmic research on this front either relied on strong separation assumptions or resulted in runtimes of the form $n^{\Omega(k)}$.

Our second main result concerns the agnostic learning of a single n-dimensional Gaussian. Our lower bound applies even for the easier problem of agnostically learning a Gaussian with unknown mean vector and identity covariance. Roughly speaking, we show that any SQ algorithm that solves this learning problem to accuracy $O(\epsilon)$ has complexity $n^{\Omega(\log^{1/4}(1/\epsilon))}$. We show (see Theorem 5.1 for a more detailed statement):

Theorem 1.2 (SQ Lower Bound for Robust Learning of Unknown Mean Gaussian)**.** *Let $\epsilon > 0$, $0 < c \leq 1/2$, and $n \geq \mathrm{poly}(\log(1/\epsilon))$. Any SQ algorithm that robustly learns an n-dimensional Gaussian $N(\mu, I)$, within total variation distance $O(\epsilon \log(1/\epsilon)^{1/2-c})$, requires $n^{\Omega(\log(1/\epsilon)^{c/2})}$ queries to an SQ oracle of precision $n^{-\Omega(\log(1/\epsilon)^{c/2})}$.*

Some comments are in order. First, Theorem 1.2 shows a *quasi-polynomial gap* between the sample complexity of agnostically learning an unknown mean Gaussian and the complexity of SQ learning algorithms for this problem. As mentioned in the introduction, $O(n/\epsilon^2)$ samples information-theoretically suffice to agnostically learn the mean to within error $O(\epsilon)$. Second, the robust learning algorithm of [DKK+16b] runs in $\mathrm{poly}(n, 1/\epsilon)$ time, can be implemented in the SQ model, and achieves error $O(\epsilon\sqrt{\log(1/\epsilon)})$ for this problem. As a consequence of Theorem 1.2, we obtain that the $O(\epsilon\sqrt{\log(1/\epsilon)})$ error guarantee of the [DKK+16b] algorithm is in fact best possible among all polynomial-time SQ algorithms.

Roughly speaking, our above theorem shows that any SQ algorithm that solves the (unknown mean Gaussian) robust learning problem to accuracy $O(\epsilon)$ needs to have sample complexity and running time at least $n^{\Omega(\log^{1/4}(1/\epsilon))}$, i.e., *quasi-polynomial* in $1/\epsilon$. An immediate open problem is whether this lower bound can be improved. Perhaps surprisingly, we show that our lower bound is qualitatively tight, by designing an SQ algorithm that uses $O_\epsilon(n^{\sqrt{\log(1/\epsilon)}})$ SQ queries of inverse quasi-polynomial precision. Moreover, we can turn this SQ algorithm into an algorithm in the sampling oracle model with similar sample complexity and running time. Specifically, we show (see Theorem 7.7 and Corollary 7.8):

Theorem 1.3 (SQ Algorithm for Robust Learning of Unknown Mean Gaussian)**.** *Let D be such that $d_{\mathrm{TV}}(D, N(\mu, I)) \leq \epsilon$ for an unknown $\mu \in \mathbb{R}^n$. There is an SQ algorithm that uses $O_\epsilon(n^{O(\sqrt{\log(1/\epsilon)})})$ statistical queries to D of precision $\epsilon/n^{O(\sqrt{\log(1/\epsilon)})}$, and outputs $\widetilde{\mu} \in \mathbb{R}^n$ such that $d_{\mathrm{TV}}(N(\widetilde{\mu}, I), N(\mu, I)) \leq O(\epsilon)$. The SQ algorithm can be turned into an algorithm (in the sample model) with the same error guarantee that has sample complexity and running time $O_\epsilon(n^{O(\sqrt{\log(1/\epsilon)})})$.*

Theorems 1.2 and 1.3 give a qualitatively tight characterization of the complexity of robustly learning an unknown mean Gaussian in the standard agnostic model, where the noisy distribution D is such that $d_{\mathrm{TV}}(D, N(\mu, I)) \leq \epsilon$. Equivalently, D satisfies $(1-\epsilon_1)D + \epsilon_1 N_1 = (1-\epsilon_2)N(\mu, I) + \epsilon_2 N_2$, where N_1, N_2 are unknown (arbitrary) distributions and $\epsilon_1 + \epsilon_2 \leq \epsilon$. A weaker error model, considered in the statistics literature [Hub64, HRRS86, HR09], prescribes that the noisy distribution D is of the form $D = (1-\epsilon)N(\mu, I) + \epsilon N$, where N is an unknown distribution. Intuitively, the difference is that in the former model the adversary is allowed to subtract good samples and add corrupted ones, while in the latter it is only allowed to add corrupted ones.

We note that the lower bound of Theorem 1.2 does not apply for the weaker model of corruptions. This holds for a reason: Concurrent work of the authors with Kamath, Li, and Moitra [DKK+16a] gives a $\mathrm{poly}(n/\epsilon)$ time robust algorithm with $O(\epsilon)$ error in the weak model. Hence, as a consequence, we establish an interesting separation between these two natural models of corruptions. We provide an intuitive justification in Section 1.3.

We now turn to our information-theoretic lower bounds on the sample complexity of the corresponding high-dimensional testing problems. For the robust Gaussian mean testing problem, we show (see Theorem 6.2 for a more detailed statement):

Theorem 1.4 (Sample Complexity Lower Bound for Robust Testing of Unknown Mean Gaussian)**.** *For any constant $c > 0$, there is a constant $\epsilon > 0$ such that the following holds: Any algorithm with sample access to a distribution D on $\mathbb{R}^n$ which satisfies either (a) $D = N(0, I)$ or (b) $d_{\mathrm{TV}}(D, N(\mu, I)) \leq \epsilon/100$, and $\|\mu\|_2 \geq \epsilon$, and distinguishes between the two cases with probability at least $2/3$ requires $\Omega(n^{1-c})$ samples.*

As stated in the Introduction, without the robustness requirement, for any constant $\epsilon > 0$, the Gaussian mean testing problem can be solved with $O_\epsilon(\sqrt{n})$ samples. Hence, the conceptual message of Theorem 1.4 is that robustness makes the Gaussian mean testing problem *information-theoretically* harder. In particular, the sample complexity of robust testing is essentially as hard as that of the corresponding learning problem, which can be solved with $O(n)$ samples. Theorem 1.4 can be viewed as a surprising fact because it implies that *the effect of robustness can be very different for testing versus learning* of the same distribution family. Indeed, recall that the sample complexity of robustly learning an ϵ-corrupted unknown mean Gaussian, up to error $O(\epsilon)$, is $O(n/\epsilon^2)$ – i.e., the same as in the noiseless case.

As a final application of our techniques, we show a sample complexity lower bound for the problem of testing whether a GMM is close to a Gaussian (see Theorem 6.4 for the detailed statement):

Theorem 1.5 (Sample Complexity Lower Bound for Testing a GMM)**.** *Any algorithm with sample access to a distribution D on $\mathbb{R}^n$ which satisfies either (a) $D = N(0, I)$, or (b) $D = \sum_i w_i N(\mu_i, \Sigma_i)$, where the components of D are almost non-overlapping, and distinguishes between the two cases with probability at least $2/3$ requires $\Omega(n^{1-5/(4k-2)})$ samples.*

An interesting open question is whether there exists a tester for the above task that uses fewer samples than the sample-optimal learning algorithm.

1.3 Our Approach and Techniques In this section, we provide a detailed outline of our approach and techniques in tandem with a high-level sketch of our proofs. The structure of this section is as follows: We start by describing our Generic Lower Bound Construction, followed by its two specific instantiations for the problems of Learning GMMs and Robustly Learning an Unknown Mean Gaussian. We continue with our Sample Complexity Testing Lower Bounds, which rely on essentially the same hard instances. We conclude with a detailed sketch of our nearly-matching SQ Algorithm for Robustly Learning an Unknown Mean Gaussian.

Generic Lower Bound Construction. The main idea of our lower bound construction is quite simple: We construct a family of distributions $\mathcal{D}$ that are standard Gaussians in all but one direction, but are somewhat different in the remaining direction (Definition 3.1). Effectively, *we are hiding the interesting information about our distributions in this unknown choice of direction*. By exploiting the simple fact that it is possible to find exponentially many nearly-orthogonal directions (Lemma 3.8), we show that any SQ algorithm with insufficient precision needs many queries in order to learn the distribution in question.

To prove our generic SQ lower bound, we bound from below the SQ-dimension of our hard family of distributions. Roughly speaking, the SQ-dimension of a distribution family (Definition 2.12) corresponds to the number of nearly uncorrelated distributions (with respect to some fixed distribution) in the family (see Definitions 2.10 and 2.11). It is known that a lower bound on the SQ-dimension implies a corresponding lower bound on the number and precision of queries of any SQ algorithm (see Lemma 2.13).

At this point, we describe our lower bound instances more concretely. Given a distribution A on the real line, we produce a family of high-dimensional distributions $\mathbf{P}_v(x)$, for v a unit n-dimensional vector. The

distribution $\mathbf{P}_v$ gives a copy of A in the v-direction, while being an independent standard Gaussian in the orthogonal directions. Our hard family of instances will be the set $\mathcal{D} = \{\mathbf{P}_v \mid v \in \mathbb{S}_n\}$.

For the sake of the intuition, we make two observations about this construction. First, we note that if A and $N(0,1)$ have substantially different moments of degree at most m, for some m, then $\mathbf{P}_v$ and $N(0,I)$ can be easily distinguished by comparing their m^{th} moment tensors. Since these tensors can be approximated in approximately n^m queries (and time), the aforementioned construction will necessarily fail unless the low-order moments of A match the corresponding low-order moments of G. We show that, aside from some mild technical conditions (see Condition 3.2), this moment-matching condition is essentially sufficient for our purposes. If the degree at most m tensors agree, we need to approximate tensors of degree $m+1$. Intuitively, in order to extract useful information from these higher degree tensors, one needs to approximate essentially all of the n^{m+1} many such tensor entries. Second, a natural approach to distinguish between $\mathbf{P}_v$ and $N(0,I)$ would be via random projections. As a critical component of our proof, we show (see Lemma 3.5) that a random projection of $\mathbf{P}_v$ will be exponentially close to $N(0,1)$ with high probability. Therefore, a random projection-based algorithm would require exponentially many random directions until it found a good one.

We now proceed with a somewhat more technical description of our proof. To bound from below the SQ-dimension of our hard family of distributions, we proceed as follows: The definition of the pairwise correlation implies we need to show that $\int \mathbf{P}_v \mathbf{P}_{v'}/G \approx 1$, where $G \sim N(0,I)$ is the Gaussian measure, for any pair of unit vectors v, v' that are nearly orthogonal. By construction of the distributions $\mathbf{P}_v, \mathbf{P}_{v'}$, it follows that in the directions perpendicular to both v and v', the relevant factors integrate to 1. Letting $y = v \cdot \mathbf{x}$ and $z = v' \cdot \mathbf{x}$ and letting y', z' be the orthogonal directions to y and z, we need to consider the integral

$$\int A(y)A(z)G(y')G(z')/G(\mathbf{x}) \ .$$

Fixing y, and integrating over the orthogonal direction, we get

$$\int A(y)/G(y) \int A(z)G(z')dy' \ .$$

Now, if v and v' are (exactly) orthogonal, $z = y'$ and the inner integral equals $G(y)$. When this is not the case, the $A(z)$ term is not quite vertical and the $G(z')$ term not quite horizontal, so instead what we get is only *nearly* Gaussian. In general, the inner integral is equal to

$$U_{v \cdot v'} A(y) \ ,$$

where U_t is the member of the Ornstein–Uhlenbeck semigroup, $U_t f(z) = \mathbf{E}[f(tz + \sqrt{1-t^2}G)]$. We would like to show that this quantity is close to a Gaussian, when $v \cdot v'$ is close to 0 (see Lemma 3.4).

The core idea of the analysis relies on the fact that $U_t A$ is a smeared out version of A. As such, it only keeps the most prominent features of A, namely its low-order moments. In fact, we are able to show that if A and G agree in their first m moments, then $U_t A$ is $O_m(t^m)$-close to a Gaussian (see Lemma 3.5), and thus the integral in question is $O_m((|v \cdot v'|)^m)$-close to 1. This intuition is borne out in a particularly clean way by writing A/G in the basis of Hermite polynomials. The moment-matching condition implies that the decomposition involves none of the Hermite polynomials of degrees 1 through m. However, the Ornstein–Uhlenbeck operator, U_t, is diagonalized by the basis $H_n G$ with eigenvalue t^n. Thus, if $A - G$ can be written in this basis with no terms of degree less than m, applying U_t decreases the size of the function by a multiple of approximately t^m.

Since, we are able to pack $2^{\Omega(n^c)}$ unit vectors v onto the n-sphere so that their pairwise inner products are at most $n^{c-1/2}$ (see Lemma 3.8), we obtain a lower bound on the SQ-dimension of our hard family of distributions. In particular, in order to learn the distribution $\mathbf{P}_v$, for the unknown v, any SQ algorithm

requires either $2^{\Omega(n^c)}$ queries, or queries of accuracy better than $n^{m(c-1/2)}$. This completes the proof sketch of our generic construction.

In our two applications, we can construct one-dimensional distributions A satisfying the necessary moment-conditions for m taken to be super-constant, thus obtaining super-polynomial SQ lower bounds. In the following two paragraphs, we explain how we apply the aforementioned framework to bound the SQ dimension for: (i) learning n-dimensional k-GMMs to constant accuracy, and (ii) robustly learning a single ϵ-corrupted Gaussian $N(\mu, I)$ to accuracy $O(\epsilon)$. In both cases, it suffices to construct a distribution A on the real-line that satisfies the necessary moment-matching conditions such that the high-dimensional family $\mathcal{D} = \{\mathbf{P}_v \mid v \in \mathbb{S}_n\}$ is a k-GMM for (i), and an ϵ-corrupted Gaussian for (ii).

SQ Lower Bound for Learning k-GMMs. The properties of our one-dimensional distribution A are summarized in Proposition 4.2. Specifically, we construct a distribution A on the real line that is a k-mixture of one-dimensional "skinny" Gaussians, A_i, that agrees with $N(0,1)$ on the first $m = 2k-1$ moments (condition (i)). For technical reasons, we require that the chi-squared divergence of A to $N(0,1)$ is bounded from above by an appropriate quantity (condition (iv)). The Gaussian components, A_i, have the same variance and appropriately bounded means (condition (ii)). We can also guarantee that the components A_i are almost non-overlapping (condition (iii)). This guarantees that the corresponding high-dimensional distributions $\mathbf{P}_v, \mathbf{P}'_v$ will be at constant total variation distance from each other when the directions v, v' are nearly orthogonal, and moreover their means will be sufficiently separated.

To establish the existence of a distribution A with the above properties, we proceed in two steps: First, we construct (Lemma 4.3) a discrete one-dimensional distribution B supported on k points, lying in an $O(\sqrt{k})$ length interval, that agrees with $N(0,1)$ on the first k moments. The existence of such a distribution B essentially follows from standard tools on Gauss-Hermite quadrature. The distribution A is then obtained (Corollary 4.4) by adding a zero-mean skinny Gaussian to an appropriately rescaled version of B. Additional technical work (Lemmas 4.5 and 4.6) gives the other conditions.

Our family of hard high-dimensional instances will consist of GMMs that look like almost non-overlapping "parallel pancakes" and is reminiscent of the family of instances considered in [BV08]. For the case of $k = 2$ components, consider a 2-GMM where both components have the same covariance that is far from spherical, the vector between the means is parallel to the unit eigenvector with smallest eigenvalue, and the distance between the means is a large multiple of the standard deviation in this direction (but a small multiple of that in the orthogonal direction). This family of instances was considered in [BV08] who gave an efficient spectral algorithm to learn them.

Our lower bound construction can be thought of as k "parallel pancakes" in which the means lie in a one-dimensional subspace, corresponding to the smallest eigenvalue of the identical covariance matrices of the components. All $n-1$ orthogonal directions will have an eigenvalue of 1, which is much larger than the smallest eigenvalue. In other words, for each unit vector v, the k-GMM $\mathbf{P}_v$ will consist of k "skinny" Gaussians whose mean vectors all lie in the direction of v. Moreover, each pair of components will have total variation distance very close to 1 and their mean vectors are separated by $\Omega(1/\sqrt{k})$.

SQ Lower Bounds for Robustly Learning Unknown Mean Gaussian. To design robust algorithms for the agnostic model, there are two types of adversarial noise to handle: *subtractive noise* – corresponding to the good samples removed by the adversary – and *additive noise* – corresponding to the bad points added by the adversary. The approach of [DKK+16b] does not do anything to address subtractive noise, but it is shown that this type of noise can incur "small" error, namely at most $O(\epsilon\sqrt{\log(1/\epsilon)})$. For additive noise, [DKK+16b] uses the maximum eigenvalue–eigenvector pair of the covariance matrix to create a filter.

Achieving error $O(\epsilon)$ in this model is hard for the following reason: the two types of noise can collude so that the first few moments of the corrupted distribution are indistinguishable from those of a Gaussian

whose mean vector has distance $\Omega(\epsilon\sqrt{\log(1/\epsilon)})$ from the true mean.

To formalize this intuition, for our robust SQ learning lower bound, we construct a distribution A on the real line that agrees with $N(0,1)$ on the first $m = \log^{1/4}(1/\epsilon)$ moments and is $\epsilon/100$-close in total variation distance to $G' = N(\epsilon, 1)$ (see Proposition 5.2). We achieve this by taking A to be the Gaussian $N(\epsilon, 1)$ outside its effective support, while in the effective support we add an appropriate degree-m univariate polynomial p satisfying the appropriate moment conditions. By expressing this polynomial as a linear combination of appropriately scaled Legendre polynomial, we can prove that its L_1 and L_∞ norms within the effective support of G' are much smaller than ϵ (see Lemma 5.11). This result is then used to bound from above the distance of A from G', which gives our SQ lower bound.

Sample Complexity Testing Lower Bounds. Our sample complexity lower bounds follow from standard information-theoretic arguments, and rely on the same lower bound constructions and correlation bounds (i.e., bounds on $\int \mathbf{P}_v \mathbf{P}_{v'}/G$) established in our SQ lower bounds. In particular, we consider the problem of distinguishing between the distribution G and the distribution $\mathbf{P}_v$ for a randomly chosen unit vector $v \in \mathbb{S}_n$ using N independent samples. Let $G^{\otimes N}$ denote the distribution on N independent samples from G, and $\mathbf{P}_v^{\otimes N}$ the distribution obtained by picking a random v and then taking N independent samples from $\mathbf{P}_v$. If it is possible to reliably distinguish between these cases, it must be the case that the chi-squared divergence $\chi(\mathbf{P}_v^{\otimes N}, G^{\otimes N})$ is substantially larger than 1. This is $\int_{v,v',x_i} \prod_{i=1}^N \mathbf{P}_v(x_i)\mathbf{P}_{v'}(x_i)/G(x_i) dv dv' dx_i$. Note that after fixing v and v' the above integral separates as a product, giving

$$\int_{v,v'} \left(\int \mathbf{P}_v(x)\mathbf{P}_{v'}(x)/G(x)dx \right)^N dv dv' \,. \tag{1}$$

Note that the inner integral was bounded from above by roughly $(1 + (v \cdot v')^m)$. By standard results, $(v \cdot v') \ll n^{-c}$ except with probability $\exp(-\Omega(n^{1-2c}))$. So if $k \gg 1/c$, the integrand of (1) is very close to 1. The integrand is only large (though it might be $\exp(O(N))$ large) with probability $\exp(-n^{1-2c})$. Thus, so long as $N < n^{1-O(1)/m}$, we can expect to show that the chi-squared divergence is close to 1, and thus that this testing problem is impossible.

Algorithm for Robustly Learning an Unknown Mean Gaussian. We show that our SQ lower bound for robustly learning the mean of a Gaussian has a nearly-matching upper bound. As our main positive result, we give a robust SQ algorithm with $O(\epsilon)$-error. Our algorithmic approach builds upon the filter-approach of [DKK+16b], generalizing it to the more involved setting of higher-order tensors.

As is suggested by our SQ lower bounds, the obstacle to learning the mean robustly, is that there are ϵ-noisy Gaussians that are $\Omega(\epsilon)$ in variation distance from a target Gaussian G, and yet match G in all of their first $O(\log^{1/4}(1/\epsilon))$ moments. For our algorithm to circumvent this difficulty, it will need to approximate all of the t^{th}-moment tensors for $t \leq k = \Omega(\log^{1/4}(1/\epsilon))$. Note that this already requires n^k queries.

The first thing we will need to show is that k moments are enough, for an appropriate parameter k. Because of our lower bound construction, we know that k needs to be at least $\Omega(\log^{1/4}(1/\epsilon))$. We in fact show that $k = O(\log^{1/2}(1/\epsilon))$ suffices. Specifically, we prove a one-dimensional moment-matching lemma (Lemma 7.1) establishing the following: If an ϵ-noisy one-dimensional Gaussian approximately matches a reference Gaussian G in all of its first k moments, where $k = O(\log^{1/2}(1/\epsilon))$ (i.e., quadratically larger than our lower bound), then it must be $O(\epsilon)$-close to G in variation distance. We note that it suffices to prove this in the one-dimensional case, as we can just project onto the line between the means.

We now proceed to describe our algorithm: Using the algorithm from [DKK+16b], we start by learning the true mean to error $O(\epsilon\sqrt{\log(1/\epsilon)})$. By translating, we can assume that the mean is this close to 0. We need to approximate the low-order moments of our target Gaussian G'. This is complicated by the fact that even a small fraction of errors can have a huge impact on the moments of the distribution. However,

any large errors are easily detectable. In particular, if any t^{th} moment tensor differs substantially from that of the standard Gaussian, it will necessarily imply the presence of errors. In particular, it will allow us to construct a polynomial p so that $\mathbf{E}[p(X)] - \mathbf{E}[p(G')]$ (where X is a noisy version of G') is much larger than $\epsilon\|p(G')\|_2$. If this is the case, then many of our errors, x, must have $p(x)$ very far from the mean. By standard concentration inequalities, this will allow us to identify these points as almost certainly being errors. This in turn lets us build a filter to clean up our distribution X, making it closer to G'.

Repeatedly applying filters as necessary, we can reduce to the case where the higher-order moments of X are close to the higher-order moments of G. This will tell us that, in almost all directions, the first k moments of X match the corresponding moments of G. By our moment-matching lemma, this will imply that the mean of G' is close to 0 in these directions. We will then only need to approximate the mean of the projection of G' onto the low-dimensional subspace V in which these moments fail to match. This approximation can be done in a brute force manner (in time exponential in $\dim(V)$, which is still relatively small), completing the description of the algorithm.

1.4 Discussion and Open Problems This work studies learning and testing high-dimensional structured distributions. Distribution learning and testing are two of the most fundamental inference tasks in statistics with a rich history (see, e.g., [NP33, BBBB72, DG85, Sil86, Sco92, DL01, LR05]) that date back to Karl Pearson. The main criteria to evaluate the performance of an estimator are its sample complexity and its computational complexity. Despite intensive investigation by several decades by different communities, the (sample and/or computational) complexity of many learning and testing problems is still not well understood, even for some surprisingly simple high-dimensional settings. In the past few decades, a long line of work within TCS [KMR$^+$94, Das99, FM99, AK01, VW02, CGG02, MR05, BV08, KMV10, MV10, BS10, DDS12a, DDS12b, CDSS13, DDO$^+$13, CDSS14a, CDSS14b, ADLS15, DDS15, DDKT16, DKS15, DKS16a] has focused on designing efficient estimators in a variety of settings. We have already mentioned the most relevant references for the specific questions we consider in Section 1.1.

With respect to computational lower bounds for unsupervised estimation problems, the most relevant references are the works [FGR$^+$13, FPV15, KV16] that show SQ lower bounds for the planted clique and related planted-like problems. It should be noted that, beyond the fact that we also use the concept of SQ dimension, our techniques are entirely different than theirs. Prior work by Feldman, O'Donnell, and Servedio [FOS08] implicitly showed an SQ lower bound of $n^{\Omega(\log k)}$ for the problem of learning k-mixtures of product distributions over $\{0,1\}^n$. This was obtained by a straightforward reduction from the problem of learning k-leaf decision trees over n Boolean variables. Our lower bound construction for learning GMMs is entirely different from [FOS08] that relied on the obvious combinatorial structure of the discrete setting. It should also be noted that our lower bound construction crucially exploits the structure of the covariance matrices of the Gaussian components.

A related line of work establishes statistical-computational tradeoffs for sparse PCA [BR13a, BR13b, MW15, WBS16], and sparse linear regression [ZWJ14] based on various hardness assumptions. An important difference between these works and the SQ lower bounds of this paper is that the aforementioned sparse problems are known to be tractable if we increase the sample size by a small polynomial factor (typically quadratic) beyond the information-theoretic limit. In contrast, both our SQ lower bounds show a *super-polynomial* gap between the information-theoretic limit and the sample complexity of any SQ algorithm.

Finally, we remark that in the supervised setting of PAC learning Boolean functions, a number of hardness results are known based on various complexity assumptions, see, e.g., [KKMS08, KS06, FGKP06, KK14, DNS14, Dan16] for the problems of learning halfspaces and learning intersections thereof.

The main contribution of this paper is a technique that gives essentially tight SQ lower bounds for two fundamental high-dimensional learning problems: learning GMMs and robustly learning a single Gaussian. To the best of our knowledge, these are the first such lower bounds for high-dimensional distribution learning problems in the continuous setting. As a consequence, we provide a rigorous explanation of the observed

(super-polynomial) gap between the sample complexity of these problems and the best known algorithms.

Our work raises a number of interesting open questions. A natural open problem is to extend our lower bound technique to broader families of high-dimensional distributions. More concretely, is there an $2^{\Omega(k)}\mathrm{poly}(n)$ SQ lower bound for learning k-mixtures of n-dimensional *spherical* Gaussians? Note that our n^k lower bound does not apply for the spherical case, as it crucially exploits the structure of the covariance matrices. In fact, faster learning algorithms for the spherical case are known [SOAJ14], albeit with exponential dependence on the number k of components.

1.5 Organization The structure of this paper is as follows: In Section 2, we introduce basic notation, definitions, and a number of useful facts that will be required throughout the paper. Our SQ lower bounds are established in Sections 3–5. Specifically, in Section 3, we give our high-dimensional SQ lower bound construction, assuming the existence of a one-dimensional density satisfying the necessary conditions. In Sections 4 and 5 we construct the appropriate one-dimensional densities for the problems of learning GMMs and robustly learning an unknown Gaussian, respectively. Section 6 gives our testing sample complexity lower bounds. Finally, in Section 7 we present our robust learning and testing algorithms.

2 Definitions and Preliminaries

2.1 Notation and Basic Definitions For $n \in \mathbb{Z}_+$, we denote by $[n]$ the set $\{1, 2, \ldots, n\}$. We will denote by $\mathbb{S}_n$ the euclidean unit sphere in $\mathbb{R}^n$. If v is a vector, we will let $\|v\|_2$ denote its Euclidean norm. If M is a matrix, we will let $\|M\|_2$ denote its spectral norm, and $\|M\|_F$ denote its Frobenius norm.

Our basic object of study is the Gaussian (or Normal) distribution and finite mixtures of Gaussians:

Definition 2.1. The n-dimensional *Gaussian distribution* $N(\mu, \Sigma)$ with mean vector $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$ is the distribution with probability density function

$$f(x) = (2\pi)^{-n/2} \det(\Sigma)^{1/2} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

Definition 2.2. An n-dimensional *k-mixture of Gaussians* (k-GMM) is a distribution on $\mathbb{R}^n$ with probability density function defined by $F(x) = \sum_{j=1}^k w_j N(\mu_j, \Sigma_j)$, where $w_j \geq 0$, for all j, and $\sum_{j=1}^k w_j = 1$.

Throughout the paper, we will make extensive use of the pdf of the standard one-dimensional Gaussian $N(0, 1)$, which we will denote by $G(x)$.

Definition 2.3. The *total variation distance* between two distributions (with probability density functions) $\mathbf{P}, \mathbf{Q} : \mathbb{R}^n \to \mathbb{R}_+$ is defined to be $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \stackrel{\mathrm{def}}{=} (1/2) \cdot \|\mathbf{P} - \mathbf{Q}\|_1 = (1/2) \cdot \int_{x\in\mathbb{R}^n} |\mathbf{P}(x) - \mathbf{Q}(x)|dx$. If X and Y are random variables, their total variation distance $d_{\mathrm{TV}}(X, Y)$ is defined as the total variation distance between their distributions.

The following simple fact bounds from above the total variation distance between two Gaussians with identity covariance in terms of the Euclidean distance between their mean vectors:

Fact 2.4. *For any $\mu_1, \mu_2 \in \mathbb{R}^n$, we have that $d_{\mathrm{TV}}(N(\mu_1, I), N(\mu_2, I)) \leq \frac{1}{2}\|\mu_2 - \mu_1\|_2$.*

2.2 Formal Problem Definitions We record here the formal definitions of the problems that we study. Our first problem of interest is learning a mixture of k arbitrary high-dimensional Gaussians:

Definition 2.5 (Density Estimation/Proper Learning of GMMs)**.** Let $\mathcal{G}_{n,k}$ be the family of n-dimensional k-GMMs. The problem of *density estimation* for $\mathcal{G}_{n,k}$ is the following: Given $\epsilon > 0$ and sample access to an unknown $\mathbf{P} \in \mathcal{G}_{n,k}$, with probability $9/10$, output a hypothesis distribution $\mathbf{H}$ such that $d_{\mathrm{TV}}(\mathbf{H}, \mathbf{P}) \leq \epsilon$. The problem of *proper learning* for $\mathcal{G}_{n,k}$ is the same with the additional requirement that $\mathbf{H} \in \mathcal{G}_{n,k}$.

Our next problem is the question of robust (agnostic) learning of a Gaussian with unknown mean vector and identity covariance matrix:

Definition 2.6 (Robust Learning of an Unknown Mean Gaussian)**.** The problem of robust (agnostic) learning of an unknown mean Gaussian is the following: Given $\epsilon > 0$ and sample access to an unknown distribution D with $d_{\mathrm{TV}}(D, N(\mu, I)) \leq \epsilon$, where $\mu \in \mathbb{R}^n$ is unknown, output a hypothesis distribution $\mathbf{H}$ such that $d_{\mathrm{TV}}(\mathbf{H}, N(\mu, I))$ is as small as possible.

Definition 2.7 (Robust Testing of an Unknown Mean Gaussian)**.** The problem of robust (agnostic) testing of an unknown mean Gaussian is the following: Given $\epsilon > 0$ and sample access to an unknown distribution D with the promise that one of the following two cases is satisfied: (i) $D = N(0, I)$, or (ii) $d_{\mathrm{TV}}(D, N(\mu, I)) \leq \epsilon/100$, where $\|\mu\|_2 \geq \epsilon$, the goal is to correctly distinguish between the two cases with confidence probability $2/3$.

Definition 2.8 (Testing GMMs)**.** The problem of robust (agnostic) testing of a GMM is the following: Given $\epsilon > 0$ and sample access to an unknown distribution D with the promise that one of the following two cases is satisfied: (a) $D = N(0, I)$, or (b) D is a k-GMM $\sum_i w_i N(\mu_i, \Sigma_i)$ in $\mathbb{R}^n$ which satisfies $d_{\mathrm{TV}}(\mathbf{P}_i, \mathbf{P}_j) \geq 1 - \epsilon$, for all $i \neq j$, distinguish between the two cases with probability at least $2/3$.

2.3 Basics on Statistical Query Algorithms over Distributions We begin by recording the necessary definitions of Statistical algorithms for problems over distributions. All the definitions and facts in this section are from [FGR$^+$13]. We start by defining a general search problem over distributions.

Definition 2.9 (Search problems over distributions)**.** Let $\mathcal{D}$ be a set of distributions over $\mathbb{R}^n$, let $\mathcal{F}$ be a set of *solutions* and $\mathcal{Z} : \mathcal{D} \to 2^{\mathcal{F}}$ be a map from a distribution $D \in \mathcal{D}$ to a subset of solutions $\mathcal{Z}(D) \subseteq \mathcal{F}$ that are defined to be valid solutions for D. The *distributional search problem* $\mathcal{Z}$ over $\mathcal{D}$ and $\mathcal{F}$ is to find a valid solution $f \in \mathcal{Z}(D)$ given access to (an oracle or samples from) an unknown $D \in \mathcal{D}$.

For general search problems over a distribution, we define SQ algorithms as algorithms that do not see samples from the distribution but instead have access to an SQ oracle. We consider two types of SQ oracles from the literature.

1. $\mathrm{STAT}(\tau)$: For a tolerance parameter $\tau > 0$ and any bounded function $f : \mathbb{R}^n \to [-1, 1]$, $\mathrm{STAT}(\tau)$ returns a value $v \in [\mathbf{E}_{x\sim D}[f(x)] - \tau, \mathbf{E}_{x\sim D}[f(x)] + \tau]$.
2. $\mathrm{VSTAT}(t)$: For a sample size parameter $t > 0$ and any bounded function $f : \mathbb{R}^n \to [0, 1]$, $\mathrm{VSTAT}(t)$ returns a value $v \in [\mathbf{E}_{x\sim D}[f(x)] - \tau, \mathbf{E}_{x\sim D}[f(x)] + \tau]$, where $\tau = \max\left\{\frac{1}{t}, \sqrt{\frac{\mathbf{Var}_{x\sim D}[f(x)]}{t}}\right\}$, where $\mathbf{Var}_{x\sim D}[f(x)] = \mathbf{E}_{x\sim D}[f(x)]\,(1 - \mathbf{E}_{x\sim D}[f(x)])$.

The first oracle was defined by Kearns [Kea98] and the second was introduced in [FGR$^+$13]. These oracles are known to be polynomially equivalent [FGR$^+$13]. Also note that these oracles can return any value within the given tolerance, and therefore can make adversarial choices.

The main technical tool that allows us to prove unconditional lower bounds on the complexity of SQ algorithms is an appropriate notion of Statistical Query (SQ) dimension. Such a notion was defined in the context of PAC learning of Boolean functions in [BFJ$^+$94], and subsequently generalized to search problems over distributions in [FGR$^+$13]. We will require the simpler definition from Section 3 of that work that relies on pairwise correlations:

Definition 2.10 (Pairwise Correlation)**.** The pairwise correlation of two distributions with probability density functions $D_1, D_2 : \mathbb{R}^n \to \mathbb{R}_+$ with respect to a distribution with density $D : \mathbb{R}^n \to \mathbb{R}_+$, where the support of D contains the supports of D_1 and D_2, is defined as $\chi_D(D_1, D_2) \stackrel{\mathrm{def}}{=} \int_{\mathbb{R}^n} D_1(x) D_2(x)/D(x) dx - 1$.

We remark that when $D_1 = D_2$ in the above definition, the pairwise correlation is identified with the χ^2-divergence between D_1 and D, i.e., $\chi^2(D_1, D) \stackrel{\text{def}}{=} \int_{\mathbb{R}^n} D_1(x)^2/D(x)dx - 1$.

We will also need the following definition:

Definition 2.11. We say that a set of m distributions $\mathcal{D} = \{D_1, \ldots, D_m\}$ over $\mathbb{R}^n$ is (γ, β)-correlated relative to a distribution D over $\mathbb{R}^n$ if

$$|\chi_D(D_i, D_j)| \leq \begin{cases} \gamma \text{ if } i \neq j \\ \beta \text{ if } i = j. \end{cases}$$

We are now ready to define our notion of dimension:

Definition 2.12 (Statistical Query Dimension)**.** For $\beta, \gamma > 0$, a search problem $\mathcal{Z}$ over a set of solutions $\mathcal{F}$, and a class of distributions $\mathcal{D}$ over $\mathbb{R}^n$, let m be the maximum integer such that there exists a reference distribution D over $\mathbb{R}^n$ and a finite set of distributions $\mathcal{D}_D \subseteq \mathcal{D}$ such that for any solution $f \in \mathcal{F}$, $\mathcal{D}_f = \mathcal{D}_D \setminus \mathcal{Z}^{-1}(f)$ is (γ, β)-correlated relative to D and $|\mathcal{D}_f| \geq m$. We define the *statistical (query) dimension* with pairwise correlations (γ, β) of $\mathcal{Z}$ to be m and denote it by $\mathrm{SD}(\mathcal{Z}, \gamma, \beta)$.

Our lower bounds proceed by bounding from below the statistical query dimension of the considered distribution learning problems. The corresponding lower bounds on the complexity of SQ algorithms for these problems are a corollary of the following result from [FGR$^+$13]:

Lemma 2.13 (Corollary 3.12 in [FGR$^+$13])**.** *Let $\mathcal{Z}$ be a search problem over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over $\mathbb{R}^n$. For $\gamma, \beta > 0$, let $m = \mathrm{SD}(\mathcal{Z}, \gamma, \beta)$. For any $\gamma' > 0$, any SQ algorithm requires at least $m \cdot \gamma'/(\beta - \gamma)$ queries to the $\mathrm{STAT}(\sqrt{\gamma + \gamma'})$ or $\mathrm{VSTAT}(1/(3(\gamma + \gamma')))$ oracles to solve $\mathcal{Z}$.*

3 Statistical Query Lower Bounds: From One Dimension to High Dimensions

Both our statistical query lower bounds are shown in two steps: We first construct a one-dimensional density satisfying certain technical conditions, and then construct a high-dimensional distribution which is Gaussian in all but one directions. The second step is the same for both problems that we consider. To formally define it, we have the following construction:

Definition 3.1 (High-Dimensional Hidden Direction Distribution)**.** For a distribution A on $\mathbb{R}$ with probability density function $A(x)$ and a unit vector $v \in \mathbb{R}^n$, consider the distribution over $\mathbb{R}^n$ with probability density function

$$\mathbf{P}_v(x) = A(v \cdot x) \exp\left(-\|x - v^T x v\|_2^2/2\right)/(2\pi)^{(n-1)/2}.$$

That is, $\mathbf{P}_v$ is the product distribution whose orthogonal projection onto the direction of v is A, and onto the hyperplane perpendicular to v is the standard $(n-1)$ dimensional normal distribution.

Condition 3.2. *The distribution A on $\mathbb{R}$ is such that (i) the first m moments of A agree with the first m moments of $N(0,1)$, and (ii) $\chi^2(A, N(0,1))$ is finite.*

Note that (ii) implies that the distribution A has a probability density function (pdf), which we will denote $A(x)$. We will henceforth blur the distinction between a distribution and its pdf.

The main result of this section is the following:

Proposition 3.3. *Given a distribution A that satisfies Condition 3.2, consider the set of distributions $\{\mathbf{P}_v\}_{v \in \mathbb{S}_n}$, for $n \geq \Omega(m^8)$. For a given $\epsilon > 0$, suppose that $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\epsilon$ whenever $|v \cdot v'|$ is smaller than $1/8$. Then, any SQ algorithm which, given access to $\mathbf{P}_v(\mathbf{x})$ for an unknown $v \in \mathbb{S}_n$, outputs a hypothesis $\mathbf{Q}$ with $d_{\mathrm{TV}}(\mathbf{Q}, \mathbf{P}_v) \leq \epsilon$ needs at least $2^{n^{2/15}} \geq n^m$ queries to $\mathrm{STAT}(O(n)^{-m/6}\sqrt{\chi^2(A, N(0,1))})$ or to $\mathrm{VSTAT}(O(n)^{m/3}/\chi^2(A, N(0,1)))$.*

The rest of this section is devoted to the proof of this proposition. In Section 3.1, we prove a correlation bound which is the main technical ingredient for the proof. In Section 3.2, we show a simple packing for unit vectors over the sphere and put the pieces together to complete the proof.

3.1 Main Correlation Lemma

The main technical result of this section is the following:

Lemma 3.4. *If A agrees with the first m moments of $N(0,1)$, then for all $v, v' \in \mathbb{R}^n$, we have that*

$$\chi_{N(0,I)}(\mathbf{P}_v, \mathbf{P}_{v'}) \leq |v \cdot v'|^{m+1} \chi^2(A, N(0,1)) \ .$$

Note that we may assume that $\chi^2(A, N(0,1))$ is finite, otherwise the lemma statement is trivial. Hence, we can henceforth assume that Condition 3.2 is satisfied. In particular the distributions A, $\mathbf{P}_v$ and $\mathbf{P}_{v'}$ all have probability density functions.

We first bound the χ^2-divergence between the one-dimensional projection of $\mathbf{P}_v$ onto v' and $N(0,1)$. As well as being a critical component towards the proof of Lemma 3.4, this fact can be used to show that random projections of $\mathbf{P}_v$ are close to $N(0,1)$ with high probability.

Lemma 3.5. *Let $\mathbf{Q}$ be the distribution of $v' \cdot X$, for $X \sim \mathbf{P}_v$. Then, we have that*

$$\chi^2(\mathbf{Q}, N(0,1)) \leq (v \cdot v')^{2(m+1)} \chi^2(A, N(0,1)) \ .$$

Proof. Let θ be the angle between v and v'. Let x, y be orthogonal coordinates for the plane spanned by v and v', with the x-axis in the v' direction. Note that $\mathbf{P}_v$ is a product of a distribution on this plane and a standard Gaussian perpendicular to it. On this plane, $\mathbf{P}_v$ is a product of A and $N(0,1)$. Thus, we have that

$$\begin{aligned}\mathbf{Q}(x) &= \int_{\mathbf{x}: v' \cdot \mathbf{x} = x} \mathbf{P}_v(\mathbf{x}) d\mathbf{x} \\ &= \int_{y \in \mathbb{R}} A(x\cos\theta + y \sin\theta) G(x \sin\theta - y\cos\theta) dy \ .\end{aligned}$$

Let U_θ be the linear operator that maps $f : \mathbb{R} \to \mathbb{R}$ to

$$\int_{y \in \mathbb{R}} f(x\cos\theta + y\sin\theta) G(x\sin\theta - y\cos\theta) dy \ ,$$

so that $\mathbf{Q} = U_\theta(A)$. We will show that we can expand A as a linear combination of eigenfunctions of U_θ.

Let $He_i(x)$ denote the i-th (probabilists') Hermite polynomial. We note that the functions $He_i(x)G(x)/\sqrt{i!}$, for $i \in \mathbb{N}$, are orthonormal with respect to the inner product $\langle f, g \rangle = \int_{x \in \mathbb{R}} f(x)g(x)/G(x)dx$. Indeed, by using the fact that the Hermite functions, which can be written as $He(x)\sqrt{G(x)}$, are a complete orthonormal family for $L_2(\mathbb{R})$, we get that any function $f : \mathbb{R} \to \mathbb{R}$ such that $\int_{\mathbb{R}} f(x)g(x)/G(x)dx < \infty$ is almost everywhere equal to a linear combination of these $He_i(x)G(x)/i!$. Since $\int_{-\infty}^{\infty} A(x)^2/G(x)dx = 1 + \chi^2(A, N(0,1))$ is finite, we can write

$$A(x) = \sum_{i=0}^{\infty} a_i He_i(x) G(x)/\sqrt{i!} \ .$$

Using the orthogonality of $He_i(x)$ we can extract these coefficients, since

$$\mathbf{E}_{X \sim A}\left[He_i(x)/\sqrt{i!}\right] = \int_{\mathbb{R}} a_i He_i(x)^2 G(x)/i! dx = a_i \ .$$

Since A agrees with the first m moments of the standard Gaussian, for $0 \le i \le m$, we have that

$$\mathbf{E}_{X\sim A}[He_i(x)/\sqrt{i!}] = \mathbf{E}_{X\sim A}[He_i(x)/\sqrt{i!}] = \delta_{i,0} \ .$$

This implies that $a_0 = 1$ and $a_1, \ldots, a_m = 0$. Thus, we have

$$A(x) = G(x) + \sum_{i=m+1}^{\infty} a_i He_i(x)G(x)/\sqrt{i!} \ . \tag{2}$$

The orthonormality of these functions with respect to the inner product $\langle f, g\rangle = \int_{\mathbb{R}} f(x)g(x)/G(x)dx$ also allows us to express the χ^2-divergence in terms of these coefficients:

$$\begin{aligned}
\chi^2(A, N(0,1)) &= \int_{-\infty}^{\infty} (A(x) - G(x))^2/G(x)dx \\
&= \int_{-\infty}^{\infty} \left(\sum_{i=m+1}^{\infty} a_i He_i(x)G(x)/\sqrt{i!} \right)^2 /G(x)dx \\
&= \sum_{i=m+1} a_i^2 \ .
\end{aligned}$$

Now we consider the effect of U_θ on this orthogonal family. From the definition of U_θ, we have

$$U_\theta(He_iG)(x) = \int_{-\infty}^{\infty} He_i(x\cos\theta + y\sin\theta)G(x\sin\theta - y\cos\theta)dy \ .$$

We can expand this 2-variable polynomial as follows:

Claim 3.6.

$$He_i(x\cos\theta + y\sin\theta) = \sum_{j=1}^{i} \binom{i}{j} \cos^j\theta \sin^{i-j}\theta He_j(x)He_{i-j}(y) \ .$$

Proof. The $He_i(x)$ are monic polynomials: the lead term is x^i with coefficient 1. Thus, all the degree-i terms of $He_i(x\cos\theta + y\sin\theta)$ are given by

$$(x\cos\theta + y\sin\theta)^i \sum_{j=1}^{i} \binom{i}{j} \cos^j\theta \sin^{i-j}\theta x^j y^{i-j} \ .$$

It follows that the degree-i terms of the LHS and RHS of the lemma agree. Therefore, we have

$$He_i(x\cos\theta + y\sin\theta) = p(x,y) + \sum_{j=1}^{i} \binom{i}{j} \cos^j\theta \sin^{i-j}\theta He_j(x)He_{i-j}(y) \ ,$$

for some polynomial $p(x,y)$ of degree at most $i-1$. We need to show that $p(x,y)$ is identically zero. To show this we consider $\mathbf{E}[He_i(X\cos\theta + Y\sin\theta)^2]$, for $(X,Y) \sim N(0,I)$. Since the Gaussian is unaltered by rotations, by a change of coordinates we have that:

$$\begin{aligned}
\mathbf{E}[He_i(X\cos\theta + Y\sin\theta)^2] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} He_i(x\cos\theta + y\sin\theta)^2 G(x)G(y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} He_i(x')^2 G(x')G(y')dx'dy' = i! \ .
\end{aligned}$$

However, pairs of distinct $He_j(x)He_{i-j}(y)$ are orthogonal to each other and they are all orthogonal to the lower degree polynomial $p(x, y)$. Thus, we have

$$\begin{aligned}
\mathbf{E}[He_i(X\cos\theta + Y\sin\theta)^2] &= \mathbf{E}[p(X,Y)^2] + \sum_{j=1}^{i}\binom{i}{j}^2 \cos^{2j}\theta\sin^{2(i-j)}\theta\mathbf{E}[He_j(X)^2He_{i-j}(Y)^2] \\
&= \mathbf{E}[p(X,Y)^2] + \sum_{j=1}^{i}\binom{i}{j}^2 \cos^{2j}\theta\sin^{2(i-j)}\theta i!(i-j)! \\
&= \mathbf{E}[p(X,Y)^2] + i!\sum_{j=1}^{i}\binom{i}{j} \cos^{2j}\theta\sin^{2(i-j)}\theta \\
&= \mathbf{E}[p(X,Y)^2] + i!(\cos^2\theta + \sin^2\theta)^i = \mathbf{E}[p(X,Y)^2] + i! \ .
\end{aligned}$$

We must therefore have that $\mathbf{E}[p(X,Y)^2] = 0$. Since the Gaussian has positive pdf everywhere, this implies that $p(X,Y)$ is identically zero. □

We can now show that conveniently $He_i(x)G(x)$ is an eigenfunction of U_θ:

Corollary 3.7. *We have that:*

$$U_\theta(He_iG)(x) = \cos^i(\theta)He_i(x)G(x) \ .$$

Proof.

$$\begin{aligned}
U_\theta(He_iG)(x) &= \int_{-\infty}^{\infty} He_i(x\cos\theta + y\sin\theta)G(x\cos\theta + y\sin\theta)G(x\sin\theta - y\cos\theta)dy \\
&= \int_{-\infty}^{\infty} He_i(x\cos\theta + y\sin\theta)G(x)G(y)dy \\
&= \sum_{j=1}^{i}\binom{i}{j}\cos^j\theta\sin^{i-j}\theta\int_{-\infty}^{\infty} He_j(x)He_{i-j}(y)G(x\sin\theta - y\cos\theta)dy \\
&= \cos^i\theta He_i(x)G(x) \ ,
\end{aligned}$$

since $\int_{-\infty}^{\infty} He_{i-j}(y)G(y)dy = \delta_{ij}$. □

We can now use these eigenfunctions and eigenvalues with equation (2) to get an expression for $U_\theta A$:

$$U_\theta A(x) = G(x) + \sum_{i=m+1}^{\infty} a_i\cos^i\theta He_i(x)G(x)/\sqrt{i!} \ ,$$

which can be used to express its χ^2-divergence:

$$\begin{aligned}
\chi^2(U_\theta A, N(0,1)) &= \sum_{i=m+1}^{\infty} a_i^2\cos^{2i}\theta \\
&\le \cos^{2(m+1)}\theta\sum_{i=m+1}^{\infty} a_i^2 \\
&= \cos^{2(m+1)}\theta\cdot\chi^2(A, N(0,1)) \ .
\end{aligned}$$

Recalling that $\mathbf{Q} = U_\theta A$ and $\cos\theta = v\cdot v'$, this completes the proof of Lemma 3.5. □

Proof of Lemma 3.4. We first show that the correlation between the high-dimensional $\mathbf{P}_v$ and $\mathbf{P}_{v'}$ needed for Lemma 3.4 can be reduced to a one-dimensional correlation.

Just as in the proof of Lemma 3.5, let $\theta = \arccos(v \cdot v')$ and let x, y be coordinates for the plane spanned by v and v' with the x-axis in the v' direction. Each of $\mathbf{P}_v$ and $\mathbf{P}_{v'}$is a product of a distribution on this plane and a standard Gaussian perpendicular to it. On this plane, they are products of A and $N(0,1)$ with different rotations applied. Thus, we have that

$$\begin{aligned}
\chi_{N(0,I)}(\mathbf{P}_v, \mathbf{P}_{v'}) + 1 &= \int_{\mathbb{R}^n} \mathbf{P}_v(\mathbf{x})\mathbf{P}_{v'}(\mathbf{x})/G(\mathbf{x})d\mathbf{x} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A(x)G(y)A(x\cos\theta + y\sin\theta)G(x\sin\theta - y\cos\theta)/G(x)G(y)dxdy \\
&= \int_{-\infty}^{\infty} A(x)/G(x) \cdot \int_{-\infty}^{\infty} A(x\cos\theta + y\sin\theta)G(x\sin\theta - y\cos\theta)dydx \\
&= 1 + \chi_{N(0,I)}(A, U_\theta A)\ .
\end{aligned}$$

Now we can use the Cauchy-Schwarz inequality to bound from above this correlation in terms of the χ^2-divergences of both distributions from $N(0,1)$, one of which we can bound using Lemma 3.5:

$$\begin{aligned}
\chi_{N(0,I)}(A, U_\theta A) &= \int_{-\infty}^{\infty} (A(x) - G(x))(U_\theta A(x) - G(x))/G(x)dx \\
&\le \sqrt{\int_{-\infty}^{\infty} (A(x) - G(x))^2/G(x)dx} \cdot \sqrt{\int_{-\infty}^{\infty} (U_\theta A(x) - G(x))^2/G(x)dx} \\
&= \sqrt{\chi^2(A, N(0,1))\chi^2(U_\theta A, N(0,1))} \\
&\le \cos^{m+1}(\theta) \cdot \chi^2(A, N(0,1)\ .
\end{aligned}$$

The proof of Lemma 3.4 is now complete. □

3.2 Proof of Proposition 3.3 We note that our distribution learning problem can be expressed as a search problem in the sense of [FGR$^+$13]. Consider the following search problem $\mathcal{Z}$: find a distribution f given samples from $\mathbf{P}_v$ for an unknown unit vector v such that $d_{\mathrm{TV}}(\mathbf{P}_v, f) \le \epsilon$. Thus, for us, F is the set of all distributions on $\mathbb{R}^n$ and $\mathcal{D} \subseteq F$ is the set of $\mathbf{P}_v$ for all unit vectors $v \in \mathbb{S}_n$. For a unit vector v, $\mathcal{Z}(\mathbf{P}_v)$ is the set of all distributions f such that $d_{\mathrm{TV}}(\mathbf{P}_v, f) \le \epsilon$. For a distribution f on $\mathbb{R}^d$, $\mathcal{Z}^{-1}(f)$ is the set of $\mathbf{P}_v$ such that $d_{\mathrm{TV}}(f, \mathbf{P}_v) \le \epsilon$.

To prove a lower bound on the statistical dimension, we will take $D = N(0, I)$, $\mathcal{D} = \{\mathbf{P}_v : v \in \mathbb{S}^n\}$ and construct a suitable finite set $\mathcal{D}_D$.

Lemma 3.8. *There is a set S of at least $2^{n^{1/7}}$ unit vectors such that for $v, v' \in S$, $|v \cdot v'| \le O(n^{-1/3})$.*

Proof. If we take $|S|$ unit vectors at random uniformly from the unit sphere in n dimensions, then this happens with positive probability. We use the following lemma:

Lemma 3.9 (Proposition 1 from [CFJ13])**.** *Given any $0 < \epsilon < \pi/2$, let θ be the angle between two random unit vectors uniformly distributed over S_n. Then we have that:*

$$\mathbf{Pr}[|\theta - \pi/2| \ge \epsilon] \le O(\sqrt{n}(\cos\epsilon)^{n-2}).$$

As a corollary, we have:

Corollary 3.10. *Let θ be the angle between two random unit vectors uniformly distributed over $\mathbb{S}_n$. Then we have that:*

$$\mathbf{Pr}[|\cos\theta| \geq \Omega(n^{-\alpha})] \leq \exp\left(-\Omega(n^{1-2\alpha})\right) ,$$

for any $0 \leq \alpha \leq 1/2$.

Proof. We apply Lemma 3.9 with $\epsilon = n^{-\alpha}$. If $n^{-\alpha} = O(1)$, the result is trivial, so we may assume that $\epsilon \leq 1/100$. Then we have:

$$\begin{aligned}\cos\epsilon &\leq 1 - \epsilon^2/2 + \epsilon^2/24 \\ &\leq 1 - \epsilon^2/3 \\ &\leq \exp(\epsilon^2/4) .\end{aligned}$$

Lemma 3.9 now gives that

$$\mathbf{Pr}[|\theta - \pi/2| \geq n^{-c}] \leq O(\sqrt{n}(\exp(-n^{-2\alpha}/4)^{n-2}) \leq \exp(-n^{1-2\alpha}/5) .$$

Note that if $|\theta - \pi/2| \geq n^{-\alpha}]$, $|\cos\theta| \leq |\theta - \pi/2| \leq 1/n^{-\alpha}$. □

By a union bound, the probability that $v, v' \in S$ have $|v, v'| \leq O(n^{-1/3})$ is

$$|S|^2 2^{-\Omega(n^{1/3})} = 2^{n^{2/7} - \Omega(n^{1/3})} < 1 ,$$

for $n = \Omega(1)$. For $n = O(1)$, the result is trivial since then $|v \cdot v'| \leq 1 \leq O(n^{-1/3})$. □

Proof of Proposition 3.3. We take S to be as in Lemma 3.8. Let $\mathcal{D}_D$ be the set of $\mathbf{P}_v$ for $v \in S$. Then, by Lemma 3.4, we have that, for $v, v' \in S$ with $v \neq v'$, it holds

$$\chi_{N(0,I)}(\mathbf{P}_v, \mathbf{P}_{v'}) \leq |v \cdot v'|^m \chi^2(A, N(0,1)) = \Omega(n)^{-m/3} \chi^2(A, N(0,1)) .$$

If $v = v'$, then $\chi_{N(0,I)}(\mathbf{P}_v, \mathbf{P}_v) = \chi^2(\mathbf{P}_v, N(0,I)) = \chi^2(A, N(0,1))$. We thus have that $\mathcal{D}_D$ is

$$(\Omega(n)^{-m/3}\chi^2(A, N(0,1)), \chi^2(A, N(0,1)))$$

correlated with respect to $D = N(0,I)$.

Since $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\epsilon$ for distinct v and v' in S, for any distribution f over $\mathbb{R}^n$, we have that $\mathcal{Z}^{-1}(f) = \{\mathbf{P}_v : v \in S \text{ and } d_{\mathrm{TV}}(\mathbf{P}_v, f) \leq \epsilon\}$ has $|\mathcal{Z}^{-1}(f)| \leq 1$ using the triangle inequality for d_{TV}. We thus have that $\mathcal{D}_D \setminus \mathcal{Z}^{-1}(f)$ is

$$(\Omega(n)^{-m/3}\chi^2(A, N(0,1)), \chi^2(A, N(0,1)))$$

correlated with respect to $D = N(0,I)$, and $|\mathcal{D}_D \setminus \mathcal{Z}^{-1}(f)| \geq |S| - 1 = \Omega(2^{n^{1/7}})$. That is,

$$SD(\mathcal{Z}, \Omega(n)^{-m/3}\chi^2(A, N(0,1)), \chi^2(A, N(0,1))) \geq \Omega(2^{n^{1/7}}) .$$

If we apply Lemma 2.13 with $\gamma = \Omega(n)^{-m/3}\chi^2(A, N(0,1))$ and $\gamma' = \Omega(n^{-m/3})\chi^2(A, N(0,1))$, we obtain that it requires at least $\Omega(2^{n^{1/7}} n^{-m/3})$ calls to the

$$\mathrm{STAT}(O(n)^{-m/6}\sqrt{\chi^2(A, N(0,1))})$$

or

$$\mathrm{VSTAT}(O(n)^{m/3}/\chi^2(A, N(0,1)))$$

oracle to solve $\mathcal{Z}$. Since $n \geq 2(4m\log_2 n/3)^7$, we have that it needs at least $2^{n^{2/15}} \geq n^m$ calls. This completes the proof. □

4 SQ Lower Bound for Learning Gaussian Mixtures

The main result of this section is the following:

Theorem 4.1. *Fix $0 < \epsilon < 1$. Any SQ algorithm that given SQ access to a mixture $\mathbf{P}$ of k Gaussians $N(\mu_1, \Sigma_1), \ldots, N(\mu_k, \Sigma_k)$ in $\mathbb{R}^n$, for $n \geq \Omega(k^8 \log(1/\epsilon)^3)$, which are promised to have $d_{\mathrm{TV}}(\mathbf{P}_i, \mathbf{P}_j) \geq 1 - \epsilon$, for all $i \neq j$, and moreover satisfy*

$$\max\left\{\max_{i,j} \|\mu_i - \mu_j\|_2, \max_i \|\Sigma_i\|_2^{1/2}\right\} \leq \mathrm{poly}(n, k, \log(1/\epsilon)) \left(\min_i 1/\|\Sigma_i^{-1}\|_2^{1/2}\right) ,$$

outputs a distribution $\mathbf{Q}$ with $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \leq 1/2$, needs at least $2^{n^{2/15}} \geq n^k$ calls to $\mathrm{STAT}\left(O(n)^{-k/6}\right)$ or to $\mathrm{VSTAT}\left(O(n)^{k/3}\right)$.

We remark that the technical condition in the above theorem (i.e., that the distances between the means and the largest and smallest eigenvalues of any covariance matrix are bounded) is required so that an SQ algorithm with a bounded number of SQ queries (in the worst case) is possible.

The proof follows from the results of the previous section and the following proposition:

Proposition 4.2. *For any $\epsilon > 0$, there exists a distribution A that is a mixture of k Gaussians A_i, $1 \leq i \leq k$, that satisfies the following conditions:*

(i) A agrees with $N(0, 1)$ on the first $2k - 1$ moments.

(ii) A is a mixture of Gaussians each with variance $\Theta\left(\frac{1}{k^2 \log(k+1/\epsilon)}\right)$ and means of magnitude $O(\sqrt{k})$.

(iii) It holds $d_{\mathrm{TV}}(A_i, A_j) \geq 1 - \epsilon$, for all $i \neq j$.

(iv) We have $\chi^2(A, N(0, 1)) \leq O(\exp(O(k)) \log(1/\epsilon))$.

(v) In the high-dimensional construction, we have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) \geq 1/2$ whenever $|v \cdot v'| \leq 1/2$.

Given the proposition, the proof of the theorem is easy.

Proof of Theorem 4.1. First note that the distribution A given by Proposition 4.2, satisfies Condition 3.2 for $m = 2k - 1$. By Proposition 3.3, any algorithm that is given SQ access to $\mathbf{P}_v$, for an unknown $v \in \mathbb{S}_n$, and outputs a distribution $\mathbf{Q}$ with $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) \leq \epsilon$, needs at least $2^{n^{2/15}} \geq n^k$ calls to

$$\mathrm{STAT}\left(O(n)^{-(2k-1)/6}/(\exp(O(k))\sqrt{\log(1/\epsilon)})\right)$$

or to

$$\mathrm{VSTAT}\left(O(n)^{(2k-1)/3}/(\exp(O(k))\log(1/\epsilon))\right) .$$

For $n = \Omega(1 + \log(1/\epsilon)^3)$, we have $n^{k-1} \geq n^{k/2} \geq \exp(O(k)) \log(1/\epsilon)^3$, and so we need precision $O(n)^{-k/6}$ for STAT or $O(n)^{k/3}$ for VSTAT.

It remains to show that $\mathbf{P}_v$ is a mixture of k Gaussians that satisfies the necessary conditions. Note that $\mathbf{P}_v$, when expressed in an appropriate basis, is a product of the mixture of k univariate Gaussians and the standard $(n-1)$-dimensional normal distribution. The product of two Gaussians is a Gaussian. If A is a mixture of k Gaussians, $N(\mu_i', \delta)$ with weight w_i, then P_v is a mixture of k Gaussians $N(v\mu_i', I - (1-\delta)vv^T)$ with weights w_i. Indeed, we have that

$$d_{\mathrm{TV}}\left(N(v\mu_i', I - (1-\delta)vv^T), N(v\mu_j', I - (1-\delta)vv^T)\right) = d_{\mathrm{TV}}\left(N(\mu_i', \delta), N(\mu_j', \delta)\right) \geq 1 - \epsilon ,$$

by Proposition 4.2 (iii). Also, we have that

$$\frac{\max\left\{\max_{i,j}\|\mu_i-\mu_j\|_2, \max_i \|\Sigma_i\|_2^{1/2}\right\}}{\left(\min_i 1/\|\Sigma_i^{-1}\|_2^{1/2}\right)} = \frac{\max\left\{\max_{i,j}\|\mu_i-\mu_j\|_2, 1\right\}}{\delta} \leq O(\sqrt{k}/\delta) \leq \mathrm{poly}(k\log(1/\epsilon)) \ .$$

□

The proof of the proposition is deferred to the following subsection.

4.1 Proof of Proposition 4.2 We start with the following lemma:

Lemma 4.3. *There is a discrete distribution B on the real line, supported on k points, that agrees with $N(0,1)$ on the first $2k-1$ moments. All points x in the support of B have $|x| = O(\sqrt{k})$.*

Proof. This mostly follows from standard techniques for Gaussian quadrature. Given a (possibly infinite) interval $[a,b]$, a weighting function $\omega(x)$, and an integer $k>0$, we can find x_i and w_i for $1\leq i\leq k$ such that

$$\int_a^b \omega(x)p(x)dx = \sum_{i=1}^k w_i p(x_i) \ ,$$

for all polynomials $p(x)$ of degree at most $2k-1$. The Gauss-Hermite quadrature is a standard implementation of this on the interval $(-\infty,\infty)$ with $\omega(x)=e^{-x^2}$. Here, we take the x_i's to be the roots of the k-th (physicist's) Hermite polynomial $H_k(x)$. Then, we have that $w_i = \frac{2^{k-1}k!\sqrt{\pi}}{k^2 H_{k-1}(x_i)^2}$.

We would like to take $\omega(x) = G(x) := \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, the pdf of $N(0,1)$. To do this, we need to rescale the above w_i and x_i, and use the probabilist's Hermite polynomials $He_k(x) := 2^{-k/2}H_k(x/\sqrt{2})$. We claim that we can take the x_i's to be the roots of $He_k(x)$ and $w_i = \frac{k!}{k^2 He_{k-1}(x_i)^2}$. Indeed, we have

$$\begin{aligned}
\sum_{i=1}^k w_i p(x_i) &= \sum_{i=1}^k \frac{k!}{k^2 He_{k-1}(x_i)^2}p(x_i) \\
&= \sum_{i=1}^k \frac{2^{k-1}k!\sqrt{\pi}}{k^2 H_{k-1}(\sqrt{2}x_i)^2}\cdot\frac{1}{\sqrt{\pi}}\cdot p(\sqrt{2}x_i) \\
&= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} p(\sqrt{2}y)e^{-y^2}dy \\
&= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} p(x)e^{-x^2/2}/\sqrt{2}dx \\
&= \int_{-\infty}^{\infty} p(x)G(x)dx \ ,
\end{aligned}$$

for all polynomials $p(x)$ of degree at most $2k-1$.

Note that all the weights are nonnegative, since from the definition $w_i = \frac{2^{k-1}k!\sqrt{\pi}}{k^2 H_{k-1}(x_i)^2}$. Also note that $\sum_{i=1}^k w_i = \int_{-\infty}^{\infty} 1\cdot G(x) = 1$. We take B to be the probability distribution with probability w_i of being x_i, for each $1\leq i\leq k$. Then we have

$$\mathbf{E}_{X\sim B}[X^j] = \sum_{i=1}^k w_i x_i^j = \int_{-\infty}^{\infty} x^j G(x)dx = \mathbf{E}_{X\sim N(0,1)}[X^j] \ ,$$

for all integers $1 \leq j \leq k$.

It is known (see, e.g., [Sze89]) that all roots of $H_k(x)$ have absolute value $O(\sqrt{k})$, and so all roots of He_k. Hence, all points x in the support of B have $|x| = O(\sqrt{k})$. This completes the proof. □

On the other hand, if we want $\chi^2(A, N(0,1))$ to be finite, we need to have a mixture of Gaussians each with positive variance $\delta > 0$.

Corollary 4.4. *For any $0 < \delta < 1$, there is a distribution A that is a mixture of k Gaussians each with variance δ that agrees with $N(0,1)$ on the first $2k-1$ moments. The means of all the Gaussians have magnitude $O(\sqrt{k})$.*

Proof. By rescaling the distribution B given by Lemma 4.3, we can find a discrete distribution supported on k points with absolute value no bigger than $\sqrt{k}$ that agrees with the first $2k-1$ moments of $N(0, 1-\delta)$. This has weights $w_i/(1-\delta)$ at the points $x_i(1-\delta)$ for $1 \leq i \leq k$. Let X be a random variable with this distribution. Let X' be a random variable distributed as $N(0, 1-\delta)$. Let Y be a random variable distributed as $N(0, \delta)$ that is independent of X and X'. We take A to be the distribution of $X + Y$. Then we have

$$\mathbf{E}[(X+Y)^j] = \sum_{i=0}^{j} \binom{i}{j} \mathbf{E}[X^i]\mathbf{E}[Y^{j-i}] = \sum_{i=0}^{j} \binom{i}{j} \mathbf{E}[X'^i]\mathbf{E}[Y^{j-i}] = \mathbf{E}[(X'+Y)^j] \ ,$$

for all integers $1 \leq j \leq k$. By standard results about Gaussians, $X' + Y$ is distributed as $N(0,1)$. Finally, note that the distribution of $X + Y$ is a mixture of k Gaussians $N(\sqrt{1-\delta}x_i, \delta)$ with weights w_i. □

To set δ, we need to consider the high-dimensional construction:

Lemma 4.5. *For $v, v \in \mathbb{S}_n$ with $|v \cdot v'| \leq 1/2$, we have that $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) \geq 1 - O\left(k\sqrt{\delta}\log(1/\delta)\csc\theta\right)$.*

Proof. We write A_i, for $1 \leq i \leq k$, for the Gaussians $N(\mu_i, \delta)$ that A is a mixture of. Fix $\epsilon > 0$. By a Chernoff bound, A_i is within the interval $[\mu_i - a, \mu_i + a]$, where $a = 2\sqrt{\delta \log(1/\epsilon)}$ with probability at least $1 - \epsilon$.

We again consider the plane spanned by v and v'. Let x, y be the orthogonal coordinates with v in the direction of the x-axis. Similarly, let x',y' be the orthogonal coordinates with v' in the direction of the x'-axis. Let θ be the angle between v and v'.

We have that

$$\begin{aligned}
\int_{\mathbf{x}} \min\{\mathbf{P}_v(\mathbf{x}), \mathbf{P}_{v'}(\mathbf{x})\} d\mathbf{x} &= \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} \min\{A(x)G(y), A(x')G(y')\} dxdy \\
&= \int_{x=-\infty}^{\infty}\int_{x'=-\infty}^{\infty} \min\{A(x)G(y), A(x')G(y')\} \csc\theta dxdx' \\
&\leq k \max_{i,j} \int_{x=-\infty}^{\infty}\int_{x'=-\infty}^{\infty} \min\{A_i(x)G(y), A_j(x')G(y')\} \csc\theta dxdx' \\
&\leq k\epsilon + k \max_{i,j} \int_{x=\mu_i-a}^{\mu_i+a}\int_{x'=\mu_j-a}^{\mu_j+a} \min\{A_i(x)G(y), A_j(x')G(y')\} \csc\theta dxdx' \\
&\leq k\epsilon + k \max_{i,j} a^2 \csc\theta \max_{x,x'\in\mathbb{R}} \min\{A_i(x)G(y), A_j(x')G(y')\} \\
&\leq k\epsilon + ka^2 \csc\theta/(2\pi\sqrt{\delta}) \\
&= k\epsilon + k\sqrt{\delta}\csc\theta\log(1/\epsilon)/\pi \ .
\end{aligned}$$

Taking $\epsilon = \sqrt{\delta}$, we obtain that $\int_x \min\{\mathbf{P}_v(x), \mathbf{P}_{v'}(x)\}dx \leq O(k\sqrt{\delta}\log(1/\delta)\csc\theta)$. On the other hand,

$$\begin{aligned}
d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) &= \frac{1}{2}\int_x |\mathbf{P}_v(x) - \mathbf{P}_{v'}(x)|dx \\
&= \frac{1}{2}\int_x \max\{\mathbf{P}_v(x), \mathbf{P}_{v'}(x)\} - \min\{\mathbf{P}_v(x), \mathbf{P}_{v'}(x)\}dx \\
&= \frac{1}{2}\int_x \mathbf{P}_v(x) + \mathbf{P}_{v'}(x) - 2\min\{\mathbf{P}_v(x), \mathbf{P}_{v'}(x)\}dx \\
&= 1 - \int_x \min\{\mathbf{P}_v(x), \mathbf{P}_{v'}(x)\}dx \\
&\geq 1 - O(k\sqrt{\delta}\log(1/\delta)\csc\theta)\ .
\end{aligned}$$

This completes the proof. □

This gives an upper bound on δ. We don't want δ to be too small, because of the following lemma:

Lemma 4.6. *We have that* $\chi^2(A, N(0,1)) \leq \exp(O(k))/\sqrt{2\delta}$.

Proof. Each component A_i, for $1 \leq i \leq k$, satisfies the following:

$$\begin{aligned}
&1 + \chi^2(A_i, N(0,1)) \\
&= \int_x A_i(x)^2/G(x)dx \\
&= 1/(\sqrt{2\pi}\delta)\int_x \exp(-(x-\mu_i)^2/\delta + x^2/2) \\
&= 1/(\sqrt{2\pi}\delta)\int_x \exp(-x^2(1/\delta - 1/2) + 2\mu_i x/\delta - \mu_i^2/\delta) \\
&= 1/(\sqrt{2\pi}\delta)\int_x \exp(-(x - 2\mu_i/(2-\delta))^2((2-\delta)/2\delta) + 2\mu_i^2/(\delta(2-\delta)) - \mu_i^2/\delta) \\
&= 1/(\sqrt{2\pi}\delta)\int_x \exp(-(x - 2\mu_i/(2-\delta))^2((2-\delta)/2\delta) + 2\mu_i^2/(\delta(2-\delta)) - (2-\delta)\mu_i^2/\delta(2-\delta)) \\
&= \sqrt{2/(2-\delta)\delta}\exp(\mu_i^2/(2-\delta))\int_x 1/\sqrt{2\pi(2\delta/(2-\delta))}\exp(-(x - 2\mu_i/(2-\delta))^2(1/\delta - 1/2)dx \\
&= \sqrt{2/(2-\delta)\delta}\exp(\mu_i^2/(2-\delta))) \\
&\leq \exp(O(k))/\sqrt{2\delta}\ .
\end{aligned}$$

Thus, for the mixture A we have that:

$$\begin{aligned}
1 + \chi^2(A, N(0,1)) &= \sum_i \sum_j w_i w_j/(1-\delta)\int_x A_i(x)A_j(x)/G(x)dx \\
&\leq \sum_i \sum_j w_i w_j/(1-\delta)\sqrt{(1+\chi^2\left(A_i, N(0,1)\right)(1+\chi^2\left(A_j, N(0,1)\right)} \\
&\leq \sum_i \sum_j w_i w_j/(1-\delta)\cdot\exp(O(k))/\sqrt{2\delta} \\
&= \exp(O(k))/\sqrt{2\delta}\cdot\sum_i\sum_j w_i w_j \\
&= \exp(O(k))/\sqrt{2\delta}\cdot 1\ .
\end{aligned}$$

This completes the proof. □

We could take $\delta = \Theta(1/k^2 \log^2(k))$. But we'd like to enforce the condition that the Gaussians are well-separated:

Lemma 4.7. *Given $\epsilon > 0$, if $\delta = O(1/\sqrt{k}\log(1/\epsilon))$, then $d_{\mathrm{TV}}(A_i, A_j) \geq 1 - \epsilon$.*

Proof. It is known (see, e.g., [Sze89]) that the difference between two roots of $H_k(x)$ is $\Omega(1/\sqrt{k})$. Thus, the same is true of $He_k(x)$ and by our construction, we have that $|\mu_i - \mu_j| \geq \Omega((1-\delta)/\sqrt{k})$, for $i \neq j$. By standard concentration bounds, with probability at least $1 - \epsilon/2$, A_i lies in the range $(\mu_i - \sqrt{2\delta\ln(2/\epsilon)}, \mu_i + \sqrt{2\delta\ln(2/\epsilon)})$. Similarly, with probability at least $1 - \epsilon/2$, A_j lies in the range $(\mu_j - \sqrt{2\delta\ln(2/\epsilon)}, \mu_j + \sqrt{2\delta\ln(2/\epsilon)})$. If these intervals are disjoint, we have $d_{\mathrm{TV}}(A_i, A_j) \geq 1 - \epsilon$. This holds when $(1-\delta)/\sqrt{k}) = \Omega(\sqrt{\delta\log(1/\epsilon)})$, which is true when $\delta = O(1/\sqrt{k}\log(1/\epsilon))$. □

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

Proof of Proposition 4.2. We take $\delta = C/(k^2\log^2(k+1/\epsilon))$ for a sufficiently small constant C. That is (ii). For (i), by Lemma 4.4, A agrees with $N(0,1)$ on the first $2k-1$ moments. Lemma 4.7 gives (iii), Lemma 4.6 gives (iv), and Lemma 4.5 gives (v). □

5 SQ Lower Bound for Robust Learning of a Gaussian

In this section, we use the framework of Section 3 to prove the following theorem:

Theorem 5.1. *Let $\epsilon > 0$, $0 < c \leq 1/2$, and $n \geq \Omega(\log(1/\epsilon)^4)$. Any algorithm that, given SQ access to a distribution $\mathbf{P}$ on $\mathbb{R}^n$ which has $d_{\mathrm{TV}}(\mathbf{P}, N(\mu, I)) \leq \epsilon$ for some vector $\mu \in \mathbb{R}^n$ with $\|\mu\|_2 \leq \mathrm{poly}(n/\epsilon)$, and returns a vector $\tilde{\mu}$ with $\|\tilde{\mu} - \mu\|_2 \leq \epsilon\log(1/\epsilon)^{1/2-c}$, requires at least $2^{n^{2/15}} \geq n^{\log(1/\epsilon)^{c/2}}$ calls to $\mathrm{STAT}\left(O(n)^{-\log(1/\epsilon)^{c/2}/6}\right)$ or to $\mathrm{VSTAT}\left(O(n)^{\log(1/\epsilon)^{c/2}/3}\right)$.*

The theorem will follow from the following proposition:

Proposition 5.2. *For any $\delta > 0$ and integer $m > 0$, there is a distribution A on $\mathbb{R}$ satisfying the following conditions:*

(i) *A and $N(0,1)$ agree on the first m moments.*

(ii) *$d_{\mathrm{TV}}(A, N(\delta, 1)) \leq O(\delta m^2/\sqrt{\log(1/\delta)})$.*

(iii) *$\chi^2(A, N(0,1)) = O(\delta)$.*

We defer the proof of the proposition for the following subsection and show how Theorem 5.1 follows from it.

Proof of Theorem 5.1. We apply Propositions 3.3 and 5.2 with $m = \ln(1/\epsilon)^{c/2}$ and $\delta = C\epsilon\ln 1/\epsilon^{1/2-c}$, where $C > 0$ is a sufficiently large constant. By Proposition 5.2, we have that (i) A and $N(0,1)$ agree on the first $\log(1/\epsilon)^{c/2}$ moments, (ii) $d_{\mathrm{TV}}(A, N(\delta,1)) \leq O(\delta m^2/\sqrt{\log(1/\delta)}) = O(\epsilon)$ and (iii) $\chi^2(A, N(0,1)) = O(\delta)$. Note that for any unit vector $v \in \mathbb{S}_n$, it holds that $d_{\mathrm{TV}}(\mathbf{P}_v, N(v\delta, I)) = d_{\mathrm{TV}}(A, N(\delta,1)) \leq O(C\epsilon)$. We thus have that for any unit vectors v and v' with $|v \cdot v'| \leq 1/8$, that

$$\begin{aligned} d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) &\geq d_{\mathrm{TV}}(N(\delta v, I), N(\delta v', I)) - d_{\mathrm{TV}}(\mathbf{P}_v, N(v\delta, I)) - d_{\mathrm{TV}}(\mathbf{P}_{v'}, N(v'\delta, I)) \\ &\geq \Omega(\delta\|v - v'\|_2) - O(C\epsilon) \\ &= \Omega(\delta\sqrt{2 - 2v\cdot v'}) - O(C\epsilon) \\ &= \Omega(\delta) - O(\delta/(\log 1/\delta)^{1/2-c}) = \Omega(\delta)\ . \end{aligned}$$

We want to apply Proposition 3.3 with ϵ taken to be $\Omega(\delta)$. Our condition on n, $n \geq \ln(1/\epsilon)^4$, implies the condition $n \geq \Omega(m^8)$. We conclude that it requires at least $2^{n^{2/15}} \geq n^{\log(1/\delta)^{c/2}}$ calls to STAT $\left(O(n)^{-\log(1/\delta)^{c/2}/6}\sqrt{\delta}\right)$ or to VSTAT $\left(O(n)^{\ln(1/\delta)^{c/2}/3}/\delta\right)$ to produce a hypothesis distribution $\mathbf{Q}$ with

$$d_{\mathrm{TV}}(\mathbf{Q}, \mathbf{P}_v) \leq O(\delta) \leq \epsilon \ln(1/\epsilon)^{1/2-c} ,$$

where we used the assumption that C is sufficiently large. □

5.1 Proof of Proposition 5.2 The proof proceeds as follows: For some $C = \Theta(\sqrt{\log(1/\delta)})$, we will have $A(x) = G(x-\delta)$ outside of $[-C, C]$. On $[-C, C]$, we will have $A(x) = G(x-\delta) + p(x)$, where $p(x)$ is the degree-m polynomial, which is unique after fixing m, C and δ, with $\int_{-C}^{C} p(x)dx = 0$ and

$$\int_{-C}^{C} p(x)x^i dx = \int_{-\infty}^{\infty} (G(x) - G(x-\delta))x^i dx ,$$

for $1 \leq i \leq m$. We need to show that we can find appropriate m, C, and δ such that the L_1-norm of $p(x)$ is at most $O(\delta m^2/\sqrt{\log(1/\delta)})$ and that $A(x)$ is non-negative.

We will express $p(x)$ as a linear combination of (appropriately scaled) Legendre polynomials, a family of orthogonal polynomials on $[-C, C]$. Rather than showing that the first m moments agree directly, we will instead want that the expectations of the first m scaled Legendre polynomials agree. Bounds on the coefficients of the Legendre polynomials in $p(x)$ allow us to obtain bounds on the L_1 and L_∞ norm of $p(x)$ on $[-C, C]$. Choosing m, δ and C appropriately will then show the proposition.

We record here the basic properties of Legendre polynomials that we will need:

Fact 5.3. *[Sze89] The Legendre polynomials, $P_k(x)$, for integers $k \geq 0$ satisfy the following properties:*

(i) $P_k(x)$ *is a polynomial of degree* k. $P_0(x) = 0$ *and* $P_1(x) = x$.

(ii) $\int_{-1}^{1} P_i(x)P_j(x)dx = (2/(2i+1))\delta_{i,j}$ *for all* $i, j \geq 0$.

(iii) $|P_k(x)| \leq 1$ *for all* $|x| \leq 1$.

(iv) $P_k(x) = (-1)^k P_k(-x)$.

(v) $P_k(x) = 1/2^k \sum_{i=0}^{\lfloor k/2 \rfloor} \binom{k}{i}\binom{2k-2i}{k} x^{k-2i}$.

As a simple corollary we obtain the following lemma:

Lemma 5.4. *We have:*

(i) $|P_k(x)| \leq (4|x|)^k$ *for all* $|x| \geq 1$.

(ii) $\int_{-1}^{1} |P_k(x)| \leq O(1/\sqrt{k})$.

First we note that we can express $p(x)$ as a linear combination of scaled Legendre polynomials whose coefficients are given by integrals:

Lemma 5.5. *We can write* $p(x) = \sum_{k=0}^{m} a_k P_k(x/C)$, *where* $a_k = ((2k+1)/2C) \int_{-C}^{C} P_k(x/C)p(x)dx$.

Proof. Since $p(x)$ has degree at most m and the set of $P_k(x/C)$ for $0 \leq k \leq m$ contains a polynomial of each degree from 0 to m, there exists a_k such that $p(x) = \sum_{k=0}^{m} a_k P_k(x/C)$.

It follows from Fact 5.3 (ii) and a change of variables that $\int_{-C}^{C} P_i(x/C)P_j(x/C)dx = (2C/(2i+1))\delta_{i,j}$, for all $i, j \geq 0$. We can use this to extract the a_k's. For $1 \leq k \leq m$, we have

$$\begin{aligned}\int_{-C}^{C} P_k(x/C)p(x)dx &= \int_{-C}^{C} P_k(x/C)\sum_{i=0}^{m} a_i P_i(x/C)dx \\ &= \sum_{i=0}^{m} a_i \int_{-C}^{C} P_k(x/C)P_i(x/C)dx \\ &= (2C/(2i+1))a_k \ .\end{aligned}$$

□

Note that knowing the first m moments fixes these integrals:

Lemma 5.6. $\int_{-C}^{C} P_k(x/C)p(x)dx = \int_{-\infty}^{\infty}(G(x) - G(x-\delta))P_k(x/C)dx$, *for any* $0 \leq k \leq m$.

Since we will apply this with $1/\delta$ exponential in k and C, we will be able to ignore $O(\delta^2)$ terms. We use Taylor's theorem to expand $(G(x) - G(x-\delta))$ to get the δ and $O(\delta^2)$ terms.

Lemma 5.7. $(G(x) - G(x-\delta)) = xG(x)\delta + (\xi(x)^2 - 1)/2 \cdot G(\xi(x))\delta^2$ *for some* $x \leq \xi(x) \leq x + \delta$.

Lemma 5.8. *For* $k \leq 4C$, $\int_{-\infty}^{\infty} P_k(x/C)xG(x)dx \leq O(\sqrt{k}/C)$.

Proof. When k is even, using Fact 5.3 (iv), we have that $P_k(x/C)xG(x) = -(P_k(-x/C) \cdot (-x)G(-x))$, and so the integral is zero. When k is odd, we can rewrite Fact 5.3 (v) in ascending order of terms, by using the change of variables $j = (k+1)/2 - i$, as

$$P_k(x) = 1/2^k \sum_{j=1}^{(k+1)/2} \binom{k}{(k-1)/2+j}\binom{k+2j-1}{2j-1}x^{2j-1}.$$

By standard results about the moments of Gaussians, we have that for all $j \geq 1$, $\int_{-\infty}^{\infty} x^{2j}G(x)dx = (2j+1)!! := \prod_{i=1}^{j}(2i+1)$. Thus, we have

$$\begin{aligned}\int_{-\infty}^{\infty} P_k(x/C)xG(x)dx &= \int_{-\infty}^{\infty} 1/2^k \sum_{j=1}^{(k+1)/2} \binom{k}{(k-1)/2+j}\binom{k+2j-1}{2j-1}x^{2j}G(x)/C^{2j-1}dx \\ &= 1/2^k \sum_{j=1}^{(k+1)/2} \binom{k}{(k-1)/2+j}\binom{k+2j-1}{2j-1}(2j+1)!!/C^{2j-1} \ .\end{aligned}$$

Note that this is non-negative. We can bound it from above as follows:

$$\begin{aligned}
\int_{-\infty}^{\infty} P_k(x/C)xG(x)dx &= 1/2^k \sum_{j=1}^{(k+1)/2} \binom{k}{(k-1)/2+j}\binom{k+2j-1}{2j-1}(2j+1)!!/C^{2j-1} \\
&\leq \sum_{j=1}^{(k+1)/2} 1/\sqrt{k}\binom{k+2j-1}{2j-1}(2j+1)!!/C^{2j-1} \\
&\leq \sum_{j=1}^{(k+1)/2} 1/\sqrt{k}(k+2j-1)^{2j-1}(2j+1)!!/C^{2j-1}(2j-1)! \\
&\leq \sum_{j=1}^{(k+1)/2} 1/\sqrt{k}2(k+2j-1)^{2j-1}/C^{2j-1} \\
&\leq 2/\sqrt{k} \sum_{j=1}^{(k+1)/2} (2k/C)^{2j-1} \leq 8\sqrt{k}/C \ .
\end{aligned}$$

□

Lemma 5.9. *For any integer* $1 \leq k \leq 4C$*,*

$$\int_{-\infty}^{\infty} P_k(x)(\xi(x)^2-1)/2 \cdot G(\xi(x))dx \leq O(1)$$

for any function $\xi(x)$ *with* $x \leq \xi(x) \leq x+\delta$*, for all* $x \in \mathbb{R}$*.*

Proof. We separate this integral into the interval $[-C, C]$ and the tails. We can use Fact 5.3 (iii) to bound the integral on $[-C, C]$, as follows:

$$\begin{aligned}
\left|\int_{-C}^{C} P_k(x)(\xi(x)^2-1)/2 \cdot G(\xi(x))dx\right| &\leq \left|\int_{-C}^{C} (\xi(x)^2-1)/2 \cdot G(\xi(x))dx\right| \\
&\leq |\int_{-C}^{C} (|x|+\delta+1)^2/2 \cdot G(\min\{0, |x|-\delta\})dx| \\
&\leq O(\delta) + |\int_{-C-\delta}^{C+\delta} (|x|+2\delta+1)^2/2 \cdot G(x)dx| \\
&\leq O(\delta + \mathbf{E}_{X\sim G}[1] + \mathbf{E}_{X\sim G}[|X|] + \mathbf{E}_{X\sim G}[X^2]) \\
&= O(1) \ .
\end{aligned}$$

For the tails, we need Lemma 5.4(i). For the right tail, $[C, \infty)$, we have

$$\begin{aligned}
\left|\int_C^\infty P_k(x)(\xi(x)^2-1)/2 \cdot G(\xi(x))dx\right| &\leq \left|\int_C^\infty (4|x|/C)^k(\xi(x)^2-1)/2 \cdot G(\xi(x))dx\right| \\
&\leq \left|\int_C^\infty (4|x|/C)^k(x+\delta)^2 G(x-\delta)dx\right| \\
&\leq \left|\int_C^\infty (4/C)^k|x+2\delta|^{k+2} G(x-\delta)dx\right| \\
&\leq \left|\int_{C-\delta}^\infty (4/C)^k|x|^{k+2} \cdot (1+2\delta/C)^k G(x-\delta)dx\right| \\
&\leq 2\left|\int_{-\infty}^\infty (4/C)^k|x|^{k+2} G(x-\delta)dx\right| \\
&\leq O((4/C)^k(k+3)!!) \\
&\leq O((4\sqrt{k}/C)^k) \leq O(1)\ .
\end{aligned}$$

A similar bound holds for the left tail, which completes the proof. □

Putting everything together, we have:

Corollary 5.10. *We can write $p(x) = \sum_{k=0}^m a_k P_k(x/C)$, where $a_k = O(\delta k^{3/2}/C^2)$, for $0 \leq k \leq m$.*

Thus, we can get bounds on the L_1 and L_∞ norm of $p(x)$ on $[-C, C]$:

Lemma 5.11. *We have that: $\int_{-C}^C |p(x)|dx \leq O(\delta m^2/C)$ and $|p(x)| \leq \delta m^{5/2}/C^2$, for all $x \in [-C, C]$.*

We can now bound from above the desired χ^2-divergence:

Lemma 5.12. $\chi^2(A, N(0,1)) = O\left(\delta^2 + \delta m^{5/2}/C^2 \cdot (C^2\delta + \max_{|x|\leq C} |p(x)|/G(x))\right)$.

Proof. We have the following:

$$\begin{aligned}
\chi^2(A, N(0,1)) &= \int_{-\infty}^\infty A(x)^2/G(x)dx - 1 \\
&= \int_{-\infty}^\infty G(x-\delta)^2/G(x)dx - 1 + \int_{-C}^C 2p(x)G(x-\delta)/G(x)dx + \int_{-C}^C p(x)^2/G(x)dx\ .
\end{aligned}$$

For the first term, we can write:

$$\begin{aligned}
\int_{-\infty}^\infty G(x-\delta)^2/G(x)dx &= 1/\sqrt{2\pi}\int_{-\infty}^\infty \exp(-x^2/2 + 2\delta x - \delta^2)dx \\
&= \int_{-\infty}^\infty G(x-2\delta)\exp(\delta^2)dx \\
&= \exp(\delta^2) \leq 1 + 2\delta^2\ .
\end{aligned}$$

We bound the second term from above as follows:

$$\begin{aligned}
\left|\int_{-C}^C p(x)G(x-\delta)/G(x)dx\right| &= \left|\int_{-C}^C 2p(x)\exp(x\delta - \delta^2/2)dx\right| \\
&= \left|\int_{-C}^C p(x)\cdot(1 + x\delta + O(C^2\delta^2))dx\right| \\
&\leq C|a_0| + O(C\delta|a_1|) + C^2\delta^2\int_{-C}^C |p(x)dx \\
&\leq 0 + O(\delta^2/C) + O(\delta^3 m^{5/2})\ .
\end{aligned}$$

Finally, for the third term we have:

$$\int_{-C}^{C} p(x)^2/G(x)dx \leq \delta m^{5/2}/C^2 \max_{x\in[-C,C]} |p(x)|/G(x) \ .$$

This completes the proof of the lemma. □

To prove the proposition, we need to set C appropriately and check the bounds on m needed for $A(x)$ to satisfy the necessary properties.

Proof of Proposition 5.2. Note that unless δ is sufficiently small and $m^2 \leq O(\sqrt{\log(1/\delta)})$, taking $A = N(0,1)$, instead of using our construction, satisfies the Proposition. We will take $C = \Theta(\sqrt{\log(1/\delta)})$, and so we can assume that $m \leq \sqrt{C}$.

Recall that $A(x)$ is defined to be $G(x-\delta)+p(x)$ on $[-C,C]$ and $G(x-\delta)$ outside of $[-C,C]$. Firstly, $A(x)$ needs to be the pdf of a distribution. Since

$$\int_{-C}^{C} P_k(x/C)p(x)dx = \int_{-\infty}^{\infty} (G(x)-G(x-\delta))P_k(x/C)dx$$

for $k=0$, when $P_k(x)=1$, we have that $\int_{-\infty}^{\infty} A(x)dx = 1$. We also need that $A(x)$ is non-negative, i.e., that on $[-C,C]$, $A(x) = G(x-\delta)-p(x) \geq 0$. Note that

$$G(x)-p(x) \geq G(C+\delta) - \delta m^{5/2}/C^2 \ ,$$

using Lemma 5.11. Since $m^2 \leq C$, we need $G(C+\delta) \geq \delta C^{3/4}$. This holds when $C = \sqrt{\ln(1/\delta)} - \delta$, since then we have $G(C+\delta) = \sqrt{\delta/2\pi} \geq \delta\sqrt{\ln(1/\delta)}^{3/4}$ for sufficiently small δ. Note that this also implies that $A(x) \leq 2G(x-\delta)$ for all x, and $|p(x)| \leq G(x)$ for all $-C \leq x \leq C$.

The second of these and Lemma 5.12 imply (iii). For (i), by construction, we have that the first m moments agree.

A satisfies (ii), since by Lemma 5.11 ,

$$d_{\mathrm{TV}}(A, N(\delta,1)) = \frac{1}{2}\int_{-C}^{C} |p(x)|dx \leq O(\delta m^2/C) = O(\delta m^2/\sqrt{\log(1/\delta)}).$$

The proof is now compete. □

6 Sample Complexity Lower Bounds for High-Dimensional Testing

In this section, we use our framework to prove information-theoretic lower bounds on the sample complexity of our two high-dimensional testing problems: (i) robustly testing the mean of a single Gaussian, and (ii) (non-robustly) testing between a Gaussian k-mixture and a single Gaussian. Both these statements follow from the structural results established in the previous sections using the following proposition:

Proposition 6.1. *Let A be a distribution on $\mathbb{R}$ satisfying Condition 3.2. Then, there is no algorithm that, for any n, given $O(n^{1-\frac{5}{2m}})$ samples from a distribution D over $\mathbb{R}^n$ which is either $N(0,I)$ or $\mathbf{P}_v$, for some unit vector $v \in \mathbb{R}^n$, correctly distinguishes between the two cases with probability at least $2/3$.*

The proof of the proposition uses the structure of the set of $\mathbf{P}_v$'s and standard information-theoretic arguments, and is deferred to the following subsection.

Using Proposition 6.1, we establish the two main results of this section:

Theorem 6.2 (Sample Complexity Lower Bound for Robustly Testing a Gaussian)**.** *There is no algorithm that, for every $\epsilon > 0$ and integer n, given $O(n^{1-\ln(1/\epsilon)^{-1/3}})$ samples from a distribution $\mathbf{P}$ on $\mathbb{R}^n$ which is promised to satisfy either (a) $\mathbf{P} = N(0, I)$ or (b) $d_{\mathrm{TV}}(\mathbf{P}, N(\mu, I)) \leq \epsilon/100$, where $\|\mu\|_2 \geq \epsilon$, can distinguish between the two cases with probability at least $2/3$.*

Theorem 6.2 follows from the following slightly stronger statement:

Lemma 6.3. *Fix $\epsilon > 0$ and $0 < c \leq 1/2$. There is no algorithm that, given $O\left(n^{1-3\log(1/\epsilon)^{-c/2}}\right)$ samples from a distribution $\mathbf{P}$ on $\mathbb{R}^n$ which is promised to satisfy either (a) $\mathbf{P} = N(0, I)$ or (b) $d_{\mathrm{TV}}(\mathbf{P}, N(\mu, I)) \leq \epsilon$, where $\|\mu\|_2 \geq \epsilon \log(1/\epsilon)^{1/2-c}$, can distinguish between the two cases with probability at least $2/3$.*

Proof. We take $m = \ln(1/\epsilon)^{c/2}$ and $\delta = C\epsilon \ln 1/\epsilon^{1/2-c}$, for some sufficiently large constant C, and apply Proposition 5.2. Then, we apply Proposition 6.1 instead of Proposition 3.3. Again, Proposition 5.2 yields that

$$d_{\mathrm{TV}}(A, N(\delta, 1)) \leq O(\delta m^2/\sqrt{\log(1/\delta)}) \leq \epsilon \,,$$

for sufficiently large C.

We then apply Proposition 6.1, which gives that no algorithm that takes $O(n^{1-1/(5\log(1/\epsilon)^{c/2}}/2))$ samples from a distribution $\mathbf{P}$ on $\mathbb{R}^n$, which is either $\mathbf{P}_v$ for a some unit vector v, or $N(0, I)$ can distinguish between the two with probability at least $2/3$. Since $d_{\mathrm{TV}}(\mathbf{P}_v, N(\delta v, I) = d_{\mathrm{TV}}(\mathbf{P}, N(\mu, I)) \leq \epsilon$, this proves the lemma. □

Theorem 6.4 (Sample Complexity Lower Bound for Testing GMMs)**.** *Fix $\epsilon > 0$, $k \in \mathbb{N}$, and $n \geq \Omega(k^8 \log(1/\epsilon)$. There is no algorithm that given $O(n^{1-5/(4k-2)})$ samples from a distribution on $\mathbb{R}^n$ that is promised to be either (a) $N(0, I)$, or (b) a Gaussian k-mixture $\sum_i w_i N(\mu_i, \Sigma_i)$ in $\mathbb{R}^n$ which satisfies $d_{\mathrm{TV}}(\mathbf{P}_i, \mathbf{P}_j) \geq 1 - \epsilon$, for all $i \neq j$, distinguishes between the two cases with probability at least $2/3$.*

Proof. The distribution A given by Proposition 4.2 satisfies the assumptions of Proposition 6.1. This gives that there is no algorithm, such that given $O(n^{1-5/(4k-2)}))$ samples from a distribution $\mathbf{P}$ which is either $N(0, I)$ or $\mathbf{P}_v$ for some unit vector v, distinguishes between the two with probability at least $2/3$.

By the same argument as 4.1, for any v, $\mathbf{P}_v$ is a mixture satisfying the necessary conditions. This completes the proof. □

6.1 Proof of Proposition 6.1 Suppose that, after fixing the dimension n, the algorithm takes at most N samples. We can consider the testing algorithm as a (possibly randomized) function from N-tuples of samples to its output. For a distribution D, let $D^{\otimes N}$ denote the distribution over independent N-tuples drawn from D. We write $f(D^{\otimes N})$ for the Bernoulli distribution that gives the output of the algorithm given a single sample of $D^{\otimes N}$. Let $\mathbf{Q}_N$ be the distribution obtained by choosing v uniformly at random over the unit sphere $\mathbb{S}_n$, and then drawing N samples from $\mathbf{P}_v$. Then, $f(\mathbf{Q}_N)$ should be "NO" with probability at least $2/3$, since the probability that each $f(\mathbf{P}_v^{\otimes N})$ is "NO" is at least $2/3$. On the other hand, $f(N(0, I)^{\otimes N})$ is "YES" with probability at least $2/3$. By the data processing inequality, it follows that

$$d_{\mathrm{TV}}(\mathbf{Q}_N, N(0, I)^{\otimes N}) \geq d_{\mathrm{TV}}(f(\mathbf{Q}_N), f(N(0, I)^{\otimes N})) \geq 1/3 \,.$$

Suppose for the sake of contradiction that $N \leq O(n^{1-5/(2m)})$ and n is sufficiently large compared to

m and $\chi^2(A, N(0,1))$. Then, we claim that $d_{\mathrm{TV}}(\mathbf{Q}_N, N(0,I)^{\otimes N}) < 1/3$. Indeed, we have that:

$$4d_{\mathrm{TV}}(\mathbf{Q}_N, N(0,I)^{\otimes N})^2 + 1 \leq \chi^2(\mathbf{Q}_N, N(0,I)^{\otimes N}) + 1 =$$

$$= \int_{\mathbf{x}^{(1)}\in\mathbb{R}^n} \dots \int_{\mathbf{x}^{(N)}\in\mathbb{R}^n} \mathbf{Q}_N(\mathbf{x}^{(1)}, \dots \mathbf{x}^{(N)})^2 / \prod_{i=1}^N G(\mathbf{x}^{(i)}) d\mathbf{x}^{(N)} \dots d\mathbf{x}^{(1)}$$

$$= \int_{\mathbf{x}^{(1)}\in\mathbb{R}^n} \dots \int_{\mathbf{x}^{(N)}\in\mathbb{R}^n} \int_{v\in\mathbb{S}_n} \int_{v'\in\mathbb{S}_n} \mathbf{P}_v^N(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})) \mathbf{P}_{v'}^N(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})) / \prod_{i=1}^N G(\mathbf{x}^{(i)}) dv' dv d\mathbf{x}^{(N)} \dots d\mathbf{x}^{(1)}$$

$$= \int_{v\in\mathbb{S}_n} \int_{v'\in\mathbb{S}_n} \int_{\mathbf{x}^{(1)}\in\mathbb{R}^n} \dots \int_{\mathbf{x}^{(N)}\in\mathbb{R}^n} \prod_{i=1}^N \mathbf{P}_v(\mathbf{x}^{(i)}) \mathbf{P}_{v'}(\mathbf{x}^{(i)}) / G(\mathbf{x}^{(i)}) d\mathbf{x}^{(N)} \dots d\mathbf{x}^{(1)} dv' dv$$

$$= \int_{v\in\mathbb{S}_n} \int_{v'\in\mathbb{S}_n} (1 + \chi_{N(0,I)}(\mathbf{P}_v, \mathbf{P}_{v'}))^N dv' dv \ .$$

Consider choosing v and v' independently uniformly at random over $\mathbb{S}_n$. By Corollary 3.10, we have that

$$\mathbf{Pr}\left[|v \cdot v'| \geq \Omega(n^{-\alpha})\right] \leq \exp(-\Omega(n^{1-2\alpha})) \ ,$$

for some $0 \leq \alpha \leq 1/2$. By Lemma 3.4, we have

$$|\chi_{N(0,I)}(\mathbf{P}_v, \mathbf{P}_{v'})| \leq |v \cdot v'|^{m+1} \chi^2(A, N(0,1)) \ .$$

For any v, v' we have

$$|\chi_{N(0,I)}(\mathbf{P}_v, \mathbf{P}_{v'})| \leq \chi^2(A, N(0,1)) \ .$$

When $|v \cdot v'| = O(n^{-\alpha})$, we have

$$|\chi_{N(0,I)}(\mathbf{P}_v, \mathbf{P}_{v'})| \leq n^{-\alpha m} \cdot O(1)^m \cdot \chi^2(A, N(0,1)) \chi^2(A, N(0,1)) \ .$$

Thus, we get that

$$\begin{aligned}
&4d_{\mathrm{TV}}(\mathbf{Q}_N, N(0,I)^{\otimes N}))^2 + 1 \\
\leq &\int_{v\in\mathbb{S}_n} \int_{v'\in\mathbb{S}_n} (1 + \chi_{N(0,I)}(\mathbf{P}_v, \mathbf{P}_{v'}))^N dv' dv \\
\leq &\mathbf{Pr}\left[|v \cdot v'| \leq O(n^{-\alpha})\right] \left(1 + n^{-\alpha m} O(1)^m \chi^2(A, N(0,1))\right)^N + \mathbf{Pr}[|v \cdot v'| \geq \Omega(n^{-\alpha})](1 + \chi^2(A, N(0,1)))^N \\
\leq &\left(1 + n^{-\alpha m} O(1)^m \chi^2(A, N(0,1))\right)^N + \exp(-\Omega(n^{1-2\alpha})) \left(1 + \chi^2(A, N(0,1))\right)^N \ ,
\end{aligned}$$

for any $0 < \alpha \leq 1/2$. We now set $\alpha = 6/(5m)$. For $n = \exp(\Omega(m)) \chi^2(A, N(0,1))^5$, it holds that $n^{-6/5} O(1)^m \chi^2(A, N(0,1)) \leq 1/n \leq 1/N$, and therefore we have that

$$\begin{aligned}
\left(1 + n^{-\alpha m} O(1)^m \chi^2(A, N(0,1))\right)^N &\leq 1 + N n^{-6/5} O(1)^m \chi^2(A, N(0,1)) \\
&\leq 1 + n^{-1/5} O(1)^m \chi^2(A, N(0,1)) < 11/9 \ .
\end{aligned}$$

For $n \geq \log(1 + \chi^2(A, N(0,1)))^{10m}$, we have that

$$\begin{aligned}
\exp(-\Omega(n^{1-2\alpha})) \left(1 + \chi^2(A, N(0,1))\right)^N &= \exp\left(-\Omega\left(n^{1-3/5m} - N \log(1 + \chi^2(A, N(0,1)))\right)\right) \\
&= \exp\left(-\Omega\left(n^{1-3/5m}\left(n^{-1/10m} - \log(1 + \chi^2(A, N(0,1)))\right)\right)\right) \\
&\leq \exp(-\Omega(n^{1-5/2m})) < 2/9 \ .
\end{aligned}$$

Thus,

$$4d_{\mathrm{TV}}\left(\mathbf{Q}_N, N(0,I)^{\otimes N}\right)^2 + 1 < 11/9 + 2/9 \ ,$$

and so $d_{\mathrm{TV}}\left(\mathbf{Q}_N, N(0,I)^{\otimes N}\right) < 1/3$.

7 SQ Algorithms for Robustly Learning and Testing a Gaussian

The structure of this section is as follows: In Section 7.1, we prove a moment matching structural result that forms the basis of our algorithms. In Section 7.2, we present our robust testing algorithm, and in Section 7.3 we give our robust learning algorithm.

7.1 One-Dimensional Moment Matching Lemma The main result of this section is the following structural result:

Lemma 7.1. *Let $G \sim N(0,1)$. For $\delta > \epsilon > 0$, define $k = 2\lceil \epsilon\sqrt{\ln(1/\epsilon)}/\delta \rceil$. Let G' be an ϵ-noisy one-dimensional Gaussian with unit variance so that for all $t \leq k$ we have that the t^{th} moments of G and G' agree to within an additive $(t-1)!(\delta/\epsilon)^t\epsilon/t$. Then, we have that $d_{\mathrm{TV}}(G, G') = O(\delta)$.*

Proof. Let $\widetilde{G} = N(\mu, 1)$ be such that $d_{\mathrm{TV}}(G', \widetilde{G}) \leq \epsilon$. We can assume without loss of generality that $\mu > 0$. Looking at just the mean and variance suffices to get $d_{\mathrm{TV}}(G, G') = O(\delta\sqrt{\log(1/\delta)})$ using techniques similar to the proof of the $O(\epsilon\sqrt{\log(1/\epsilon)})$ filter algorithm from [DKK+16b]. This allows us to focus on the case that $\mu \geq 1$. Formally, we have the following claim:

Claim 7.2. *Lemma 7.1 holds when the mean μ of $\widetilde{G}$ is at least 1.*

Proof. We start by noting that we can assume $\epsilon > 0$ is smaller that a sufficiently small universal constant. Assuming otherwise and recalling that $\delta > \epsilon$ gives that $d_{\mathrm{TV}}(G, G') \leq 1 = O(\delta)$, in which case the lemma statement is trivial.

Let μ' be the mean of G'. We can write

$$G' = \widetilde{G} + \epsilon' E - \epsilon' L \,, \tag{3}$$

for distributions E, L with disjoint supports where $\epsilon' = d_{\mathrm{TV}}(G', \widetilde{G}) \leq \epsilon$. Moreover, it holds

$$\widetilde{G} \geq \epsilon' L \,. \tag{4}$$

Since $k \geq 2$ by definition, the lemma assumptions imply that $|\mu'| \leq \delta$ and $\mathbf{E}_{X\sim G'}[X^2] \leq \delta^2/\epsilon$. By (3) we have that $\mu' = \mu + \epsilon'\mathbf{E}_{X\sim E}[X] - \epsilon'\mathbf{E}_{X\sim L}[X]$, and similarly $\mathbf{E}_{X\sim G'}[(X-\mu')^2] = 1 + (\mu-\mu')^2 + \epsilon'\mathbf{E}_{X\sim E}[(X-\mu')^2] - \epsilon'\mathbf{E}_{X\sim L}[(X-\mu')^2]$. Therefore, we get

$$\epsilon'\mathbf{E}_{X\sim E}[(X-\mu')^2] \leq \delta^2/\epsilon - 1 - (\mu'-\mu)^2 + \epsilon'\mathbf{E}_{X\sim L}[(X-\mu)^2]$$

Note that $\mathbf{Pr}_{X\sim L}[|X-\mu'| \geq T] \leq \mathbf{Pr}_{X\sim\widetilde{G}}[|X-\mu'| \geq T]/\epsilon'$. As in Corollary 8.8 of the high dimensional agnostic paper [DKK+16b], we have that

$$\begin{aligned}
\mathbf{E}_{X\sim L}[(X-\mu')^2] &= \int_0^\infty \mathbf{Pr}_{X\sim L}[|X-\mu'| \geq T] 2T dT \\
&\leq \int_0^\infty 2T \min\left\{1, \mathbf{Pr}_{X\sim N(\mu,1)}[|X-\mu'| \geq T]/\epsilon'\right\} dT \\
&\leq \int_0^\infty 2T \min\{1, \exp(-((T-|\mu-\mu'|)^2/2)2T/\epsilon'\} dT \\
&= \int_0^{\sqrt{2\ln(1/\epsilon')}+|\mu-\mu'|} 2T dT + \int_{\sqrt{2\ln 1/\epsilon'}}^\infty \exp(-T^2/2)2(T+|\mu-\mu'|)/d_{\mathrm{TV}}(G', \widetilde{G}) dT \\
&= O(\ln(1/\epsilon') + 1 + |\mu-\mu'|^2) \,.
\end{aligned}$$

By combining the above, we get that

$$\epsilon' \mathbf{E}_{X\sim E}[(X-\mu')^2] \leq \delta^2/\epsilon - 1 - (1-O(\epsilon'))|\mu-\mu'|^2 + O(\epsilon' \ln(1/\epsilon')) \ ,$$

and thus $\mathbf{E}_{X\sim E}[(X-\mu')^2] \leq \delta^2/(\epsilon\epsilon')$.

However, the means of L and E, μ_L and μ_E have $|\mu_L - \mu'|^2 \leq \mathbf{E}_{X\sim L}[(X-\mu')^2]$ and $|\mu_E - \mu'|^2 \leq \mathbf{E}_{X\sim E}[(X-\mu')^2]$, and therefore $\epsilon'|\mu_L - \mu'| \leq O(\epsilon'\sqrt{\ln 1/\epsilon'} + \epsilon'|\mu'-\mu|) \leq O(\epsilon\sqrt{\ln 1/\epsilon} + \epsilon|\mu'-\mu|)$ and $\epsilon'|\mu_E - \mu| \leq O(\epsilon'\sqrt{\delta^2/\epsilon\epsilon'}) = O(\delta)$.

We thus have that $\mu' - \mu = \epsilon'\mu_E - \epsilon'\mu_L$ has $|\mu'-\mu| \leq O(\delta) + O(\epsilon\sqrt{\ln 1/\epsilon}) + O(\epsilon|\mu-\mu'|)$. Since ϵ is sufficiently small, we have $|\mu'-\mu| \leq O(\delta) + 1/2$. Since $|\mu'| \leq \delta$, We thus have that $\mu \leq O(\delta) + 1/2$. Since we assumed that $\mu \geq 1$, we thus have that $\delta = \Omega(\mu - 1/2) = \Omega(1/2)$ and so $\mu = O(\delta)$. We now have that $d_{\mathrm{TV}}(G, G') = O(\mu) = O(\delta)$. □

We will henceforth assume that $\mu \leq 1$ and thus we have that $d_{\mathrm{TV}}(G, \widetilde{G}) = \Theta(\mu)$. Suppose for the sake of contradiction that $d_{\mathrm{TV}}(G, G') = \Omega(\delta)$, for a sufficiently large constant in the big-Ω. Then, $d_{\mathrm{TV}}(G, \widetilde{G}) \geq d_{\mathrm{TV}}(G, \widetilde{G}) - \epsilon \gg \delta$. We may assume that $\mu \geq C^2\delta$ for a sufficiently large constant $C > 0$. Thus, we have that $\epsilon \leq \mu/C^2 \leq 1/C^2$.

The proof will proceed as follows: Let $f(x) = \sin(xC\epsilon/\mu)$. We note that $f(x)$ has a simple expectation under G or $\widetilde{G}$, and we can easily get a lower bound on their difference. We will also use the Taylor series for $f(x)$ and our moment bounds to derive an upper bound on this difference which contradicts this lower bound.

For $x \in \mathbb{R}_+$, we want an expression for $\mathbf{E}_{X\sim N(x,1)}[f(X)]$. By standard facts on the Fourier transform, we have that

$$\mathbf{E}_{X\sim N(x,1)}[\exp(-i\omega X)] = \exp(-i\omega x)\mathbf{E}_{X\sim N(0,1)}[\exp(-i\omega X)] = \exp(-\omega^2/2 - i\omega x) \ ,$$

for any $\omega \in \mathbb{R}$. Since $\sin(xC\epsilon/\mu) = (\exp(-ixC\epsilon/\mu) - \exp(ixC\epsilon/\mu))/2i$, we obtain that

$$\mathbf{E}_{X\sim N(x,1)}[f(X)] = \exp(-(C\epsilon/\mu)^2/2)\sin(xC\epsilon/\mu) \ .$$

Therefore, $\mathbf{E}_{X\sim N(0,1)}[f(X)] = 0$ and $\mathbf{E}_{X\sim N(\mu,1)}[f(X)] \geq \exp(1/C)\sin(C\epsilon) > (C/2)\epsilon$, and thus

$$\mathbf{E}_{X\sim G'}[f(X)] \geq \mathbf{E}_{X\sim\widetilde{G}}[f(X)] - \epsilon > (C/3)\epsilon. \tag{5}$$

Let h be the degree-$(k-1)$ Taylor polynomial of f plus the term $(Cx\epsilon/\mu)^k/k!$. By the Lagrange form of the remainder in Taylor's theorem, we have that $\left|h(x) - (Cx\epsilon/\mu)^k/k! - f(x)\right| \leq f^{(k)}(\xi)x^k/k!$, for some $\xi \in [0, x]$. Since k is even, we have that the k-th derivative of f, $|f^{(k)}(\xi)| = (C\epsilon/\mu)^k|\sin(\xi C\epsilon/\mu)| \leq (C\epsilon/\mu)^k$. Thus, we get

$$f(x) \leq h(x) \leq f(x) + 2(xC\epsilon/\mu)^k/k! \ .$$

Our goal will be to show that $\mathbf{E}_{X\sim G'}[h(X)]$ is substantially larger than $\mathbf{E}_{X\sim N(0,1)}[h(X)]$, which will contradict the assumption about approximately matching moments. We start by considering $\mathbf{E}_{X\sim N(\mu,1)}[h(X)]$ versus $\mathbf{E}_{X\sim N(0,1)}[h(X)]$. We can write

$$\begin{aligned}
&\mathbf{E}_{X\sim N(\mu,1)}[h(X)] - \mathbf{E}_{X\sim N(0,1)}[h(X)] \\
&= \mathbf{E}_{X\sim N(\mu,1)}[f(X)] - \mathbf{E}_{X\sim N(0,1)}[f(X)] + O\left(\mathbf{E}_{X\sim N(\mu,1)}\left[(XC\epsilon/\mu)^k/k!\right] - \mathbf{E}_{X\sim N(0,1)}\left[(XC\epsilon/\mu)^k/k!\right]\right) \\
&\geq (C/3)\epsilon + O\left(\mathbf{E}_{X\sim N(\mu,1)}\left[(XC\epsilon/\mu)^k/k!\right] - \mathbf{E}_{X\sim N(0,1)}\left[(XC\epsilon/\mu)^k/k!\right]\right) \ ,
\end{aligned}$$

where the last inequality follows from (5). To bound this latter term, we make the following claim:

Claim 7.3. *We have that*

$$|G(x-\mu) - G(x)| \leq O(\mu)G(x/\sqrt{2}).$$

Proof. First, we note that $G(x-\mu)/G(x) = \exp(-2x\mu + \mu^2/2)$.

Recalling our assumption that $0 \leq \mu < 1$, for $|x| \leq O(1/\mu)$, we have that $|G(x-\mu) - G(x)| \leq O(|x|\mu + \mu^2)G(x) \leq O((|x|+1)\mu)G(x)$. Then, since $G(x)/G(x/\sqrt{2}) = O(G(x)) \leq O(1/(|x|+1))$, we get $|G(x-\mu) - G(x)| \leq O(\mu)G(x/\sqrt{2})$.

For $|x| \geq 5/\mu$, since $\sqrt{2\ln(1/\mu)} + 2 \leq 2\ln(1/\mu) + 3 \leq 2/\mu + 3 \leq x$, we have that $G(|x|-2) \leq \mu$. Thusm $G(x)/G(x/\sqrt{2}) \leq O(G(x)) \leq O(\mu)$ and

$$\begin{aligned} G(x-\mu)/G(x/\sqrt{2}) &= O(\exp(-x^2/4 + \mu x - \mu^2/2)) \\ &= O(G((x-2\mu)/\sqrt{2})\exp(\mu^2/2)) \\ &\leq O(G(|x|-2)) \leq O(\mu)\ , \end{aligned}$$

and hence $|G(x-\mu) - G(x)| \leq O(\mu)G(x/\sqrt{2})$. □

Using this claim, we note that

$$\int_{-\infty}^{\infty} (xC\epsilon/\mu)^k/k!\,|G(x-\delta) - G(x)|\,dx = O(\epsilon)\mathbf{E}_{X\sim N(0,1/2)}[(C\epsilon/\mu)^k/k!] = O(\epsilon/(C/\sqrt{2})^k) = O(\epsilon).$$

Therefore, we have that

$$\mathbf{E}_{X\sim\widetilde{G}}[h(X)] \geq \mathbf{E}_{X\sim G}[h(X)] + (C/3)\epsilon.$$

Next, we wish to compare $\mathbf{E}_{X\sim G'}[h(X)]$ to $\mathbf{E}_{X\sim\widetilde{G}}[h(X)]$. Using (3), we have the following for $\epsilon' = d_{\mathrm{TV}}(G', \widetilde{G})$:

$$\begin{aligned} \mathbf{E}_{X\sim G'}[h(X)] &= \mathbf{E}_{X\sim\widetilde{G}}[h(X)] + \epsilon'\left(\mathbf{E}_{X\sim E}[h(X)] - \mathbf{E}_{X\sim L}[h(X)]\right) \\ &\geq \mathbf{E}_{X\sim\widetilde{G}}[h(X)] + \epsilon'\left(\mathbf{E}_{X\sim E}[f(X)] - \mathbf{E}_{X\sim L}[f(X)] - \mathbf{E}_{X\sim L}\left[(X\epsilon/\delta)^k/k!\right]\right) \\ &= \mathbf{E}_{X\sim\widetilde{G}}[h(X)] + O(\epsilon) - \epsilon'\mathbf{E}_{X\sim L}\left[(XC\epsilon/\mu)^k/k!\right]\ . \end{aligned}$$

From (4), i.e., $\widetilde{G} \geq \epsilon' L$, it follows that L satisfies the concentration inequality

$$\Pr_{X\sim L}[|X-\mu| > T] \leq 2\exp(-T^2/2)/\epsilon'\ .$$

We now proceed to bound the subtractive term from above:

$$\begin{aligned} &\epsilon'\mathbf{E}_{X\sim L}\left[(XC\epsilon/\mu)^k/k!\right] \\ &\leq \epsilon'\sum_{J=0}^{\infty}\Pr_{X\sim L}\left[|X-\mu| \geq 2J\sqrt{\log(1/\epsilon)}\right](2(J+1)\sqrt{\log(1/\epsilon)} + \mu)^k(C\epsilon/\mu)^k/k! \\ &\leq \epsilon(3\sqrt{\log(1/\epsilon)}C\epsilon/\mu)^k/k! + \sum_{J=1}^{\infty} 2\epsilon^J(2(J+2)\sqrt{\log(1/\epsilon)}C\epsilon/\mu)^k/k! \\ &\leq \epsilon\cdot O\left(\sqrt{\log(1/\epsilon)}C\epsilon/\mu\right)^k/k!\cdot\left(1 + \sum_{J=1}^{\infty}\epsilon^J(J+2)^k\right) \\ &\leq \epsilon\cdot O\left(\sqrt{\log(1/\epsilon)}\epsilon/C\delta\right)^k/k! \leq \epsilon\cdot O\left(\sqrt{\log(1/\epsilon)}\epsilon/C\delta\right)^k/(\sqrt{2\pi k}(k/e)^k) \\ &\leq \epsilon O(\sqrt{\log(1/\epsilon)}\epsilon/C\delta k)^k/\sqrt{k} \leq O(\epsilon/\sqrt{k}C^k) \leq O(\epsilon)\ . \end{aligned}$$

Therefore, we conclude that

$$\mathbf{E}_{X\sim G'}[h(X)] \geq \mathbf{E}_{X\sim N(\mu,1)}[h(X)] + O(\epsilon) \geq \mathbf{E}_{X\sim N(0,1)}[h(X)] + (C/4)\epsilon \ .$$

On the other hand, recalling that h is the degree-$(k-1)$ Taylor expansion of $f(x) = \sin(xC\epsilon/\mu)$, we can write $h(x) = \sum_{i=0}^{k-1} a_i x^i$, with $a_i = O((C\epsilon/\mu)^i/i!) = O((\epsilon/C\delta)^i/i!$. Therefore, the difference $\mathbf{E}_{X\sim G'}[h(X)] - \mathbf{E}_{X\sim N(0,1)}[h(X)]$ is the sum over i of a_i times the difference in the i^{th} moments, which by assumption is at most

$$\epsilon \sum_{i=0}^{k-1} (i-2)!(\delta/C\epsilon)^i O((\epsilon/\delta)^i/i!) = O(\epsilon)\sum_{i=1}^{k-1}(1/i)^2 = O(\epsilon) \ .$$

This contradicts the fact that their difference is at least $(C/4)\epsilon$, and concludes the proof. □

7.2 Robust Testing Algorithm In this subsection, we give a robust testing algorithm, i.e., an algorithm that distinguishes between an ϵ-noisy Gaussian and $N(0, I)$. This algorithm will form the basis for our robust learning algorithm of the following subsection.

Theorem 7.4 (Robust Testing Algorithm)**.** *Let G' be an ϵ-noisy version of an n-dimensional Gaussian with identity covariance. Let δ be at least a sufficiently large constant multiple of ϵ. There exists an SQ algorithm that makes $O(n^k)$ queries to $\mathrm{STAT}(\epsilon \cdot O(n\log(n/\epsilon)^2)^{-k})$ where $k = 2\lceil O(\epsilon\sqrt{\log(1/\epsilon)}/\delta)\rceil$, and distinguishes between the cases that G' is the standard normal distribution $N(0, I)$, and the case that G' is at least δ-far from $N(0, I)$. The algorithm has running time $n^{O(k)}$.*

By simulating the statistical queries with samples, we obtain:

Corollary 7.5. *Given sample access to G', an ϵ-noisy version of an n-dimensional Gaussian with identity covariance and $\epsilon, \delta > 0$ with δ be at least a sufficiently large constant multiple of ϵ, there is an algorithm that with probability $9/10$ distinguishes between the cases that G' is the standard normal distribution $N(0, I)$, and the case that G' is at least δ-far from $N(0, I)$ and requires at most $(n\log(1/\epsilon))^{O(k)}/\epsilon^2$ samples and running time where $k = 2\lceil O(\epsilon\sqrt{\log(1/\epsilon)}/\delta)\rceil$.*

Proof. In the case when G' is δ-far from $N(0, I)$, let $\widetilde{G} = N(\mu, I)$ be such that $d_{\mathrm{TV}}(G', \widetilde{G}) \leq \epsilon$. We need to show that we can distinguish between the cases $G' = N(0, I)$ and G' is an ϵ-noisy version of a Gaussian $\widetilde{G} = N(\mu, I)$ with $d_{\mathrm{TV}}(\widetilde{G}, G) \geq \delta - \epsilon$. We assume from now on that the completeness case is to show that $d_{\mathrm{TV}}(\widetilde{G}, G) \geq \delta$, since replacing δ with $\delta + \epsilon$ does not affect the statement of the theorem.

The algorithm is quite simple. Let C be a sufficiently large universal constant such that the $O(\delta)$ total variation distance bound in Lemma 7.1 is less than $C\delta$. We assume that $\delta > 4C\epsilon$.

Robust Testing Algorithm:

- For each coordinate axis $1 \leq i \leq n$, use STAT to approximate $\Pr_{X\sim G'}[X \leq \epsilon]$ and $\Pr_{X\sim G'}[X \geq \epsilon]$ to within $\epsilon/2$. If any of these approximations are bigger than $1/2 + \epsilon$, then output "NO".
- Let $k = 2\lceil 2C\epsilon\sqrt{\ln(1/\epsilon)}/\delta\rceil$. Let C' be a sufficiently large constant.
- Using access to STAT, find all the mixed moments of $X \sim G'$, conditioned on $\|X\|_2 \leq C'k\sqrt{n\log(n/\epsilon)}$, of order at most k to within error $n^{-k/2}\epsilon/2$.
- If the difference between any moment of order $t \leq k$ that we measured and that of $N(0, I)$ is more than $((t-1)!(\delta/2C\epsilon)^t/t - 1) \cdot n^{-k/2}\epsilon$, then output "NO".

- Otherwise, output "YES".

The idea is to use Lemma 7.1 with the approximations the moments. However, we have the issue that the STAT oracle can only be used to approximate the expectation of a bounded function. Using the condition $\|X\|_2 \leq C'k\sqrt{n\log(n/\epsilon)}$ allows us to avoid this. But we first need to show that conditioning on it does not affect the moments too much and does not move the distribution far in total variational distance.

Note that the first step of the algorithm will reject if the median of $N(\mu, I)$ projected onto any coordinate axis is outside of the interval $[-\epsilon, \epsilon]$. If this occurs, then $\mu \neq 0$. If this step does not reject, then μ projected onto any coordinate axis is $O(\epsilon)$, and so $\|\mu\|_2 \leq O(\epsilon\sqrt{n})$.

The condition $\|\mathbf{x}\|_2 \leq C'k\sqrt{n\log(n/\epsilon)}$ ensures that the any degree less than k monomial in X is at most $(C'nk^2\log(n/\epsilon)))^{k/2}$. Thus, we can approximate this expectation to precision $n^{-k/2}\epsilon$ using $\mathrm{STAT}(\epsilon \cdot O(n\log(n/\epsilon)^2)^{-k})$. However, since $\|\mu\|_2 \leq O(\epsilon\sqrt{n})$, by standard concentration bounds, the probability that $X \sim N(\mu, I)$ does not satisfy this condition is at most $\epsilon \cdot (C'nk^2\log(n/\epsilon)))^{-k}$. Let G'' be G' conditioned on $\|X\|_2 \leq C'k\sqrt{n\log(n/\epsilon)}$. If $G' = N(0, I)$, then we need to show that the moments of G'' and G' are within $\ln^{-k/2}(\epsilon/2)$. To show this, we use the following lemma:

Lemma 7.6. *For $0 \leq \delta \leq \exp(-k)$, the difference between any mixed moment of degree at most k of $N(0, I)$ and $N(0, I)$ conditioned on $\|x\|_2 \leq O(\sqrt{nk\log(1/\delta)})$, is at most δ.*

Proof. Let X be distributed as $N(0, I)$. Let ϵ be $\delta/(k + \ln 1/\delta)^{(k-1)}C$ for a sufficiently large constant C. Thus, $\ln(1/\epsilon) = \ln(1/\delta) + \ln C + (k-1)\ln(k + \ln(1/\delta))$. Let T be $\sqrt{2n\log 1/\epsilon}$. Then, $T = O_C(\sqrt{nk\log(1/\delta)})$ and by standard concentration inequalities, we have that $\Pr[\|X\|_2 \geq T] \leq \epsilon$.

For $\mathbf{a} \in \mathbb{N}^n$ with $\|\mathbf{a}\|_1 \leq k$, consider the monomial of degree at most k, $m_{\mathbf{a}}(\mathbf{x}) = \prod_{i=1}^n x_i^{a_i}$. First we consider its mean and variance. If any a_i is odd, $\mathbf{E}(m_{\mathbf{a}})(X) = \prod_{i=1}^n \mathbf{E}(X_i^{a_i}) = 0$, since the odd moments of $X_i \sim N(0, 1)$ are zero. If all a_i are even, then $\mathbf{E}(m_{\mathbf{a}})(X) = \prod_{i=1}^n \mathbf{E}(X_i^{a_i}) = \prod_i 2^{a_i/2}(a_i/2)! \leq 2^{k/2}(k/2)!$. For the variance, we have $\mathbf{Var}[m_{\mathbf{a}}(X)] \leq \mathbf{E}[m_{\mathbf{a}}(X)^2] = \prod_{i=1}^n \mathbf{E}(X_i^{2a_i}) = \prod_i 2_i^a a_i! \leq 2^k k!$. Let $p_{\mathbf{a}}(X) = m_{\mathbf{a}}(X)/2^{k/2}\sqrt{k!}$. Then we have that $0 \leq \mathbf{E}[p_{\mathbf{a}}(X)] \leq 1$ and $\mathbf{Var}[p_{\mathbf{a}}(X)] \leq 1$.

By the standard concentration inequality given in Lemma 7.16 below, we have that for all $t > 0$, $\Pr[|p_{\mathbf{a}}(X)| \geq t + 1] \leq \exp(2 - (t/R)^{2/k})$ for some $R > 0$. Thus, we have $\Pr[|p_{\mathbf{a}}(X)| \geq c + 1] \leq \epsilon$, for $c = R(\ln(1/\epsilon) - 2)^{k/2}$. Let $I(\mathbf{x})$ be the indicator function of $\|\mathbf{x}\|_2 \geq T$. Then we have that

$$
\begin{aligned}
|\mathbf{E}[I(X)p_{\mathbf{a}}(X)]| &\leq \mathbf{E}[I(X)|p_{\mathbf{a}}(X)|] \\
&= \int_0^\infty \Pr[I(X)|p_{\mathbf{a}}(X)| \geq t]dt \\
&\leq \int_0^{c+1} \epsilon dt + \int_{c+1}^\infty \exp(2 - (t - 1/R)^{2/k})dt \\
&= (c+1)\epsilon + \int_c^\infty \exp(2 - (t/R)^{2/k})dt \\
&= (c+1)\epsilon + \int_{\ln(1/\epsilon)}^\infty \exp(2 - x)(dt/dx)dx \qquad \text{(where } x = (t/R)^{2/k}\text{)} \\
&= (c+1)\epsilon + (Rk/2)\int_{\ln(1/\epsilon)}^\infty \exp(2 - x)x^{k/2-1}dx \\
&= (c+1)\epsilon + (Rk/2)\epsilon \cdot \sum_{j=0}^{k/2-1} ((k/2)!/(k/2 - j)!)\ln(1/\epsilon)^{k/2-j} \\
&\leq R\ln(1/\epsilon)^{k/2}\epsilon + (k^2/8)\epsilon(k + \log(1/\epsilon))^{k/2-1} \\
&\leq O(k^2\epsilon(k + \log(1/\epsilon))^{k/2-1}) \ ,
\end{aligned}
$$

where the integral $\int_{\ln(1/\epsilon)}^{\infty} \exp(2-x)x^{k/2-1}dx$ is calculated explicitly below in Claim 7.18. In terms of $m_a(\mathbf{x})$, we have

$$|\mathbf{E}[I(X)m_{\mathbf{a}}(X)]| \leq O(k^2\epsilon(k+\log(1/\epsilon))^{k/2-1}2^{k/2}\sqrt{k!}) \leq O(\epsilon(k+\log(1/\epsilon)^{(k-1)}) \leq O(\delta/C) \leq \delta/2 \; .$$

Then, for X' distributed as $N(0,I)$ conditioned on $\|X'\|_2 \leq T$, we have

$$\begin{aligned}
|\mathbf{E}[m_{\mathbf{a}}(X)] - \mathbf{E}[m_{\mathbf{a}}(X')]| &= |\mathbf{E}[m_{\mathbf{a}}(X)] - \mathbf{E}[m_{\mathbf{a}}(X)(1-I(X))]/(1-\Pr[\|X\|_2 > T])| \\
&= |(\mathbf{E}[I(X)m_{\mathbf{a}}(X)] - \mathbf{E}[m_{\mathbf{a}}(X)]\Pr[\|X\|_2 > T])\,/1-\Pr[\|X\|_2 > T])| \\
&\leq \left(\delta/2 + 2^{k/2}\sqrt{k}\epsilon\right)/(1-\epsilon) \\
&\leq 2\delta/3(1-\epsilon) \leq \delta \; .
\end{aligned}$$

□

Applying Lemma 7.6 for $\delta = n^{-k/2}\epsilon$, noting that $C'k\sqrt{n\log(n/\epsilon)} = \Omega(C'\sqrt{nk\log(1/\delta)})$ yields that the moments of G'' and $G' = N(0,I)$ are within $\ln^{-k/2}(\epsilon/2)$. Thus, in this case, the approximations of the moments of G'' are within $n^{-k/2}\epsilon$ of the moments of G'.

For the soundness case, we just note that since $(t-1)!(\delta/2C\epsilon)^t\epsilon/t - 1 \geq (\delta/2C\epsilon)/2 - 1 \geq 1$, the bounds on the moments we need to fail are bigger than the precision of the statistical queries we use to approximate them, and therefore we never output "NO" when $G' = N(0,I)$.

Now suppose that G' is an ϵ-noisy version of an identity covariance Gaussian $\widetilde{G}$. Then G'' is a 2ϵ-noisy version of $\widetilde{G}$. We will denote μ the mean vector of $\widetilde{G}$ and will assume that $\|\mu\|_2 \geq \delta$. We need to show that the algorithm outputs "NO".

Consider the unit vector $v = \mu/\|\mu\|_2$ which has $v \cdot \mu \geq \delta$. Consider the one-dimensional distributions G''_v and $\widetilde{G}_v$ of the form $v \cdot X$, where either $X \sim G''$ or $X \sim \widetilde{G}$. Note that $\widetilde{G}_v = N(\|\mu\|_2, 1)$ has mean larger than δ and that $d_{\mathrm{TV}}(G''_v, \widetilde{G}_v) \leq 2\epsilon$. We can now apply the contrapositive of Lemma 7.1 with δ/C in place of δ and 2ϵ in place of ϵ, which implies that there is a $t \leq k$ such that the t-th moment of G'_v is more than $(t-1)!(\delta/2C\epsilon)^t\epsilon/t$ far from that of $N(0,1)$. That is,

$$|\mathbf{E}_{X\sim G'}[(v\cdot X)^t] - \mathbf{E}_{X\sim N(0,I)}[(v\cdot X)^t]| \geq (t-1)!(\delta/2C\epsilon)^t\epsilon/t \; .$$

Now consider the polynomial $(v\cdot\mathbf{x})^t$. Note that the coefficient of a monomial $\prod_i x_i^{a_i}$, for $\mathbf{a} \in \mathbb{Z}^n_{>0}$, $\|\mathbf{a}\|_1 = t$, is given by the multinomial theorem as $\binom{t}{a_1,\ldots,a_n}\prod_{i=1}^n v_i^{a_i}$. The L_1-norm of its coefficients is the same as the L_1-norm of entries of the rank-t tensor $v^{\otimes t}$, which has $\mathbf{i}$-th entry $\prod_{j=1}^t v_{i_j}$, for $\mathbf{i} \in \{0,\ldots,n\}^t$, since there are $\binom{t}{a_1,\ldots,a_n}$ entries in this symmetric tensor which are given by $\prod_{i=1}^n v_i^{a_i}$ for any $\mathbf{a}$. The Frobenius norm of $v^{\otimes t}$, the L_2-norm of its entries, is $\sum_{\mathbf{i}\in\{0,\ldots,n\}^t}\prod_{j=1}^t v_{i_j}^2 = \prod_{j=1}^t\sum_{i=1}^n v_i^2 = 1$. Since there are n^t entries, the L_1-norm of its entries must be at most $n^{t/2} \leq n^{k/2}$. We can write $\mathbf{E}_{X\sim G''}[(v\cdot X)^t] - \mathbf{E}_{X\sim N(0,I)}[(v\cdot X)^t]$ as a linear combination of differences in moments with these coefficients, and we can thus bound this from above by the product of the L_1-norm of the coefficients and the L_∞-norm of the differences in moments. Hence, there must be some moment $\mathbf{E}[\prod_i X_i^{a_i}]$ with $\mathbf{a} \in \mathbb{Z}^n_{>0}$, $\|\mathbf{a}\| \leq k$, such that

$$\left|\mathbf{E}_{X\sim G'}\left[\prod_i X_i^{a_i}\right] - \mathbf{E}_{X\sim N(0,I)}\left[\prod_i X_i^{a_i}\right]\right| \geq (t-1)!(\delta/2C\epsilon)^t\epsilon/t\cdot n^{-k/2} \; .$$

This in turn means that the difference in the approximation of this moment of G'' and that of $N(0,I)$ is at most $((t-1)!(\delta/2C\epsilon)^t/t - 1)\cdot n^{-k/2}\epsilon$. Thus, the testing algorithm outputs "NO". □

7.3 Robust Learning Algorithm In this section, we build on the testing algorithm of the previous section to design our robust learning algorithm. Formally, we prove:

Theorem 7.7. *Let G' be an ϵ-noisy version of an n-dimensional Gaussian with identity covariance matrix, $N(\mu, I)$ with $\|\mu\|_2 \leq \mathrm{poly}(n, \epsilon)$. There is an algorithm that, given statistical query access to G', outputs an approximation $\widetilde{\mu}$ to the mean μ such that $\|\mu - \widetilde{\mu}\|_2 \leq O(\epsilon)$. The algorithm uses $n^{O(\sqrt{\log(1/\epsilon)})} + 2^{\log(1/\epsilon)^{O(\sqrt{\log(1/\epsilon)})}}$ calls to $\mathrm{STAT}(\epsilon/(n\ln(1/\epsilon))^{O(\sqrt{\log(1/\epsilon)})})$ and has running time $n^{O(\sqrt{\log(1/\epsilon)})} + 2^{\log(1/\epsilon)^{O(\sqrt{\log(1/\epsilon)})}}$.*

By simulating the statistical queries with samples, we obtain:

Corollary 7.8. *Given sample access to G', an ϵ-noisy version of an n-dimensional Gaussian $N(\mu, I)$, there is an algorithm that with probability $9/10$ outputs $\widetilde{\mu}$ with $\|\mu - \widetilde{\mu}\|_2 \leq O(\epsilon)$ and requires $(n\log(1/\epsilon))^{O(\sqrt{\log(1/\epsilon)})}/\epsilon^2$ samples and $n^{O(\sqrt{\log(1/\epsilon)})}/\epsilon^2 + 2^{\log(1/\epsilon)^{O(\sqrt{\log(1/\epsilon)})}}$ time.*

The work [DKK+16b] gives algorithms which can compute an approximation μ' with $\|\mu - \mu'\|_2 \leq O(\epsilon\sqrt{\log(1/\epsilon)})$. These algorithms can be expressed as Statistical Query algorithms. However, due to the model of adversary used for robustness in [DKK+16b], the algorithms were expressed there in terms of operations on sets of samples that were drawn before the execution of the algorithm. The filtering algorithms work by successively removing samples from this set and then computing expectations of the current set of remaining samples. The samples that are removed are those that satisfy an explicit condition, we say that they are rejected by a filter. We can implement these algorithms as SQ algorithms by replacing expectations of the current set of remaining samples with the conditional expectation of the input distribution, conditioned on all previous filters accepting. This is similar to the filtering algorithm for learning binary Bayesian networks given in [DKS16b]. Even there, we still used samples to compute the threshold for the filter. We note that using arguments similar to those we use for the algorithm below, all theses algorithms can be expressed as SQ algorithms. In particular, this is the case for Algorithm `Filter-Gaussian-Unknown-Mean`, which we will use as a black box pre-processing step to approximate the mean within $O(\epsilon\sqrt{\log(1/\epsilon)})$.

Instead of dealing with moments, i.e., the expectations of monomials, directly, we will consider expectations of Hermite polynomials, which have a simpler form for normal distributions.

Definition 7.9. We define multi-dimensional normalized Hermite polynomials as follows: for $\mathbf{a} \in \mathbb{Z}^n$, $He_{\mathbf{a}}(\mathbf{x}) = \prod_{i=1}^n He_{a_i}(\mathbf{x}_i)$. We define $n(\mathbf{a}) = \prod_{i=1}^n a_i!$.

Thus we have the following:

Fact 7.10. $\mathbf{E}_{X\sim N(0,I)}[He_{\mathbf{a}}(X)He_{\mathbf{b}}(X)] = \delta_{\mathbf{ab}}n(\mathbf{a})$.

We define a linear combination of these Hermite polynomials with degree t associated with a tensor of rank t. For $i \in \{0, \ldots, n\}^t$, we define the count vector $c(i) \in \mathbb{Z}^n_{\geq 0}$ such that $c(i)_j$ is the number of coordinates of i that are j.

Definition 7.11. For a rank-t tensor A over $\mathbb{R}^n$, we define

$$h_A(\mathbf{x}) = \sum_{\mathbf{i}\in\{0,\ldots,n\}^t} He_{c(\mathbf{a})}(\mathbf{x})/\sqrt{t!}\,.$$

We are now ready to describe our learning algorithm.

Robust Learning Algorithm:

1. Let $k = 2\lceil\sqrt{\ln(1/\epsilon)})\rceil$.

2. Compute an approximation μ' with $\|\mu - \mu\|_2 \leq O(\epsilon\sqrt{\log(1/\epsilon)})$ by iterating Algorithm `Filter-Gaussian-Unknown-Mean` from [DKK+16b]. We change the origin so that $\mu' = 0$.

3. Let N be the filter that accepts when $\|\mathbf{x}\|_2 \leq \sqrt{2n\log(1/\epsilon)}$.

4. For $1 \leq t \leq k$, let $\widetilde{P}_t$ be the rank-t tensor with $i_1, \ldots, i_t$ entry given by $\sqrt{t!}$ times the result of asking an SQ oracle for $\mathbf{E}_{X\sim G'}[h_{c(i)}(X)]$ conditioned on N accepting to within precision $\epsilon/n^{t/2}$.

5. While $\|\widetilde{P}_t\|_F \geq \epsilon\Omega(\log(1/\epsilon))^{t/2}$ for some t,

   - Let t' be the least t such that $\|\widetilde{P}_{t'}\|_F \geq \epsilon\Omega(\log(1/\epsilon))^{t'/2}$.
   - Let $A = \widetilde{P}_{t'}/\|\widetilde{P}_{t'}\|_F$. Let $h_A(x) = \sum_{i_1,\ldots i_{t'}} A_i He_{c(i)}(x)/\sqrt{t!}$ For each positive integer T, approximate
   $$\Pr_{X\sim\widetilde{G}}[|h_A(X)| \geq T+1]$$
   until one is found that is at least
   $$3\exp(2 - \Omega(T)^{2/t'})) + \epsilon/Cn^{2t'}$$
   for a sufficiently large constant C. Let F be the filter that accepts when $|h_A(x)| \leq T+1$.
   - Recalculate $\widetilde{P}_t$, for all $1 \leq t \leq k$, where all expectations are conditioned N and the filters F from all previous iterations.

6. End while.

7. For $1 \leq t \leq k$, compute the SVD of $M(\widetilde{P}_t)$, $\widetilde{P}_t$ considered as an $n \times n^{t-1}$ matrix, and let $V_t \subseteq \mathbb{R}^n$ be the subspace spanned by all right singular vectors of P_t with singular value more than ϵ.

8. Let V be the span of $V_1, \ldots, V_k$.

9. Let $S \subset V$ be a set of unit vectors of size $|\dim(V)|^{O(\dim(V))}$ such that for any unit vector $v \in V$, there is a $v' \in S$ with $\|v - v'\|_2 \leq 1/2$.

10. For each $v \in S$, compute the median m_v of $v^T X$, for $X \sim G'$, to within $\epsilon/\sqrt{\dim V}$ using bisection and statistical queries to approximate the $\Pr[v^T X \leq m]$ for $m = \mu' + O(\epsilon\sqrt{\log(1/\epsilon)})$. (We don't need to condition on any filters here).

11. Find a feasible point $\widetilde{\mu}_V$ of the LP $\widetilde{\mu}_V \in V$ with $|v^T\widetilde{\mu}_V - v^T m_v| \leq O(\epsilon)$ for all $v \in S$.

12. Return $\widetilde{\mu}_V$.

Using similar techniques to those used to express this algorithm in terms of Statistical Queries, we can run Algorithm `Filter-Gaussian-Unknown-Mean` using $\mathrm{poly}(n/\epsilon)$ time and calls to $\mathrm{STAT}(\widetilde{O}(\epsilon/\mathrm{poly}(n)))$.

Note that we can approximate conditional expectations easily as a ratio of expectations approximated by two SQ queries. Since, as we will show, our filters only throw away at most an $O(\epsilon)$ fraction of points, we will not need to increase the precision beyond a constant factor to do this.

The algorithm needs approximate expectations to within $\epsilon/n^{O(\sqrt{\log(1/\epsilon)})}$. To show that we can use the oracle $\mathrm{STAT}(\epsilon/(n\ln(1/\epsilon))^{O(\sqrt{\log(1/\epsilon)})})$ to obtain this, we need to note that the distributions we approximate the expectations of are supported in an interval of length $(n\ln(1/\epsilon))^{O(\sqrt{\log(1/\epsilon)})}$. Thanks to the naive pruning of Step 3, only $\mathbf{x}$ with $\|\mathbf{x}\|_2 \leq \sqrt{2n\log(1/\epsilon)}$ contribute to these expectations. This suffices to show that the maximum value of all polynomials we consider on any such $\mathbf{x}$ is at most $(n\ln(1/\epsilon))^{O(\sqrt{\log(1/\epsilon)})}$.

We need to show the following for the filter step of our algorithm:

Proposition 7.12. *The loop in Step 5 takes $O(n^{2k})$ iterations and all filters together accept with probability at least $1 - O(\epsilon)$.*

7.3.1 Proof of Proposition 7.12 We now proceed with the proof. By standard concentration bounds, N accepts with probability $1 - O(\epsilon)$. Let F' be the event that N and all filters F from previous iterations accept. We assume inductively that $\Pr_{G'}[F'] \leq O(\epsilon)$, and need to show that the same holds if we include the filter F produced in the current iteration.

Let $\widetilde{G} = N(\mu, I)$ be a Gaussian with $d_{\mathrm{TV}}(G', \widetilde{G}) \leq \epsilon$ such that $\|\mu\|_2 \leq O(\epsilon\sqrt{\log(1/\epsilon)})$. We write $G'|F'$ for the distribution obtained by conditioning on F'. Since $\Pr_{G'}[F'] \leq C\epsilon$, we have that

$$d_{\mathrm{TV}}(G'|F, \widetilde{G}) \leq d_{\mathrm{TV}}(G', \widetilde{G}) + d_{\mathrm{TV}}(G', G'|F') \leq (C+1)\epsilon \, .$$

Thus, for the first iteration, we have $G'|F' = \widetilde{G} + d_{\mathrm{TV}}(G'|F, \widetilde{G})E - d_{\mathrm{TV}}(G'|F, \widetilde{G})L$, for distributions E and L with disjoint supports.

For any iteration, we will write $G'|F' = w_{\widetilde{G}}\widetilde{G} + w_E E - w_L L$, for distributions E and L with disjoint supports, where $w_E + w_L = O(\epsilon)$ and $w_{\widetilde{G}} = 1 + O(\epsilon)$. In the first iteration, we will take $w_{\widetilde{G}} = 1$ and $w_E = w_L = d_{\mathrm{TV}}(G'|F', \widetilde{G})$.

We will need properties of the polynomials $h_A(x)$ for the analysis. In particular, we show the following:

Lemma 7.13. *Given a rank-t symmetric tensor A over $\mathbb{R}^n$, let $h_A(x)$ be as in Definition 7.11. Then, we have:*

(i) $\mathbf{E}_{X\sim N(0,I)}[h_A(X)^2] = \|A\|_F^2$.

(ii) If B is a rank-t tensor with $B_i = \sqrt{t!}\mathbf{E}_{X\sim\mathbf{P}}[He_{c(i)}(X)]$, for a distribution $\mathbf{P}$, then $\mathbf{E}_{X\sim\mathbf{P}}[h_A(X)] = \sum_i A_i B_i$.

(iii) We can recover A from $h_A(x)$ using $\sqrt{t!}A_{i_1,\ldots,i_t} = \frac{\partial}{\partial x_{i_1}} \cdots \frac{\partial}{\partial x_{i_t}} h_A(x)$.

(iv) If O is an orthogonal matrix, then $h_A(O\mathbf{x}) = h_B(\mathbf{x})$ for a symmetric rank-t tensor B with $\|B\|_F = \|A\|_F$.

(v) If B is a rank-t tensor with $B_i = \sqrt{t!}\mathbf{E}_{X\sim\mathbf{P}}[He_{c(i)}(X)]$, for a distribution $\mathbf{P}$, and $j > 0$, $v \in \mathbb{R}^n$, then $\mathbf{E}_{X\sim\mathbf{P}}[He_j(v \cdot X)]/\sqrt{t!} = B(v, \ldots, v)$.

Proof. For (i), we first need to get an expression for the coefficients of each $He_{\mathbf{a}}(x)$, since they appear multiple times in $h_A(\mathbf{x}) = \sum_{\mathbf{i}\in\{0,\ldots,n\}^t} A_{\mathbf{i}} He_{c(\mathbf{a})}(\mathbf{x})/\sqrt{t!}$. Let $c^{-1}(a)$ be a function mapping $\mathbf{a} \in \mathbb{Z}^t_{\geq 0}$ with $\|\mathbf{a}\|_1 = t$ to $\mathbf{i} \in \{1, \ldots, n\}^t$, with $c(c^{-1}(\mathbf{a})) = \mathbf{a}$ for all $\mathbf{a}$. Since A is symmetric, the choice of c^{-1} does not affect $A_{c^{-1}(\mathbf{a})}$ for any $\mathbf{a}$. Note that, for a given $\mathbf{a}$, there are $\binom{t}{a_1,\ldots,a_n} = t!/n(\mathbf{a})$ possible $\mathbf{i}$ with $c(\mathbf{i}) = \mathbf{a}$. Thus, we have:

$$\begin{aligned} h_A(\mathbf{x}) &= \sum_{\mathbf{i}\in\{0,\ldots,n\}^t} A_{\mathbf{i}} He_{c(\mathbf{a})}(\mathbf{x})/\sqrt{t!} \\ &= \sum_{\|\mathbf{a}\|_1 = t} t!/n(\mathbf{a}) \cdot A_{c^{-1}(\mathbf{a})} He_{\mathbf{a}}(\mathbf{x})/\sqrt{t!} \\ &= \sum_{\|\mathbf{a}\|_1 = t} \sqrt{t!}/n(\mathbf{a}) \cdot A_{c^{-1}(\mathbf{a})} He_{\mathbf{a}}(\mathbf{x}) \, . \end{aligned}$$

Now, by orthogonality of $He_{\mathbf{a}}(\mathbf{x})$ with distinct $\mathbf{a}$, we have that, for $X \sim N(0, I)$, it holds:

$$\begin{aligned}\mathbf{E}[h_A(X)^2] &= \sum_{\|\mathbf{a}\|_1=t} t!/n(\mathbf{a})^2 \cdot A^2_{c^{-1}(\mathbf{a})} \mathbf{E}[He_{\mathbf{a}}(X)^2] \\ &= \sum_{\|\mathbf{a}\|_1=t} t!/n(\mathbf{a}) \cdot A^2_{c^{-1}(\mathbf{a})} \\ &= \sum_{\mathbf{i}} A^2_{\mathbf{i}} = \|A\|_F^2 \ .\end{aligned}$$

For (ii), we now have that

$$\begin{aligned}\mathbf{E}_{X\sim\mathbf{P}}[h_A(X)] &= 1/\sqrt{t!} \cdot \sum_{\mathbf{i}} A_{\mathbf{i}} \mathbf{E}_{X\sim\mathbf{P}}[He_{c(\mathbf{i})}(x)] \\ &= \sum_{\mathbf{i}} A_{\mathbf{i}} B_{\mathbf{i}} \ .\end{aligned}$$

For (iii), note that $\mathbf{a}$ with $\|\mathbf{a}\|_1 = t$, $He_{\mathbf{a}}(x)$ has only one monomial of degree t, which is $\prod x_i^{a_i}$. Thus, given $\mathbf{i} \in \{1, \ldots, n\}^t$, there is only one $\mathbf{a}$ with $\|\mathbf{a}\|_1 = t$ and $\frac{\partial}{\partial x_{i_1}} \cdots \frac{\partial}{\partial x_{i_t}} He_{\mathbf{a}}(x) \neq 0$, which is $\mathbf{a} = c(\mathbf{i})$ and has

$$\begin{aligned}\frac{\partial}{\partial x_{i_1}} \cdots \frac{\partial}{\partial x_{i_t}} He_{c(\mathbf{i})}(x) &= \frac{\partial}{\partial x_{i_1}} \cdots \frac{\partial}{\partial x_{i_t}} \prod_j x_j^{c(\mathbf{i})_j} \\ &= \prod_j c(\mathbf{i})_j! = n(c(\mathbf{i})) \ .\end{aligned}$$

Thus, we have

$$\begin{aligned}\frac{\partial}{\partial x_{i_1}} \cdots \frac{\partial}{\partial x_{i_t}} h_A(\mathbf{x}) &= n(c(\mathbf{i})) \cdot \sqrt{t!}/n(\mathbf{a}) \cdot A_{c^{-1}}(\mathbf{a}) \\ &= \sqrt{t!} A_{\mathbf{i}} \ .\end{aligned}$$

For (iv), consider the linear transformation of rank-t tensors $O^{\otimes t}$, which for our purposes can be defined as the unique function such that $(O^{\otimes t}A)(v_1, \ldots, v_t) = A(O^T v_1, \ldots, O^T v_t)$, for all rank-$t$ tensors over $\mathbb{R}^n$, A, and vectors $v_1 \ldots v_t \in \mathbb{R}^n$. It is a fact that $\|O^{\otimes t}A\|_F = \|A\|_F$. Now, we consider the t-th order directional derivatives of a function f, $\nabla_{v_1}, \cdots \nabla_{v_t} f(\mathbf{x})$, where $\nabla_v g(\mathbf{x}) = \sum_i v_i \frac{\partial g}{\partial x_i}(\mathbf{x})$. These can be expressed in terms of the rank-t tensor F, with $F_{\mathbf{i}} = \frac{\partial}{\partial x_{i_1}} \cdots \frac{\partial}{\partial x_{i_t}} f(x)$, as $\nabla_{v_1}, \ldots \nabla_{v_t} f(\mathbf{x}) = F(v_1, \ldots v_t)$. We have that, for all $\mathbf{i}$,

$$\begin{aligned}&\frac{\partial}{\partial x_{i_1}} \cdots \frac{\partial}{\partial x_{i_t}} h_A(O\mathbf{x}) \\ &= \nabla_{\mathbf{e}_{i_1}} \cdots \nabla_{\mathbf{e}_{i_t}} h_A(O\mathbf{x}) \\ &= \nabla_{O^T\mathbf{e}_{i_1}} \cdots \nabla_{O^T\mathbf{e}_{i_t}} h_A(\mathbf{x}) \\ &= \sqrt{t!} A(O^T\mathbf{e}_{i_1}, \ldots, O^T\mathbf{e}_{i_t}) \\ &= \sqrt{t!} (O^{\otimes t}A)_i \\ &= \frac{\partial}{\partial x_{i_1}} \cdots \frac{\partial}{\partial x_{i_t}} h_{O^{\otimes t}A}(\mathbf{a}) \ .\end{aligned}$$

Since $h_A(O\mathbf{x})$ and $h_{O^{\otimes t}A}(\mathbf{x})$ are both multivariate polynomials of degree t, that these derivatives agree means that the coefficients of all monomials of degree-t agree. We thus have $h_A(O\mathbf{x}) = h_{O^{\otimes t}A}(\mathbf{x}) + p(x)$, where p is a polynomial of degree at most $t-1$. Since $h_{O^{\otimes t}A}(\mathbf{x})$ is a linear combination of Hermite polynomials of degree t, which are orthogonal to all polynomials of degree smaller than t, we have $\mathbf{E}_{X\sim N(0,I)}[h_{O^{\otimes t}A}(X)p(X)] = 0$, and so

$$\begin{aligned}
1 = \|A\|_F &= \mathbf{E}_{X\sim N(0,I)}[h_A(O\mathbf{x})^2] \\
&= \mathbf{E}_{X\sim N(0,I)}[h_{O^{\otimes t}A}(X)^2] + \mathbf{E}_{X\sim N(0,I)}[p(X)^2] + 2\mathbf{E}_{X\sim N(0,I)}[h_{O^{\otimes t}A}(X)p(X)] \\
&= \|O^{\otimes t}A\|_F + \mathbf{E}_{X\sim N(0,I)}[p(X)^2] + 0 \\
&= 1 + \mathbf{E}_{X\sim N(0,I)}[p(X)^2] \ .
\end{aligned}$$

Since $\mathbf{E}_{X\sim N(0,I)}[p(X)^2] = 0$, we must have $p(\mathbf{x}) \equiv 0$, and thus

$$h_A(O\mathbf{x}) = h_{O^{\otimes t}A}(\mathbf{x}) \ .$$

Taking $B = O^{\otimes t}A$ gives (iv).

For (v), let O be an orthogonal matrix that gives a rotation mapping e_1 to v, $\mathbf{a} = \{t, 0, \ldots, 0\}$ and T_1 be the rank-t tensor with $(1, \ldots, 1)$ entry 1 and every other entry 0. Then, we can rewrite

$$He_t(v\cdot\mathbf{x}) = He_t((O\mathbf{x})_1) = He_{\mathbf{a}}((O\mathbf{x})_1) = \sqrt{t!}h_{T_1}(O\mathbf{x}) = \sqrt{t!}h_{O^{\otimes t}T_1}(\mathbf{x}) \ .$$

For any $\mathbf{i}$, we have

$$\begin{aligned}
(O^{\otimes t}T_1)_{\mathbf{i}} &= (O^{\otimes t}T_1)(\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_t}) \\
&= T_1(O^T\mathbf{e}_{i_1}, \ldots, O^T\mathbf{e}_{i_t}) \\
&= \prod_{j=1}^{t}(O^T\mathbf{e}_{i_j})_1 \\
&= \prod_{j=1}^{t}(O\mathbf{e}_1)_{i_j} \\
&= \prod_{j=1}^{t} v_{i_j} \ .
\end{aligned}$$

Thus we have $O^{\otimes t}T_1 = v^{\otimes t}$, the tensor with entries $\prod_{j=1}^{t} v_{i_j}$, and so

$$\begin{aligned}
\mathbf{E}_{X\sim\mathbf{P}}[He_t(v\cdot X)]/\sqrt{t!} &= \mathbf{E}_{X\sim\mathbf{P}}[h_{v^{\otimes t}}(X)] \\
&= \sum_{\mathbf{i}} B_{\mathbf{i}} \prod_{j=1}^{t} v_{i_j} = B(v, \ldots, v) \ . \qquad \text{(by (ii))}
\end{aligned}$$

This completes the proof. □

We write $P_{t'}$ or $G_{t'}$ for the rank-t' tensor with entries $\sqrt{t!}\mathbf{E}[h_{c(i)}(X)]$, where X is distributed according to $G'|F'$ or $\widetilde{G}$ respectively. We know that $\|\widetilde{P}_{t'}\|_F \geq \epsilon\Omega(\log(1/\epsilon)^{t'/2})$.

Lemma 7.14. *When $\|\widetilde{P}_{t'}\|_F \geq \epsilon\Omega(\log(1/\epsilon)^{t'/2})$, we have $|\mathbf{E}_{X\sim G'}[h_A(X)]| \geq \epsilon\cdot\Omega(\log(1/\epsilon))^{t/2}$.*

Proof. The assumption on the SQ errors imply that the corresponding entries of $P_{t'}$ and $\widetilde{P}_{t'}$ are within $\epsilon/n^{t'/2}$. It follows that

$$\|P_{t'} - \widetilde{P}_{t'}\|_F \leq \sqrt{t'!}\epsilon \ .$$

Using Lemma 7.13 (iii), we have

$$\begin{aligned}
\mathbf{E}_{X\sim G'}[h_A(X)] &= \sum_{i\in[n]^{t'}} A_i (P_{t'})_i \\
&= 1/\|\widetilde{P}_t\|_F \cdot \sum_{i\in[n]^{t'}} (\widetilde{P}_{t'})_i (P_{t'})_i \\
&= 1/\|\widetilde{P}_t\|_F \cdot \left(\|\widetilde{P}_{t'}\|_F^2 + \sum_{i\in[n]^{t'}} (\widetilde{P}_{t'})_i (P_{t'} - \widetilde{P}_{t'})_i \right) \\
&\geq \|\widetilde{P}_{t'}\|_F - \|P_{t'} - \widetilde{P}_{t'}\|_F \geq \|\widetilde{P}_{t'}\|_F - \sqrt{t'!}\epsilon \\
&\geq \epsilon\Omega(\log(1/\epsilon)^{t'/2})) - \epsilon\sqrt{\log(1/\epsilon)}^{t'} \\
&\geq \epsilon\Omega(\log(1/\epsilon)^{t'/2}) \ .
\end{aligned}$$

□

Lemma 7.15. $\mathbf{E}_{X\sim\widetilde{G}}[h_A(X)] = O(\epsilon\sqrt{\log(1/\epsilon)})^t$ *and* $\mathbf{E}_{X\sim\widetilde{G}}[h_A(X)^2] = O(1)$.

Proof. Since for all $\mathbf{a} \in \mathbb{Z}^k_{>0}$, $\mathbf{E}_{X\sim N(0,I)}[He_{\mathbf{a}}(X)] = 0$, we have $\mathbf{E}_{X\sim N(0,I)}[h_A(X] = 0$. By Lemma 7.13 (i), $\mathbf{E}_{X\sim N(0,I)}[h_A(X)^2] = \|A\|_F^2 = 1$.

We need to take these expectations under $\widetilde{G} = N(\mu, I)$ instead of $N(0, I)$. Consider a rotation given by an orthogonal matrix O that maps $\|\mu\|_2\mathbf{e}_1$ to μ. By Lemma 7.13 (i), there is a symmetric rank-t tensor B with $\|B\|_F = 1$ such that $h_A(OX) = h_B(X)$. Now we have that

$$\mathbf{E}_{X\sim G'}[h_A(X)] = \mathbf{E}_{X\sim N(\|\mu\|_2\mathbf{e}_1,I)}[h_A(OX)] = \mathbf{E}_{X\sim N(\|\mu\|_2\mathbf{e}_1,I)}[h_B(X)] \ ,$$

and similarly for $h_A(X)^2$.

For $\mathbf{a} \in \mathbb{Z}^n_{\geq 0}$ with $\mathbf{a} \neq 0$, writing $\mathbf{a}_{-1} \in \mathbb{Z}^{n-1}_{\geq 0}$ for the vector dropping the first coordinate of $\mathbf{a}$, we have

$$\begin{aligned}
\mathbf{E}_{X\sim N(\|\mu\|_2 e_1,I)}[He_{\mathbf{a}}(X)] &= \mathbf{E}_{X\sim N(0,I)}[He_{\mathbf{a}_{-1}}(X)]\mathbf{E}_{X\sim N(\|\mu\|_2,1)}[He_{a_1}(X)] \\
&= \delta_{\mathbf{a}_{-1},0}\mathbf{E}_{X\sim N(0,1)}[He_{a_1}(X + \|\mu\|_2)] \ .
\end{aligned}$$

Note that there is only one index i such that $c(i)$ is zero in all except the first coordinate, and so we have

$$\mathbf{E}_{X\sim\widetilde{G}}[h_A(X)] = \mathbf{E}_{X\sim N(\|\mu\|_2 e_1,I)}[h_B(X)] = B_{1,\ldots,1}\mathbf{E}_{X\sim N(0,1)}[He_t(X + \|\mu\|_2)/\sqrt{t!}] \ .$$

Since $\|B\|_F = 1$, we have

$$\left|\mathbf{E}_{X\sim\widetilde{G}}[h_B(X)]\right| \leq |\mathbf{E}_{X\sim N(0,1)}[He_t(X + \|\mu\|_2)/\sqrt{t!}]|.$$

By standard results, we have that $\frac{dHe_i}{dx}(x) = iHe_{i-1}(x)$, and so by Taylor's theorem we have $He_i(x + \|\mu\|_2) = \sum_{j=0}^i \binom{i}{j}\|\mu\|_2^j He_{i-j}(x)$. Thus,

$$\begin{aligned}
|\mathbf{E}_{X\sim N(0,1)}[He_t(X + \|\mu\|_2)]| &= \left|\sum_{i=0}^t \binom{t}{i}\|\mu\|_2^i |\mathbf{E}_{X\sim N(0,1)}[He_{t-i}(X)]\right| \\
&= \|\mu\|_2^t \ .
\end{aligned}$$

This gives

$$\left|\mathbf{E}_{X\sim\widetilde{G}}[h_A(X)]\right| \le \|\mu\|_2^t/\sqrt{t!} = O(\epsilon\sqrt{\log(1/\epsilon)})^t/\sqrt{t!}\,,$$

as required.

Similarly, we have, for all $\mathbf{a},\mathbf{b}\in\mathbb{Z}^n_{>0}$,

$$\begin{aligned}\mathbf{E}_{X\sim N(\|\mu\|_2 e_1,I)}[He_{\mathbf{a}}(X)He_{\mathbf{b}}(X)] &= \mathbf{E}_{X\sim N(0,I)}[He_{\mathbf{a}_{-1}}(X)He_{\mathbf{b}_{-1}}(X)]\mathbf{E}_{X\sim N(\|\mu\|_2,1)}[He_{a_1}(X)He_{b_1}(X)]\\ &= \delta_{\mathbf{a}_{-1}\mathbf{b}_{-1}}\cdot\mathbf{E}_{X\sim N(0,1)}[He_{a_1}(X+\|\mu\|_2)He_{b_1}(X+\|\mu\|_2)]\,.\end{aligned}$$

When $\|\mathbf{a}\|_1 = \|\mathbf{b}\|_1 = t$, if $\mathbf{a}_{-1}=\mathbf{b}_{-1}$, then $\mathbf{a}=\mathbf{b}$ (since $a_1=b_1=t-\|\mathbf{a}\|_1$). For $1\le j\le t$, we have:

$$\begin{aligned}\mathbf{E}_{X\sim N(0,1)}[He_j(X+\|\mu\|_2)^2] &= \mathbf{E}_{X\sim N(0,1)}\left[\left(\sum_{i=0}^{j}\binom{j}{i}\|\mu\|_2^i He_{j-i}(x)\right)^2\right]\\ &= \sum_{i=0}^{j}\binom{j}{i}^2\|\mu\|_2^{2i}|\mathbf{E}_{X\sim N(0,1)}[He_{j-i}(X)^2]\\ &= \sum_{i=0}^{j}\binom{j}{i}^2\|\mu\|_2^{2i}i!\\ &= \sum_{i=0}^{j}\|\mu\|_2^{2i}(j!/(j-i)!)^2/i!\\ &\le \sum_{i=0}^{j}\|\mu\|_2^{2i}j^2 i/i!\\ &\le 2\sum_{i=0}^{j}2^{-2i} \qquad\qquad \text{(since } \|\mu\|_2\le 1/2k\le 1/2j\text{)}\\ &\le 3\,.\end{aligned}$$

Putting these together, for $\mathbf{a},\mathbf{b}$ with $\|\mathbf{a}\|_1=\|\mathbf{b}\|_1=t$, we have

$$\left|\mathbf{E}_{X\sim N(\|\mu\|_2 e_1,I)}[He_{\mathbf{a}}(X)He_{\mathbf{b}}(X)]\right| \le 3\delta_{\mathbf{ab}}\,.$$

The sum of squares of coefficients of all $He_{\mathbf{a}}(x)$ in $h_B(X)$ is $\mathbf{E}_{X\sim N(0,1)}[h_B(X)^2] = \|B\|_F = 1$, and so we have that $\mathbf{E}_{X\sim N(\|\mu\|_2 e_1,I)}[h_B(X)^2]| \le 3$. Finally, recall that $\mathbf{E}_{X\sim G}[h(A)^2] = \mathbf{E}_{X\sim N(\|\mu\|_2 e_1,I)}[h_B(X)^2]|$, and so this is $O(1)$, as required. □

Now consider the equation

$$\mathbf{E}_{X\sim G'}[h_A(X)] = w_{\widetilde{G}}\mathbf{E}_{X\sim\widetilde{G}}[h_A(X)] + w_E\mathbf{E}_{X\sim E}[h_A(X)] - w_L\mathbf{E}_{X\sim L}[h_A(X)].$$

We know that the LHS is $\Omega(\epsilon\log(1/\epsilon)^{t/2})$ and that the first term on the RHS is smaller. Therefore, one of the last two terms is small. Since $w_L L\le\widetilde{G}$, we will use standard concentration inequalities to show that $w_L\mathbf{E}_{X\sim L}[h_A(X)]$ is $O(\epsilon\log(1/\epsilon)^{t/2})$. If we cannot find a filter, then $w_E E\le G'$ must satisfy similar concentration inequalities, which would imply that $w_E\mathbf{E}_{X\sim E}[h_A(X)]$ is $O(\epsilon\log(1/\epsilon)^{t/2})$. Since some term on the RHS must be bigger than this, we can find a filter.

Lemma 7.16. *For $X\sim N(0,I)$, if $p(x)$ is a degree-d polynomial with $\mathbf{E}[p(X)^2]\le 1$, we have that*

$$\Pr\left[|p(X)|\ge T+\mathbf{E}[p(X)]\right] \le \exp(2-\Omega(T)^{2/d}))\,.$$

Lemma 7.17. *We have that* $w_L|\mathbf{E}_{X\sim L}[h_A(X)]| \leq \epsilon \cdot O(\log(1/\epsilon))^{t'/2}$.

Proof. We start with the following claim:

Claim 7.18. *For any* $R > 0$, $d \in \mathbb{Z}_+$, $\epsilon > 0$, *and* $\exp(-(a/R)^{2/d}) = \epsilon$, *we have*

$$\int_a^\infty \exp(-(T/R)^{2/d})TdT \leq (d^2/2)\epsilon(d + \ln(1/\epsilon))^{d-1} .$$

Proof. Note that $(a/R)^{2/d} = \ln(1/\epsilon)$. First, we change variables to $x = (T/R)^{2/d}$ to obtain

$$\begin{aligned}\int_{\ln(1/\epsilon)}^\infty \exp(-(T/R)^{2/d})TdT &= \int_{\ln(1/\epsilon)}^\infty \exp(-x)x^{d/2}\frac{dT}{dx}dx \\ &= \int_{\ln(1/\epsilon)}^\infty \exp(-x)x^{d/2} \cdot (Rd/2)x^{d/2-1}dx \\ &= (Rd/2)\int_{\ln(1/\epsilon)}^\infty \exp(-x)x^{d-1}dx .\end{aligned}$$

We can now integrate by parts

$$\int_{\ln(1/\epsilon)}^\infty \exp(-x)x^{d-1}dx = \epsilon \ln(1/\epsilon)^{d-1} + (d-1)\int_{\ln(1/\epsilon)}^\infty \exp(-x)x^{d-2}dx .$$

By a simple induction, we have

$$\int_{\ln(1/\epsilon)}^\infty \exp(-x)x^{d-1}dx = \epsilon \sum_{j=0}^{d-1} d!/(d-j)! \ln(1/\epsilon)^{(d-j-1)} .$$

Now we have

$$\begin{aligned}\int_a^\infty \exp(-(T/R)^{2/d})Tdt &= (d/2)\epsilon \sum_{j=0}^{d-1} d!/(d-j)! \ln(1/\epsilon)^{(d-j-1)} \\ &\leq (d/2)\exp(-(a/R)^{2/d}) \sum_{j=0}^{d-1} d^j \ln(1/\epsilon)^{(d-j-1)} \\ &\leq (d^2/2)\epsilon(d + \ln(1/\epsilon))^{d-1} .\end{aligned}$$

□

Let $\mu_h = \mathbf{E}_{X\sim\widetilde{G}}[h_A(X)]$. We have the following sequence of inequalities:

$$\begin{aligned}
w_L\mathbf{E}_{X\sim L|}[h_A(X)^2] &= \int_0^\infty T w_L \Pr_{X\sim L}[|h_A(X)| > T]dT \\
&\le \int_0^\infty T\min\{w_L, \Pr_{X\sim\widetilde{G}}[|h_A(X)| > T]\}dT \\
&\le \int_0^\infty T\min\{w_L, \exp(2-\Omega(T-\mu_h)^{2/t'}))\}dT \\
&\le \int_0^\infty T\min\{w_L, \exp(2-((T-\mu_h)/R)^{2/t'}))\}dT \qquad \text{(for some } R>0\text{)} \\
&= \int_{-\mu_h}^\infty (T+\mu_h)\min\{w_L, \exp(2-(T/R)^{2/t'}))\}dT \\
&= \int_{-\mu_h}^c (T+\mu_h)w_L dT + \int_c^\infty \exp(2-(T/R)^{2/t'})(T+\mu_h)dT \\
&\qquad\qquad \text{(where } c = (R\ln 1/w_L)^{t'/2}\text{)} \\
&\le w_L(c+\mu_h)^2/2 + e^2(c+\mu_h)/c\cdot\int_c^\infty \exp(-(T/R)^{2/t'})TdT \\
&\le w_L O(\ln(1/w_L))^{t'} + O(t'^2 w_L(t'+\ln(1/w_L))^{t'-1}) \\
&\le \epsilon\cdot O(\ln(1/\epsilon))^{t'}\ ,
\end{aligned}$$

where the last line follows from $t' \le k = O(\sqrt{\log(1/\epsilon)}) \le O(\log(1/\epsilon))$. Then, by an application of the Cauchy-Schwarz inequality, we have that

$$w_L|\mathbf{E}_{X\sim L}[h_A(X)]| \le w_L\mathbf{E}_{X\sim L}[|h_A(X)|] \le \sqrt{w_L^2\mathbf{E}_{X\sim L}[h_A(X)]^2} \le \epsilon\cdot O(\log(1/\epsilon))^{t'/2}.$$

This completes the proof of the lemma. □

Lemma 7.19. *If* $\Pr_{X\sim G'|F'}[|h_A(X)| \ge T+1] \le O(\exp(2-\Omega(T)^{2/d})) + \epsilon/(2n)^{2t'}$*, for all integers* T*, then* $w_E|\mathbf{E}_{X\sim E}[h_A(X)]| \le O(\epsilon\ln(1/\epsilon)^{t'/2})$.

Proof. Since F' includes the filter M, we have that the support of $G'|F'$ and the support of E includes only x with $\|x\|_2 \le \sqrt{2n\ln(1/\epsilon)}$:

Claim 7.20. *When* $\|x\|_2 \le \sqrt{2n\ln(1/\epsilon)}$*, then* $|h_A(x)| \le (2n\sqrt{\ln(1/\epsilon)})^t$.

Proof. Note that $t' \le k \le \sqrt{2n\log(1/\epsilon)}$. Using the explicit formula for the coefficient $He_i(x)$, we can

show that for $|x| \leq \sqrt{2n\ln(1/\epsilon)}$, with $k \geq i$, the $He_i(x)$ is dominated by its leading coefficient:

$$\begin{aligned}
|He_i(x)| &= \left| \sum_{j=0}^{\lfloor i/2 \rfloor} i!(-1)^j x^{i-2j}/j!(i-2j)!2^j \right| \\
&\leq \sum_{j=0}^{\lfloor i/2 \rfloor} j!(3/2)^j |x|^{i-2j} \\
&\leq \sum_{j=0}^{\lfloor i/2 \rfloor} j!(3/2)^j (\sqrt{2n\ln(1/\epsilon)})^{i-2j} \\
&\leq \sum_{j=0}^{\lfloor i/2 \rfloor} (2n\log(1/\epsilon))^{(i-j)/2} \\
&\leq 2 \cdot (n\log(1/\epsilon))^{i/2} \ .
\end{aligned}$$

Therefore, for any $a \in \mathbb{Z}^n_{\geq 0}$ with $\sum_i a_i = t$, and $\|x\|_2 \leq \sqrt{2n\log(1/\epsilon)}$, we have that

$$\begin{aligned}
|h_a(x)| &= \prod_{i=1}^{n} |He_{a_i}(x)/\sqrt{a_i!}| \\
&\leq \prod_{i=1}^{n} 2\sqrt{n\log(1/\epsilon)}^{a_i} \\
&\leq (2\sqrt{n\log(1/\epsilon)})^t \ .
\end{aligned}$$

Since $\|A\|_F = 1$ and A has $n^{t'}$ entries, the L_1-norm of the entries is at most $n^{t/2}$. Thus, we have that $|h_A(x)| \leq n^{t/2} \cdot (2\sqrt{n\log(1/\epsilon)})^t = (2n\sqrt{\log(1/\epsilon)})^t$. □

We note that since

$$\Pr_{X \sim G'|F'}[|h_A(X)| \geq T+1] \leq O(\exp(2-\Omega(T)^{2/d})) + \epsilon/(2n)^{2t'} \ ,$$

for integers T, that

$$\Pr_{X \sim G'|F'}[|h_A(X)| \geq T+2] \leq O(\exp(2-\Omega(T)^{2/d})) + \epsilon/(2n)^{2t'} \ ,$$

for all T.

Similarly to the proof of Lemma 7.17, we obtain:

$$\begin{aligned}
w_E \mathbf{E}_{X\sim E}[h_A(X)^2] &= \int_0^\infty T w_E \Pr_{X\sim E}[|h_A(X)| > T] dT \\
&= \int_0^{(2n\sqrt{\log(1/\epsilon)})^t} T w_E \Pr_{X\sim E}[|h_A(X)| > T] dT \\
&\le \int_0^{(2n\sqrt{\log(1/\epsilon)})^t} T \min\{w_E, \Pr_{X\sim G'}[|h_A(X)| > T]\} dT \\
&\le \int_0^{(2n\sqrt{\log(1/\epsilon)})^t} T \min\{w_E, O(\exp(2-\Omega(T-2)^{2/t}) + \epsilon/Cn^{2t'})\} dT \\
&\le \int_{-2}^{(2n\sqrt{\log(1/\epsilon)})^t - 1} (T+2) \min\{w_E, O(\exp(2-(T/R)^{2/t}) + \epsilon/Cn^{2t'})\} dT \\
&\le \int_0^{(2n\sqrt{\log(1/\epsilon)})^t} O(T\epsilon/(2n)^{2t'}) dT + \int_{-2}^{c} (T+2) w_E dT \\
&+ (c+1)/c \cdot \int_c^\infty O(\exp(-(T/R)^{2/t})) T dT \qquad \text{(where } c = (R \ln 1/w_E)^{t/2}\text{)} \\
&= (2n\sqrt{\log(1/\epsilon)})^{2t} \cdot \epsilon/(2n)^{2t'} + O(w_E c^2/2) + O(t'^2 w_E (t' + \ln(1/w_E))^{t'-1}) \\
&\le \epsilon \cdot (4\log(1/\epsilon)^t + w_E \cdot O(\ln 1/w_E)^t \\
&\le \epsilon \cdot O(\log(1/\epsilon))^t \ .
\end{aligned}$$

Then, by the Cauchy-Schwarz inequality, we conclude that

$$w_E |\mathbf{E}_{X\sim E}[h_A(X)]| \le w_E \mathbf{E}_{X\sim E}[|h_A(X)|] \le \sqrt{w_E^2 \mathbf{E}_{X\sim E}[h_A(X)]^2} \le \epsilon \cdot O(\log(1/\epsilon))^{t/2} \ .$$

This completes the proof. □

As an immediate consequence, we obtain:

Corollary 7.21. *There is an integer $0 \le T \le O(n\sqrt{\log(1/\epsilon)})^t$ such that*

$$\Pr_{X\sim G'|F'}[|h_A(X)| \ge T+1] \ge 3\exp(2-\Omega(T)^{2/d})) + 2\epsilon/Cn^{2t'} \ .$$

We can now prove the following crucial lemma:

Lemma 7.22. *The algorithm finds a T with*

$$\Pr_{X\sim G'|F'}[|h_A(X)| \ge T+1] \le 3\exp(2-\Omega(T)^{2/d})) + \epsilon/Cn^{2t'} \ .$$

Proof. By Corollary 7.21, such a T exists, and therefore our algorithm will find one after enumerating $O(n\sqrt{\log(1/\epsilon)})^t$ possibilities. □

Let F be the event that the new filter accepts. In the next iterations, we will use $G'|F' \cap F$ instead of $G'|F'$. We need to show that the parameters $w_{\widetilde{G}}$, w_E and w_L improve in such a way that we only need a bounded number of iterations:

Claim 7.23. *We can write $G'|F' \cap F = w'_{\widetilde{G}} \widetilde{G} + w'_E E' - w'_L L'$, where L' and E' have disjoint supports $w'_E, w'_L > 0$ and $w'_E + w'_L \le w_E + w_L - \epsilon/Cn^{2t'}$. The probability that the filter rejects is at most $O(w_E + w_L - w'_E - w'_L)$.*

Proof. This proof is very similar to that of Claim 26 from [DKS16b]. Let $\neg F$ be the event that the filter rejects, i.e., that $|h_A(X)| \geq T+1$. We have that $\Pr_{G'|F'}[\neg F] \geq 3\exp(2-\Omega(T)^{2/d})) + \epsilon/Cn^{2t'}$. On the other hand, by the concentration inequality, $\Pr_{\widetilde{G}}[\neg F] \leq \exp(2-\Omega(T)^{2/d}))$. Thus, we have that

$$\Pr_{G'|F'}[\neg F] \geq 3\Pr_{\widetilde{G}}[\neg F] + \epsilon/Cn^{2t'} \ .$$

However, the defining relation between $G'|F'$ and $\widetilde{G}$, E and L yields for the event $\neg F$ that

$$\Pr_{G'|F'}[\neg F] \geq w_{\widetilde{G}}\Pr_{\widetilde{G}}[\neg F] + w_E\Pr_{E}[\neg F] - w_E\Pr_{E}[\neg F] \ .$$

Since

$$\Pr_{G'|F'}[\neg F] \leq w_{\widetilde{G}}\Pr_{G'}[\neg F] + w_E\Pr_{E}[\neg F] \ ,$$

and $w_{\widetilde{G}} \leq 1 + O(\epsilon)$, we must have

$$\Pr_{G'|F'}[\neg F] \leq (2+O(\epsilon))w_E\Pr_{E}[\neg F] \ ,$$

and

$$\Pr_{G'}[\neg F] \leq (1/3+O(\epsilon))w_E\Pr_{E}[\neg F].$$

Then, we get that

$$\begin{aligned}
&\left(1-\Pr_{G'|F'}[\neg F]\right)G'|F'(x) = \left(1-\Pr_{G}{}'|F'[\neg F]\right)(G'|F'\cap\neg F)(x)\\
&= w_{\widetilde{G}}\widetilde{G}(x) + w_E\left(1-\Pr_{E}[\neg F]\right)E(x) + w_L\left(1-\Pr_{L}[\neg F]\right)L(x) - w_{\widetilde{G}}\Pr_{G'}[\neg F]\widetilde{G}(x) \ .
\end{aligned}$$

Thus, we have

$$\begin{aligned}
w'_L &= \frac{w_L\left(1-\Pr_L[\neg F]\right) - w_{\widetilde{G}}\Pr_{G'}[\neg F]}{1-\Pr_{\widetilde{P}}[\neg F]}\\
&\leq w_L + (1+O(\epsilon))\Pr_{G'}[\neg F] + O\left(\epsilon\Pr_{\widetilde{P}}[\neg F]\right)\\
&\leq w_L + (1/3+O(\epsilon))\,w_E\Pr_{E}[\neg F] + O\left(\epsilon\Pr_{G'|F'}[\neg F]\right) \ .
\end{aligned}$$

Also we have

$$\begin{aligned}
w'_E &= \frac{w_E\left(1-\Pr_E[\neg F]\right)}{1-\Pr_{G'|F'}[\neg F]}\\
&\leq w_E\left(1-\Pr_{E}[\neg F]\right) + O\left(\epsilon\Pr_{G'|F'}[\neg F]\right) \ .
\end{aligned}$$

Thus,

$$\begin{aligned}
w_L + w_E - w'_L - w'_E &\geq (2/3-O(\epsilon))\,w_E\Pr_{E}[\neg F] - O\left(\epsilon\Pr_{\widetilde{P}}[\neg F]\right)\\
&\geq (1/3-O(\epsilon))\Pr_{G'|F'}[\neg F] \geq \epsilon/Cn^{2t'} \ .
\end{aligned}$$

Note that the penultimate inequality also gives that

$$\Pr_{G'|F'}[\neg F] \leq (3+O(\epsilon))\,(w_L + w_E - w'_L - w'_E).$$

This completes the proof. □

Proposition 7.12 now follows using induction on the iterations.

7.3.2 Completing the Proof of Correctness

Lemma 7.24. $\dim(V) \leq O(\log(1/\epsilon))^k$.

Proof. After leaving the filter loop, for all $1 \leq t \leq k$, we have that $\|\widetilde{P}_t\|_F \leq \epsilon O(\log(1/\epsilon))^{t/2}$. $M(\widetilde{P}_t)$ has the same Frobenius norm, and thus the L_2-norm of its singular values, when considered as a matrix. Thus, there are at most $O(\log(1/\epsilon))^{t/2}$ singular values bigger than ϵ. So, we have that $\dim(V_t) = O(\log(1/\epsilon))^{t/2}$, and so $\dim(V) \leq \sum_{t=1}^{k} \dim V_t \leq O(\log(1/\epsilon))^k$. □

Let μ_V be the projection of μ onto the subspace V. Now we can show using our moment matching lemma that it suffices to approximate μ_V.

Lemma 7.25. *We have that* $\|\mu_V - \mu\|_2 \leq O(\epsilon)$.

Proof. Let $v = (\mu_V - \mu)/\|\mu_V - \mu\|_2$. Note that v is a unit vector perpendicular to V and we need to show that $v^T\mu \leq O(\epsilon)$. We will apply Lemma 7.1 to G'', the projection G' conditioned on the event that all the filters we produced accept F', onto v. Note that $G'|F'$ has $d_{\mathrm{TV}}(\widetilde{G}, G'|F') \leq O(\epsilon)$, and so we have that $d_{\mathrm{TV}}(G'', N(0, v^T\mu)) \leq O(\epsilon)$. We can bound the expectation of the Hermite polynomials as follows, for $1 \leq t \leq k$:

$$\begin{aligned}
|\mathbf{E}_{X\sim G''}[He_t(X)/\sqrt{t!}]| &= |\mathbf{E}_{X\sim G'}[He_t(v\cdot X)/\sqrt{t!}]| = |P^t(v,\ldots,v)| \\
&= |(v^{\otimes t-1})^T M(P^t)v| \leq \|v^{\otimes t-1}\|_2 \|M(P^t)v\|_2 \\
&\leq 1 \cdot O(\epsilon)\ .
\end{aligned}$$

On the other hand, we have $\mathbf{E}_{X\sim N(0,1)}[He_t(X)/\sqrt{t!}] = 0$, for $1 \leq t \leq k$. For $t = 0$, $He_t(X)/\sqrt{t!} = 1$, which has expectation 1 under both G'' and $N(0,1)$. We want to consider the difference in the expectations of X^t, for $1 \leq t \leq k$. We can write x^t as a linear combination of Hermite polynomials, $x^t = \sum_{i=0}^{t} a_i He_i(x)/\sqrt{i}$. Using the orthonormality of these polynomials, we have that $\mathbf{E}_{X\sim N(0,1)}[(X^t)^2] = \sum_i a_i^2$. On the other hand, by standard results, $\mathbf{E}_{X\sim N(0,1)}[(X^t)^2] = 2^t t!$. Thus, we have:

$$\begin{aligned}
|\mathbf{E}_{X\sim G''}[X^t] - \mathbf{E}_{X\sim N(0,1)}[X^t]| &= \left|\sum_{i=0}^{t} a_i \left(\mathbf{E}_{X\sim G''}[He_i(X)/\sqrt{i!}] - \mathbf{E}_{X\sim N(0,1)}[He_i(X)/\sqrt{i!}]\right)\right| \\
&\leq O(\epsilon)\cdot \sum_{i=0}^{t} a_i \\
&\leq O(\epsilon)\cdot \sqrt{t}\cdot\sqrt{2^t t!}
\end{aligned}$$

Note that for $t \geq 10$, $\sqrt{t2^t t!} \leq (t-1)!/t$. Thus, there is a constant $c > 0$ such that this $O(\epsilon)\cdot\sqrt{t}\cdot\sqrt{2^t t!}$ is smaller than $(t-1)!c^t\epsilon/t$, for all $1 \leq t \leq k$. Now we can apply Lemma 7.1 with $\delta = c\epsilon$ and obtain that $|v^T\mu| \leq O(\delta) = O(\epsilon)$. We need to set k to be a sufficiently high multiple of $\epsilon\sqrt{\ln(1/\epsilon)}$ to make this work. □

It remains to analyze the rest of the algorithm and show that $\widetilde{\mu}_V$ it produces is close to μ_V.

Lemma 7.26. *We can construct a set $S \subset V$ of unit vectors of size $\dim(V)^{O(\dim(V))}$ such that for any unit vector $v \in V$, there is a $v' \in S$ with $\|v - v'\|_2 \leq 1/2$, in time $\dim(V)^{O(\dim(V))}$.*

Proof. Let $\ell = \dim(V)$. We will construct such a cover for $\mathbb{R}^\ell$ and translate that to V by using the orthonormal basis for V given by the right singular vectors of $M(P_t)$ with singular values bigger than ϵ. We can divide the cube $[-1,1]^\ell$ into $\ell^{O(\ell)}$ cubes of side length $1/(2\sqrt{\ell})$. For each cube, we check if it has a corner with L_2-norm ≥ 1 and a corner with L_2-norm ≤ 1. If it does not, it does not contain any unit vectors so we can ignore it. If the cube does contain any unit vectors, then if its center is v, we add the normalized vector $v/\|v\|_2$ to S. Since there is a unit vector v' in the cube, and all vectors in the cube have $\|v'-v\|_2 \leq \sqrt{\ell}\|v'-v\|_\infty \leq 1/4$, we have $|\|v\|_2 - 1| \leq 1/4$, and so $\|v - v/\|v\|_2\|_2 \leq 1/4$. Thus, for any unit vector v' in this cube, we have $\|v' - v/\|v\|_2\|_2 \leq 1/4 + 1/4 \leq 1/2$. Since every unit v' is in some cube whose normalized center we added to S, we are done. □

Firstly, we note that in order to approximate $v^T\mu$, it is sufficient to find an x with $\Pr_{X\sim G'}[v^T X \geq x] = 1/2 + O(\epsilon)$:

Lemma 7.27. *For all $x \in \mathbb{R}$ with $|\Pr_{X\sim G'}[v^T X \geq x] - 1/2| \leq 3\epsilon$, we have that $|v^T\mu - x| \leq O(\epsilon)$.*

Proof. First note that the pdf of $\widetilde{G}$ projected onto v has $G(x - v^T\mu) \geq 1/2$ for all x with $|x - v^T\mu| \leq O(\epsilon)$. Supposing that $|x - v^T\mu| \geq 8\epsilon$, we have that $|\Pr_{X\sim\widetilde{G}}[v^T X \geq x] - 1/2| \geq 4\epsilon$, and so $|\Pr_{X\sim G'}[v^T X \geq x] - 1/2| \geq 4\epsilon - d_{\mathrm{TV}}(G', \widetilde{G}) \geq 3\epsilon$. □

To show that we can find such a point by bisection, we need to show that there is an interval of such points of reasonable length where we are looking for them:

Lemma 7.28. *Given a unit vector $v \in \mathbb{R}^n$, there is an interval $[a,b]$ such that*

- *for all $x \in [a,b]$, we have that $|\Pr_{X\sim G'}[v^T X \geq x] - 1/2| \leq 2\epsilon$,*
- *$b - a = \Theta(\epsilon)$,*
- *and $|a|, |b| \leq O(\epsilon\sqrt{\log 1/\epsilon})$.*

Proof. We can take $[a,b]$ to be the set of x with $|\Pr_{X\sim\widetilde{G}}[v^T X \geq x] - 1/2| \leq \epsilon$. This is an interval since $\Pr_{X\sim\widetilde{G}}[v^T X \geq x]$ is monotone. All x in it have $|\Pr_{X\sim G'}[v^T X \geq x] - 1/2| \leq \epsilon + d_{\mathrm{TV}}(G', \widetilde{G}) \leq 2\epsilon$. Thus, by the previous lemma, $|b - v^T\mu|, |a - v^T\mu| \leq O(\epsilon)$ and thus $|b - a| \leq O(\epsilon)$, and since $|v^T\mu| \leq O(\epsilon\sqrt{\log 1/\epsilon})$, we have that $|a|, |b| \leq O(\epsilon\sqrt{\log 1/\epsilon})$. □

Thus, we obtain:

Lemma 7.29. *Given a unit vector $v \in \mathbb{R}^n$, we can find an m_v with $|v^T\mu - m_v| \leq O(\epsilon)$ using $O(\log\log 1/\epsilon)$ statistical queries of precision $\epsilon/2$.*

Proof. We use bisection to find a point where our SQ approximation $\widetilde{p}$ to $\Pr_{X\sim G'}[v^T X \geq x]$ is within $5\epsilon/2$ of $1/2$. If we find such a point, it has $|v^T\mu - m_v| \leq O(\epsilon)$, by Lemma 7.27. Lemma 7.28 yields that there is an interval $[a,b]$ of length $O(\epsilon)$ containing such points in the interval $|x| \leq O(\epsilon\sqrt{\log(1/\epsilon)})$. Indeed, if our test point x has $\widetilde{p} > 1/2 + 5\epsilon/2$, then $x > b$ and if $\widetilde{p} < 1/2 - 5\epsilon/2$, then $x < a$. Thus, $[a,b]$ remains a subinterval of the interval we are considering. □

We now have that μ_v is a feasible point of the LP considered in Step 11. The following lemma completes the proof:

Lemma 7.30. *Any feasible point of the LP considered in Step 11, $\widetilde{\mu}_V$ has $\|\mu_V - \widetilde{\mu}_V\|_2 \leq O(\epsilon)$.*

Proof. Consider the vector $v = (\mu_V - \widetilde{\mu}_V)/\|\mu_V - \widetilde{\mu}_V\|_2$. Note that v is in V, since $\mu_C, \widetilde{\mu}_V$ are. Since v is a unit vector in V, there is a $v' \in S$ with $\|v - v'\|_2 \leq 1/2$. Since $\widetilde{\mu}_V$ is a solution to the LP, $v'^T(\mu_V - \widetilde{\mu}_V) \leq O(\epsilon)$. Thus, we have that

$$\begin{aligned}\|\mu_V - \widetilde{\mu}_V\|_2 &= v^T(\mu_V - \widetilde{\mu}_V) \\ &= v'^T(\mu_V - \widetilde{\mu}_V) + (v - v')^T(\mu_V - \widetilde{\mu}_V) \\ &\leq O(\epsilon) + \|\mu_V - \widetilde{\mu}_V\|_2/2 \,.\end{aligned}$$

Therefore, $\|\mu_V - \widetilde{\mu}_V\|_2 \leq O(\epsilon)$, as required. □

Proof of Theorem 7.7. Since the LP has a feasible point, we can find such a point $\widetilde{\mu}_V$ that has $\|\mu_V - \widetilde{\mu}_V\|_2 \leq O(\epsilon)$. By the previous lemma, we have that $\|\mu_V - \mu\|_2 \leq O(\epsilon)$. Thus, the algorithm is correct. All statistical queries are of the claimed precision. We need to get bounds on the running time and number of statistical queries.

Step 2 that uses the algorithm from [DKK+16b], takes $\mathrm{poly}(n/\epsilon)$ time and statistical queries. Finding $\widetilde{P}_t$, for each $1 \leq t \leq k$, takes n^t statistical queries, giving $n^O(k)$ time total. There are at most $O(n^{2k})$ iterations of the loop. Each iteration takes $\mathrm{poly}(n/\epsilon)$ time and statistical queries to find T, and n^t statistical queries to recompute $\widetilde{P}_t$. It suffices to compute the SVD to within Frobenius norm $1/\mathrm{poly}(n/\epsilon)$, which takes time $\mathrm{poly}(n/\epsilon)$. The set S has size $\dim(V)^{O(\dim(V))} = \log(1/\epsilon)^{kO(\log(1/\epsilon))^k} = 2^{\log(1/\epsilon)^{O(k)}}$. Computing it takes time $2^{\log(1/\epsilon)^{O(k)}}$. Approximating the medians takes $2^{\log(1/\epsilon)^{O(k)}}$ statistical queries and time. The LP has $2^{\log(1/\epsilon)^{O(k)}}$ constraints and $\log(1/\epsilon)^{O(k)}$ variables. The size of the LP is $2^{\log(1/\epsilon)^{O(k)}}$ bits, and so with a polynomial time LP solver, we can get $2^{\log(1/\epsilon)^{O(k)}}$ time.

We thus have that the total time and statistical queries are both at most

$$n^{O(k)}\mathrm{poly}(1/\epsilon) + 2^{\log(1/\epsilon)^{O(k)}} = n^{O(\sqrt{\log(1/\epsilon)})} + 2^{\log(1/\epsilon)^{O(\sqrt{\log(1/\epsilon)})}}.$$

□

Acknowledgements. I.D. would like to thank Andy Drucker for useful discussions on Question 1.1.

References

[ADLS15] J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt. Sample-optimal density estimation in nearly-linear time. *CoRR*, abs/1506.00671, 2015.

[AK01] S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *Proceedings of the 33rd Symposium on Theory of Computing*, pages 247–257, 2001.

[AM05] 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.

[BBBB72] R.E. Barlow, D.J. Bartholomew, J.M. Bremner, and H.D. Brunk. *Statistical Inference under Order Restrictions*. Wiley, New York, 1972.

[BFJ+94] A. Blum, M. Furst, J. Jackson, M. Kearns, Y. Mansour, and S. Rudich. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *Proceedings of the Twenty-Sixth Annual Symposium on Theory of Computing*, pages 253–262, 1994.

[BFR+00] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *IEEE Symposium on Foundations of Computer Science*, pages 259–269, 2000.

[BR13a] Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *COLT 2013 - The 26th Annual Conference on Learning Theory*, pages 1046–1066, 2013.

[BR13b] Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. *Ann. Statist.*, 41(4):1780–1815, 08 2013.

[BS10] M. Belkin and K. Sinha. Polynomial learning of distribution families. In *FOCS*, pages 103–112, 2010.

[BV08] 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.

[CDSS13] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Learning mixtures of structured distributions over discrete domains. In *SODA*, pages 1380–1394, 2013.

[CDSS14a] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In *STOC*, pages 604–613, 2014.

[CDSS14b] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Near-optimal density estimation in near-linear time using variable-width histograms. In *NIPS*, pages 1844–1852, 2014.

[CFJ13] 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.

[CGG02] M. Cryan, L. Goldberg, and P. Goldberg. Evolutionary trees can be learned in polynomial time in the two state general Markov model. *SIAM Journal on Computing*, 31(2):375–397, 2002.

[CGR15] M. Chen, C. Gao, and Z. Ren. Robust covariance matrix estimation via matrix depth. *CoRR*, abs/1506.00691, 2015.

[CQ10] S. X. Chen and Y. L. Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.*, 38(2):808–835, 04 2010.

[Dan16] A. Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the 48th Annual Symposium on Theory of Computing, STOC 2016*, pages 105–117, 2016.

[Das99] S. Dasgupta. Learning mixtures of Gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pages 634–644, 1999.

[DDKT16] C. Daskalakis, A. De, G. Kamath, and C. Tzamos. A size-free CLT for poisson multinomials and its applications. In *Proceedings of STOC'16*, 2016.

[DDO+13] C. Daskalakis, I. Diakonikolas, R. O'Donnell, R.A. Servedio, and L. Tan. Learning Sums of Independent Integer Random Variables. In *FOCS*, pages 217–226, 2013.

[DDS12a] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning k-modal distributions via testing. In *SODA*, pages 1371–1385, 2012.

[DDS12b] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning Poisson Binomial Distributions. In *STOC*, pages 709–728, 2012.

[DDS15] A. De, I. Diakonikolas, and R. Servedio. Learning from satisfying assignments. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, pages 478–497, 2015.

[DG85] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View*. John Wiley & Sons, 1985.

[DG92] D. L. Donoho and M. Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.*, 20(4):1803–1827, 12 1992.

[DK14] 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.

[DKK+16a] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Efficient and optimally robust learning of high-dimensional gaussians. In *Manuscript, available at Arxiv*, 2016.

[DKK+16b] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *Proceedings of FOCS'16*, 2016.

[DKS15] I. Diakonikolas, D. M. Kane, and A. Stewart. Optimal learning via the fourier transform for sums of independent integer random variables. *CoRR*, abs/1505.00662, 2015.

[DKS16a] I. Diakonikolas, D. M. Kane, and A. Stewart. The fourier transform of poisson multinomial distributions and its algorithmic applications. In *Proceedings of STOC'16*, 2016.

[DKS16b] I. Diakonikolas, D. M. Kane, and A. Stewart. Robust learning of fixed-structure bayesian networks. *CoRR*, abs/1606.07384, 2016.

[DL01] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics, Springer, 2001.

[DNS14] A. Daniely, N.Linial, and S. S.-Shwartz. From average case complexity to improper learning complexity. In *Symposium on Theory of Computing, STOC 2014*, pages 441–448, 2014.

[Fel16a] V. Feldman. A general characterization of the statistical query complexity. *CoRR*, abs/1608.02198, 2016.

[Fel16b] V. Feldman. Statistical query learning. In *Encyclopedia of Algorithms*, pages 2090–2095. 2016.

[FGKP06] V. Feldman, P. Gopalan, S. Khot, and A. Ponnuswami. New results for learning noisy parities and halfspaces. In *Proc. 47th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 563–576, 2006.

[FGR+13] V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proceedings of STOC'13*, pages 655–664, 2013.

[FGV15] V. Feldman, C. Guzman, and S. Vempala. Statistical query algorithms for stochastic convex optimization. *CoRR*, abs/1512.09170, 2015.

[FM99] Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In *Proceedings of the 12th Annual COLT*, pages 183–192, 1999.

[FOS06] 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.

[FOS08] J. Feldman, R. O'Donnell, and R. A. Servedio. Learning mixtures of product distributions over discrete domains. *SIAM J. Comput.*, 37(5):1536–1564, 2008.

[FPV15] V. Feldman, W. Perkins, and S. Vempala. On the complexity of random satisfiability problems with planted solutions. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC, 2015*, pages 77–86, 2015.

[GR00] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. Technical Report TR00-020, Electronic Colloquium in Computational Complexity, 2000.

[Hot31] H. Hotelling. The generalization of student's ratio. *Ann. Math. Statist.*, 2(3):360–378, 08 1931.

[HP15] 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.

[HR09] P.J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley New York, 2009.

[HRRS86] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics. The approach based on influence functions*. Wiley New York, 1986.

[Hub64] P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964.

[Kea98] M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.

[KK14] A. R. Klivans and P. Kothari. Embedding hard learning problems into gaussian space. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2014*, pages 793–809, 2014.

[KKMS08] A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[KMR$^+$94] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, and L. Sellie. On the learnability of discrete distributions. In *Proc. 26th STOC*, pages 273–282, 1994.

[KMV10] A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In *STOC*, pages 553–562, 2010.

[KS06] A. Klivans and A. Sherstov. Cryptographic hardness for learning intersections of halfspaces. In *Proc. 47th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 553–562, 2006.

[KSV08] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. *SIAM J. Comput.*, 38(3):1141–1156, 2008.

[KV16] R. Kannan and S. Vempala. Beyond spectral: Tight bounds for planted gaussians. *CoRR*, abs/1608.03643, 2016.

[LR05] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, 2005.

[LRV16] K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In *Proceedings of FOCS'16*, 2016.

[MR05] E. Mossel and S. Roch. Learning nonsingular phylogenies and Hidden Markov Models. In *To appear in Proceedings of the 37th Annual Symposium on Theory of Computing (STOC)*, 2005.

[MV10] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, pages 93–102, 2010.

[MW15] T. Ma and A. Wigderson. Sum-of-squares lower bounds for sparse pca. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 1612–1620, 2015.

[NP33] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

[Sco92] D.W. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley, New York, 1992.

[SD08] M. S. Srivastava and M. Du. A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, 99(3):386 – 402, 2008.

[Sil86] B. W. Silverman. *Density Estimation*. Chapman and Hall, London, 1986.

[SOAJ14] 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.

[Sze89] G. Szegö. *Orthogonal Polynomials*, volume XXIII of *American Mathematical Society Colloquium Publications*. A.M.S, Providence, 1989.

[Tuk75] J.W. Tukey. Mathematics and picturing of data. In *Proceedings of ICM*, volume 6, pages 523–531, 1975.

[Val84] L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[VW02] 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.

[WBS16] T. Wang, Q. Berthet, and R. J. Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *Ann. Statist.*, 44(5):1896–1930, 10 2016.

[Wil97] R. R. Wilcox. *Introduction to robust estimation and hypothesis testing*. Statistical modeling and decision science. Acad. Press, San Diego, Calif. [u.a.], 1997.

[ZB96] H. Saranadasa Z. Bai. Effect of high dimension: by an example of a two sample problem. *Statist. Sinica,*, 6:311–329, 1996.

[ZWJ14] Y. Zhang, M. J. Wainwright, and M. I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014*, pages 921–948, 2014.

Appendix

A Sample Complexity Upper Bound for Learning GMMs

In this section, we show that learning a k-mixture of n-dimensional Gaussians to variation distance error ϵ is easy information theoretically. In particular, we have:

Theorem A.1. *Given $\epsilon > 0$ and positive integers k and n, there exists an algorithm that, given a probability distribution $\mathbf{P}$ which is a k-mixture of n-dimensional Gaussians, takes $O(n^2k^3\log^2(k)/\epsilon^5)$ samples from $\mathbf{P}$ and with probability at least $2/3$ returns a distribution $\mathbf{Q}$ with $d_{\mathrm{TV}}(\mathbf{P},\mathbf{Q}) < \epsilon$.*

Note that the algorithm given in Theorem A.1 will not be computationally efficient.

The basic idea of Theorem A.1 will be to make many guesses as to the mixture, at least one of which is close, and then run a tournament to find the true answer. We approximate the mixture by first guessing approximations to the weights and then approximating each individual Gaussian. If we had polynomially many samples from a single part of the mixture, it would be easy to learn:

Lemma A.2 (Folklore)**.** *Let G be an n-dimensional Gaussian, and let $\delta > 0$. There exists a polynomial time algorithm that, given $O(n^2/\delta^2)$ independent samples from G, returns a probability distribution $\mathbf{P}$ so that, with probability at least $2/3$, $d_{\mathrm{TV}}(G,\mathbf{P}) < \delta$.*

Note that we can easily improve the success probability in Lemma A.3 to $1-\delta$ at the cost of multiplying the sample complexity by $\log(1/\delta)$. In particular, we have:

Corollary A.3. *Let G be an n-dimensional Gaussian, and let $\epsilon > 0$. There exists a polynomial time algorithm that given M independent samples from G returns a probability distribution $\mathbf{P}$ so that with probability at least $1-\exp(-\Omega(M\epsilon^2/d^2))$, we have $d_{\mathrm{TV}}(G,\mathbf{P}) < \epsilon$.*

Unfortunately, we cannot simply run this algorithm for each component in our mixture, since we do not know which samples come from which component. However, if we manage to correctly guess where each sample comes from this will not be an issue.

Proposition A.4. *Given $\epsilon > 0$ and positive integers k and n, there exists an algorithm that given a probability distribution $\mathbf{P}$, which is a k-mixture of n-dimensional Gaussians, and $\Theta(n^2k^3\log(k)/\epsilon^3)$ independent samples from $\mathbf{P}$, returns a set of $\exp(O(n^2k^3\log^2(k)/\epsilon^3))$ distributions $\mathbf{Q}_i$ so that with probability at least $2/3$ there exists an i so that $d_{\mathrm{TV}}(\mathbf{P}_i,\mathbf{Q}_i) < \epsilon$.*

Proof. If our algorithm is given N samples, it will return a q_i for each function $f : [N] \to [k]$. Intuitively, f encodes our guess as to which sample came from which component of the mixture. Note that there are only $\exp(N\log(k))$ many such f's.

The algorithm is quite simple. Let $s_1, s_2, \ldots, s_N$ be our samples, and let $S_i = \{s_j : f(j) = i\}$. Letting A be the algorithm from Corollary A.3 with δ taken to be $\epsilon/(10k)$, we let

$$\mathbf{Q}_f = \sum_{i=1}^{k} \left(\frac{|S_i|}{N}\right) A(S_i).$$

We claim that at least one of these works with probability $2/3$.

In particular, let $\mathbf{P}$ be the mixture $\sum_{i=1}^{k} w_i G_i$. Consider the case where f correctly guesses which part of the mixture each sample was taken from. In particular, $f(i) = j$ if and only if s_i was taken from G_j. We claim that, with probability at least $2/3$, *this* choice of f leads to $d_{\mathrm{TV}}(\mathbf{Q}_f,\mathbf{P}) < \epsilon$. We will henceforth use S_i to denote the set of samples actually taken from the i^{th} component of the mixture.

Firstly, note that by standard concentration bounds, we have that $\left|\frac{|S_i|}{N} - w_i\right| < \epsilon/(10k)$, for all i, with probability at least $9/10$.

Secondly, note that after conditioning on which samples of $\mathbf{P}$ were taken form which part, the samples themselves are independent samples from the appropriate G_i's, that with probability at least $9/10$ we have that $d_{\mathrm{TV}}(G_i, A(S_i)) < \epsilon/(10k)$ for all i with $|S_i| \gg n^2k^2\log(k)/\epsilon^2$.

We claim that if both of the above conditions hold (which happens with probability at least $2/3$) that $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}_f) < \epsilon$. Letting S be the set of indices i so that $w_i < \epsilon/(5k)$, we note that for $i \notin S$ that $|S_i| \gg n^2k^2\log(k)/\epsilon^2$. We then have that

$$\begin{aligned}
d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}_f) &= \frac{1}{2}\left|\sum_{i=1}^k w_i G_i - \sum_{i=1}^k \left(\frac{|S_i|}{N}\right) A(S_i)\right|_1 \\
&\leq \frac{1}{2}\sum_{i=1}^k \left|w_i G_i - \left(\frac{|S_i|}{N}\right) A(S_i)\right|_2 \\
&\leq \frac{1}{2}\sum_{i=1}^k \left|w_i - \left(\frac{|S_i|}{N}\right)\right| + \frac{1}{2}\sum_{i=1}^k \max\left(w_i, \left(\frac{|S_i|}{N}\right)\right)|G_i - A(S_i)|_1 \\
&\leq \epsilon/20 + \frac{1}{2}\sum_{i\in S}(w_i + \epsilon/(10k))|G_i - A(S_i)|_1 + \sum_{i\notin S} d_{\mathrm{TV}}(G_i, A(S_i)) \\
&\leq \epsilon/20 + \frac{1}{2}\sum_{i\in S} 3\epsilon/(10k) + \sum_{i\notin S}\epsilon/(10k) \\
&\leq \epsilon/20 + 3\epsilon/20 + \epsilon/20 \\
&< \epsilon\ .
\end{aligned}$$

This completes the proof. □

Theorem A.1 now follows immediately form a standard tournament argument (see, e.g. [DL01, DDS12b]).

B Sample Complexity Upper Bound for Parameter Estimation of Separated GMMs

Next we consider the more complicated task of parameter estimation. In particular, given samples from a distribution $\mathbf{P} = \sum_{i=1}^k G_i$, where each G_i is a weighted Gaussian, we would like to learn a distribution $\mathbf{Q}$ that is not only close to $\mathbf{P}$ but that can be written as $\mathbf{Q} = \sum_{i=1}^k H_i$ with $\|H_i - G_i\|_1$ small for all i. Now, in general, this task will require number of samples exponential in k, simply because there are pairs of mixtures that are $\epsilon^{\Omega(k)}$-close in variation distance and yet ϵ-far in terms of their individual components. However, we will show that if the components are separated, this cannot be the case and thus learning the distribution in variation distance will be sufficient.

Before we begin, we need to clarify our notion of separation. Given two pseudo-distributions, p and q, we define their overlap as $V(p,q) := \int \min(dp, dq)$. We should note that if p and q are honest distributions, then $d_{\mathrm{TV}}(p,q) = 1 - V(p,q)$. We have the following theorem:

Theorem B.1. *There exists a constant C so that if we have two mixtures $p = \sum_{i=1}^k G_i$ and $q = \sum_{i=1}^k H_i$, where p and q are normalized distributions with H_i, G_i weighted Gaussians, such that $d_{\mathrm{TV}}(p,q) < (\delta/k)^C$ for some sufficiently small $\delta > 0$, and so that for any $i \neq j$, $V(G_i, G_j), V(H_i, H_j) < (\delta/k)^C$, then there exists a permutation $\pi : [k] \to [k]$ so that $\|G_i - H_{\pi(i)}\|_1 < \delta$ for all i.*

We begin by producing a proxy for the overlap between distributions. In particular, for pseudo-distributions

p and q, we define

$$h(p,q) = -\log\left(\int \sqrt{dpdq}\right).$$

Notice that if p and q are true distributions, this is related to the Hellinger distance by $H(p,q) = 2(1 - e^{-h(p,q)})$. We also note the relationship to the overlap:

Lemma B.2. *If p and q are pseudo-distributions with L_1 norm at most 1, then*

$$V(p,q) = \exp(-\Theta(h(p,q)) + O(1)) .$$

Proof. On the one hand, there is an easy upper bound

$$V(p,q) = \int \min(dp, dq) \leq \int \sqrt{dpdq} = \exp(-h(p,q)).$$

The lower bound is by Cauchy-Schwarz

$$\exp(-h(p,q)) = \int \sqrt{dpdq} \leq \left(\int \min(dp,dq)\right)^{1/2} \left(\int \max(dp,dq)\right)^{1/2} \leq \sqrt{2V(p,q)}.$$

This completes our proof. □

Ideally we would like to show that h is nearly a metric for Gaussians. Namely that $h(A,C) = O(h(A,B) + h(B,C))$. This would imply that H_i could not have large overlap with more than one G_j, since if $V(H_i, G_a)$ and $V(H_i, G_b)$ were both large, then $h(H_i, G_a), h(H_i, G_b)$ would be small and therefore, $h(G_a, G_b)$ would be small. This would contradict our assumption that G_a and G_b have small overlap. Unfortunately, this is not true. In one dimension, a very wide Gaussian may have non-trivial overlap with two narrow Gaussians with widely separated means, neither of which overlaps the other substantially. We will need to develop techniques to deal with this circumstance.

To do this, we introduce an intermediate notation. If $G_i = w_i N(\mu_i, \Sigma_i)$ are weighted Gaussians, we define

$$h_\Sigma(G_1, G_2) := h(N(0,\Sigma_1), N(0,\Sigma_2)).$$

This is useful because it does satisfy an approximate triangle inequality.

Proposition B.3. *For F, G, H weighted Gaussians, we have that*

$$h_\Sigma(F,H) = O(h_\Sigma(F,G) + h_\Sigma(G,H)).$$

Before we prove this, we will first need to find an approximation to h_Σ.

Lemma B.4. *If G and H are weighted Gaussians with covariance matrices A and B respectively, then*

$$h_\Sigma(G,H) = \Theta\left(\Sigma_{\lambda \text{ eigenvalue of } B^{-1/2}AB^{-1/2}} \min(|\log(\lambda)|, |\log(\lambda)|^2)\right).$$

Proof. By making an appropriate change of variables, we can assume that H has identity covariance and G has covariance $B^{-1/2}AB^{-1/2}$. Thus, it suffices to consider the case where $B = I$. In this case, we may

diagonalize A to get $A = \mathrm{diag}(\lambda_i)$. We then have that

$$\begin{aligned} h_\Sigma(G,H) &= h(N(0,A),N(0,I)) \\ &= -\log\left((2\pi)^{-n/2}\prod_{i=1}^n \lambda_i^{-1/4}\int \exp\left(-\sum_{i=1}^n x_i^2/2(1/(2\lambda_i)+1/2)\right)dx\right) \\ &= -\log\left(\prod_{i=1}^n \lambda_i^{-1/4}((1+\lambda_i^{-1})/2)^{-1/2}\right) \\ &= \sum_{i=1}^n \log(\lambda_i)/4 + \log((1+\lambda_i^{-1})/2)/2. \end{aligned}$$

We claim that

$$\log(\lambda)/4 + \log((1+\lambda^{-1})/2)/2 = \Theta(\min(|\log(\lambda)|, |\log(\lambda)|^2)).$$

To see this note that when $\lambda = 1+\epsilon$ for small values of ϵ, the left hand side above is

$$(\epsilon/4 - \epsilon^2/8 + O(\epsilon^3)) + (-\epsilon/4 + 3\epsilon/16 + O(\epsilon^3)) = \epsilon^2/16 + O(\epsilon^3) = \Theta(\epsilon^2).$$

On the other hand, when $\lambda \gg 1$, this is asymptotic to $|\log(\lambda)|/4$, and when $\lambda \ll 1$, it is similarly asymptotic to $-\log(\lambda)/4$. Finally, since it is easily verified that $\log(\lambda)/4 + \log((1+\lambda^{-1})/2)/2$ is never 0 unless $\lambda = 1$, this proves the claim, from which our lemma follows easily. □

We will also need the following fact about eigenvalues of a product of matrices:

Lemma B.5. *Let A and B be symmetric matrices with eigenvalues $\nu_1 \geq \nu_2 \geq \ldots \geq \nu_n > 0$ and $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n > 0$, respectively. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{2n} > 0$ be the sorting of the ν_i and μ_i together. Let M be a matrix with $M^T M = A$. Then, the k^{th} largest eigenvalue of $M^T B M$ is at most λ_k^2.*

Proof. We need to show that there is an $(n-k+1)$-dimensional subspace V so that for $v \in V$ we have that $vA^{1/2}BA^{1/2}v \leq \lambda_k^2|v|^2$. Suppose that $\lambda_1, \ldots, \lambda_{k-1}$ contains m of the ν_i and $k-m-1$ of the μ_i. Let V be the subspace of vectors v so that v is perpendicular to the top m eigenvectors of A and so that $A^{1/2}v$ is perpendicular to the top $m-k-1$ eigenvalues of B. Then

$$vM^T BMv \leq \lambda_k |Mv|^2 = \lambda_k vAv \leq \lambda_k^2 |v|^2.$$

This completes the proof. □

We are now ready to prove Proposition B.3.

Proof. Let F, G, H have covariance matrices A, B, C respectively. Let $\Sigma_1 = A^{-1/2}BA^{-1/2}$, $\Sigma_2 = B^{-1/2}CB^{-1/2}$ and $\Sigma_3 = A^{-1/2}CA^{-1/2} = (A^{-1/2}B^{1/2})\Sigma_2(B^{1/2}A^{-1/2})$. Let the eigenvalues of Σ_i be $\lambda_1^{(i)} \geq \lambda_2^{(i)} \geq \ldots \geq \lambda_n^{(i)} > 0$. Let $f(x) = \max(0, \min(\log(x), \log^2(x)))$. We have by Lemma B.4 that

$$h_\Sigma(F,G) = \Theta\left(\sum_{i=1}^n f(\lambda_i^{(1)}) + f(1/\lambda_i^{(1)})\right),$$

$$h_\Sigma(G,H) = \Theta\left(\sum_{i=1}^n f(\lambda_i^{(2)}) + f(1/\lambda_i^{(2)})\right),$$

$$h_\Sigma(F,H) = \Theta\left(\sum_{i=1}^n f(\lambda_i^{(3)}) + f(1/\lambda_i^{(3)})\right).$$

On the other hand, Lemma B.5 says that $\lambda_i^{(3)}$ is at most the square of the i^{th} largest of the $\lambda_j^{(1)}$ and $\lambda_j^{(2)}$. Therefore,

$$\sum_{i=1}^{n} f(\lambda_i^{(3)}) = O\left(\sum_{i=1}^{n} f(\lambda_i^{(1)}) + \sum_{i=1}^{n} f(\lambda_i^{(2)})\right).$$

Similarly, by considering the inverses of these matrices, we find that

$$\sum_{i=1}^{n} f(1/\lambda_i^{(3)}) = O\left(\sum_{i=1}^{n} f(1/\lambda_i^{(1)}) + \sum_{i=1}^{n} f(1/\lambda_i^{(2)})\right).$$

Together these complete the proof. □

In addition to this, we need to know what else contributes to $h(G, H)$. We define

$$h_\mu(G, H) = h(G, H) - h_\Sigma(G, H).$$

We make the following claim:

Proposition B.6.

$$h_\mu(w_1 N(\mu_1, \Sigma_1), w_2 N(\mu_2, \Sigma_2)) = -1/2 \log(w_1 w_2) + \inf_x((x-\mu_1)\Sigma_1^{-1}(x-\mu_1) + (x-\mu_2)\Sigma_2^{-1}(x-\mu_2))/4.$$

Proof. We have that

$$\begin{aligned} &h(w_1 N(\mu_1, \Sigma_1), w_2 N(\mu_2, \Sigma_2)) \\ &= -1/2 \log(w_1 w_2) - \log\left(\int (2\pi)^{-n/2} (\det(\Sigma_1 \Sigma_2))^{-1/4} \exp(-((x-\mu_1)\Sigma_1^{-1}(x-\mu_1) + (x-\mu_2)\Sigma_2^{-1}(x-\mu_2))/4) dx\right). \end{aligned}$$

Letting x_0 achieve the minimum value of $((x-\mu_1)\Sigma_1^{-1}(x-\mu_1) + (x-\mu_2)\Sigma_2^{-1}(x-\mu_2))/4$, this is

$$\begin{aligned} &-1/2 \log(w_1 w_2) + ((x_0-\mu_1)\Sigma_1^{-1}(x_0-\mu_1) + (x_0-\mu_2)\Sigma_2^{-1}(x_0-\mu_2))/4 \\ &\qquad + \log\left(\int (2\pi)^{-n/2} (\det(\Sigma_1 \Sigma_2))^{-1/4} \exp(-(x-x_0)(\Sigma_1^{-1} + \Sigma_2^{-1})(x-x_0)/4) dx\right). \end{aligned}$$

Noting that the term at the end is simply $h_\Sigma(w_1 N(\mu_1, \Sigma_1), w_2 N(\mu_2, \Sigma_2))$ completes the proof. □

We need one further proposition from which Theorem B.1 will follow easily.

Proposition B.7. *Under the assumptions of Theorem B.1, for each i there exists at most one j so that $h(G_i, H_j) > (\delta/k)^{\sqrt{C}}$.*

To prove this, we will need one further lemma:

Lemma B.8. *If $h(G_i, H_j) > (\delta/k)^{\sqrt{C}}$, with Σ_G and Σ_H the covariance matrices of the corresponding Gaussians, then for A a sufficiently large constant (independent of C) $\Sigma_G \leq A\Sigma_H$.*

Proof. Suppose for sake of contradiction that this is not the case. By making a change of variables, we can assume that $\Sigma_G = I$. This means that Σ_H has some eigenvector v with eigenvalue less than $1/A$. Let H' be H_j translated by $C^{1/4}\sqrt{\log(k/\delta)}$ in the direction closer to the mean of G_i. We have that

$$h(H_j, H') = h_\mu(H_j, H') = \Theta(A\sqrt{C}\log(k/\delta)).$$

Therefore, $V(H_j, H') = (\delta/k)^{\Omega(A\sqrt{C})}$. On the other hand,

$$\begin{aligned} h(G_i, H') &= h_\Sigma(G_i, H') + h_\mu(G_i, H') \\ &\leq h_\Sigma(G_i, H) + O(\sqrt{C}\log(k/\delta)) \\ &\leq h(G_i, H) + O(\sqrt{C}\log(k/\delta)) \\ &= O(\sqrt{C}\log(k/\delta)). \end{aligned}$$

This means that $V(G, H') = (\delta/k)^{O(\sqrt{C})}$.

This means that $p = \sum_\ell G_\ell$ has $V(p, H') = (\delta/k)^{O(\sqrt{C})}$, and since p is close to $q = \sum_\ell H_\ell$, there must be some ℓ so that $V(H', H_\ell) > (\delta/k)^{O(\sqrt{C})}$. Note that ℓ here cannot be j. On the other hand, this implies that

$$\begin{aligned} h(H_j, H_\ell) &= h_\Sigma(H_j, H_\ell) + h_\mu(H_j, H_\ell) \\ &\leq h_\Sigma(H', H_\ell) + O(h_\mu(H', H_\ell) + A\sqrt{C}\log(k/\delta)) \\ &\leq O(h(H', H_\ell) + A\sqrt{C}\log(k/\delta)) \\ &= O(A\sqrt{C}\log(k/\delta)). \end{aligned}$$

Therefore, $V(H_\ell, H_j) = (\delta/k)^{O(A\sqrt{C})}$, which for $C \gg A^2$ contradicts our assumptions. This completes the proof. □

We are now prepared to prove Proposition B.7.

Proof. Suppose for sake of contradiction that $V(G_i, H_j), V(G_i, H_\ell) > (\delta/k)^{\sqrt{C}}$ for some $j \neq \ell$. Then, by Lemma B.8, we have that all of the covariance matrices of G_i, H_j, H_ℓ are comparable to each other (namely each is no more than a constant multiple of any other). We claim that this implies that $h_\mu(H_j, H_\ell) = O(h_\mu(H_j, G_i) + h_\mu(H_\ell, G_i) + \sqrt{C}\log(k/\delta))$. This is because, letting Σ_G be the covariance matrix of G_i, and letting $w_G = |G_i|_1, w_H = |H_j|_1, w'_H = |H_\ell|_1$, we have the following: First, each of $w_G, w_H, w'_H = \exp(O(\sqrt{C}\log(\delta/k)))$, because each distribution has large overlap with some other. Next, we have that

$$\begin{aligned} h_\mu(G_i, H_j) &= O(\sqrt{C}\log(\delta/k)) + \inf_x \Theta((x - \mu_{G_i})\Sigma_G(x - \mu_{G_i}) + (x - \mu_{H_j})\Sigma_G(x - \mu_{H_h}) \\ &= O(\sqrt{C}\log(\delta/k)) + \Theta((\mu_{G_i} - \mu_{H_j})\Sigma_G(\mu_{G_i} - \mu_{H_j})). \end{aligned}$$

Similarly,

$$h_\mu(G_i, H_\ell) = O(\sqrt{C}\log(\delta/k)) + \Theta((\mu_{G_i} - \mu_{H_\ell})\Sigma_G(\mu_{G_i} - \mu_{H_\ell})),$$

and

$$h_\mu(H_j, H_\ell) = O(\sqrt{C}\log(\delta/k)) + \Theta((\mu_{H_j} - \mu_{H_\ell})\Sigma_G(\mu_{H_j} - \mu_{H_\ell})).$$

This implies that

$$h_\mu(H_j, H_\ell) = O(h_\mu(H_j, G) + h_\mu(G, H_\ell)).$$

Therefore, we have that

$$\begin{aligned} h(H_j, H_\ell) &= h_\Sigma(H_j, H_\ell) + h_\mu(H_j, H_\ell) \\ &= O(h_\Sigma(H_j, G) + h_\mu(H_j, G) + h_\Sigma(G, H_\ell) + h_\mu(G, H_\ell)) \\ &= O(h(H_j, G) + h(G, H_\ell)) \\ &= O(\log(1/V(H_j, G)) + \log(1/V(G, H_\ell))) \\ &= O(\sqrt{C}\log(k/\delta)). \end{aligned}$$

However, this implies that $V(H_j, H_\ell) = (\delta/k)^{O(\sqrt{C})}$, a contradiction.

This completes our proof. □

We are now ready to prove Theorem B.1

Proof. For each G_i that has overlap more than $(\delta/k)^{\sqrt{C}}$ with some H_j, let $\pi(i)$ be that j. For other i, define $\pi(i)$ arbitrarily subject to π being a permutation.

Note that $V(G_i, H_j) < (\delta/k)^{\sqrt{C}}$ for any $j \neq \pi(i)$. Also note that

$$V(G_i, q) \geq V(G_i, p) - |p - q|_1 = |G_i|_1 - 2(\delta/k)^C.$$

On the other hand,

$$\begin{aligned} V(G_i, q) &\leq \sum_j V(G_i, H_j) \\ &\leq V(G_i, H_{\pi(i)}) + \sum_{j \neq \pi(i)} V(G_i, H_j) \\ &\leq V(G_i, H_{\pi(i)}) + \delta/3. \end{aligned}$$

Therefore $V(G_i, H_{\pi(i)}) \geq |G_i|_1 - \delta/2$. It is also at most $|G_i|_1 - \delta/2$. On the other hand $|G_i - H_{\pi(i)}|_1 = |G_i|_1 + |H_{\pi(i)}|_1 - 2V(G_i, H_{\pi(i)}) \leq \delta$. This completes the proof. □

C Testing the Mean of a High-Dimensional Gaussian

Theorem C.1. *There exists an algorithm that given $k = O(\sqrt{n}/\epsilon^2)$ samples from an n-dimensional Gaussian $G = N(\mu, I)$ distinguishes between the cases*

- $\mu = 0$
- $\|\mu\|_2 > \epsilon$

with probability at least $2/3$.

Proof. The tester is fairly simple. Let X_i be the i^{th} sample, and let

$$Z := \frac{1}{\sqrt{k}} \sum_{i=1}^{n} X_i.$$

The algorithm returns "YES" if $|Z|_2^2 < \epsilon^2 k/2 + n$ and "NO" otherwise.

To show correctness, note that Z is distributed as $N(\mu\sqrt{k}, I)$. If $\mu = 0$, then $|Z|_2^2$ has mean n and variance $O(n)$, and so is less than $n + \epsilon^2 k/2$ with probability at least $2/3$, assuming that k is a sufficiently large multiple of $\sqrt{n}/\epsilon^2$. On the other hand, if $\|\mu\|_2 > \epsilon$, we note that $|Z|_2^2$ has mean $n + k\|\mu\|_2^2$ and variance $O(n) + O(\sqrt{k}\|\mu\|_2)$. Thus, if $k\|\mu\|_2^2 \gg \sqrt{n}$, the algorithm rejects with probability $2/3$. Again, this happens if $\|\mu\|_2^2 > \epsilon$ and k is a sufficiently large multiple of $\sqrt{n}/\epsilon^2$. □

We also note that this tester can be implemented in the SQ model simply by verifying that each coordinate-wise median has absolute value less than $\epsilon/\sqrt{n}$, which can be verified by showing that $\Pr(x_i > 0) = 1/2 + O(\epsilon/\sqrt{n})$.

We also show that the tester above is sample-optimal, up to a constant factor:

Theorem C.2. *There is no algorithm that given $k = o(\sqrt{n}/\epsilon^2)$ samples from an n-dimensional Gaussian $G = N(\mu, I)$ distinguishes between the cases*

- $\mu = 0$
- $\|\mu\|_2 > \epsilon$

with probability at least $2/3$.

Proof. Suppose for sake of contradiction that such an algorithm does exist. Consider the following scenario: Let μ be taken from the distribution $N(0, (2\epsilon/\sqrt{n})I)$. Note that $\|\mu\|_2 > \epsilon$ with probability at least $9/10$. Let $Y_1, Y_2, \ldots, Y_k$ be independent samples taken from $N(\mu, I)$. And let $Z_1, \ldots, Z_k$ be independent samples from $N(0, I)$. Assuming that our algorithm exists, it can distinguish between a sample from $Y_1, \ldots, Y_k$ and a sample from $Z_1, \ldots, Z_k$ with probability better than $1/2$. This means that these distributions must have constant variational distance. However, note that the vector $(Z_1, \ldots, Z_k)$ is simply a standard nk-dimensional Gaussian. The vector $(Y_1, \ldots, Y_k)$ on the other hand is an nk-dimensional Gaussian with mean 0 and with

$$\mathrm{Cov}(Y_{ab}, Y_{cd}) = \begin{cases} 1 + 2\epsilon^2/n & \text{if } ab = cd \\ 2\epsilon^2/n & \text{if } b = d \text{ and } a \neq c \\ 0 & \text{otherwise.} \end{cases}$$

By standard results, $G' = N(0, \Sigma)$ has constant variational distance from $N(0, I)$ if and only if $\|\Sigma - I\|_F \gg 1$. Taking Σ to be the covariance matrix for the Y's, we have that

$$\|\Sigma - I\|_F^2 = nk(2\epsilon/\sqrt{n})^2 = 4k\epsilon^2/n = o(1) \ .$$

This implies that the distribution on Y's is close, in total variation distance, to the distribution on Z's, and gives a contradiction. □

ECCC ISSN 1433-8092
http://eccc.hpi-web.de