58th Annual IEEE Symposium on Foundations of Computer Science

Statistical Query Lower Bounds for Robust Estimation of High-Dimensional Gaussians and Gaussian Mixtures (Extended Abstract)

Ilias Diakonikolas
CS
University of Southern California
Los Angeles, CA, USA
diakonik@usc.edu

Daniel M. Kane
CSE & Math
University of California San Diego
La Jolla, CA, USA
dakane@cs.ucsd.edu

Alistair Stewart
CS
University of Southern California
Los Angeles, CA, USA
alistais@usc.edu

***Abstract*—We describe a general technique that yields the first *Statistical Query lower bounds* for a range of fundamental high-dimensional learning problems involving Gaussian distributions. Our main results are for the problems of (1) learning Gaussian mixture models (GMMs), and (2) robust (agnostic) learning of a single unknown Gaussian distribution. For each of these problems, we show a *super-polynomial gap* between the (information-theoretic) sample complexity and the computational complexity of *any* Statistical Query algorithm for the problem. 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: a wide range of known algorithmic techniques in machine learning are known to be implementable using SQs. Moreover, for the unsupervised learning problems studied in this paper, all known algorithms with non-trivial performance guarantees are SQ or are easily implementable using SQs.**

Our SQ lower bound for Problem (1) is qualitatively matched by known learning algorithms for GMMs. 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 [29] is essentially best possible among all polynomial-time SQ algorithms. On the positive side, we also give a new (SQ) learning algorithm for Problem (2) achieving the information-theoretically optimal accuracy, up to a constant factor, whose running time essentially matches our lower bound. Our algorithm relies on a filtering technique generalizing [29] that removes outliers based on higher-order tensors.

Our SQ lower bounds are attained via a unified moment-matching technique that is useful in other contexts and may be of broader interest. Our technique yields nearly-tight lower bounds for a number of related unsupervised estimation problems. Specifically, for the problems of (3) robust covariance estimation in spectral norm, and (4) robust sparse mean estimation, we establish a quadratic *statistical–computational tradeoff* for SQ algorithms, matching known upper bounds. Finally, our technique can be used to obtain tight sample complexity lower bounds for high-dimensional *testing* problems. Specifically, for the classical problem of robustly *testing* an unknown mean (known covariance) Gaussian, our technique implies an information-theoretic sample lower bound that scales *linearly* in the dimension. Our sample lower bound matches the sample complexity of the corresponding robust *learning* problem and separates the sample complexity of robust testing from standard (non-robust) testing. This separation is surprising because such a gap does not exist for the corresponding learning problem.

***Keywords*-unsupervised learning; statistical learning, statistical queries; robust algorithm**

I. Introduction

A. 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: a wide range of known algorithmic techniques in machine learning are known to be implementable using SQs. These include spectral techniques, moment and tensor methods, local search (e.g., Expectation Maximization), and many others (see, e.g., [18], [41] for a detailed discussion). Moreover, for the unsupervised learning problems studied in this paper, all known algorithms with non-trivial performance guarantees are SQ or are easily implementable using SQs.

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/bracketing 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 computational complexity of *any* SQ algorithm that solves a given learning problem. Given the ubiquity and generality of

0272-5428/17 $31.00 © 2017 IEEE
DOI 10.1109/FOCS.2017.16

SQ algorithms, an SQ lower bound provides strong evidence of the problem's computational intractability.

In this paper, we describe a general technique that yields the first *Statistical Query lower bounds* for a range of fundamental high-dimensional learning problems involving Gaussian distributions. Such problems are ubiquitous in applications across the data sciences and have been intensely investigated by different communities of researchers for several decades. Our main results are for the problems of (1) learning Gaussian mixture models (GMMs), and (2) robust (agnostic) learning of a single unknown Gaussian distribution. In particular, we show a *super-polynomial gap* between the (information-theoretic) sample complexity and the computational complexity of *any* Statistical Query algorithm for these problems. In more detail, 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). For Problem (2), we give a new (SQ) algorithm in this paper whose running time nearly matches our SQ lower bound.

Our SQ lower bounds are attained via a unified moment-matching technique that is useful in other contexts and may be of broader interest. Our technique yields nearly-tight lower bounds for a number of related unsupervised estimation problems. Specifically, for the problems of (3) robust covariance estimation in spectral norm, and (4) robust sparse mean estimation, we establish a quadratic *statistical–computational tradeoff* for SQ algorithms, matching known upper bounds.

Finally, we use our technique to obtain tight sample complexity lower bounds for high-dimensional *testing* problems. Specifically, for the classical problem of robustly *testing* an unknown mean (known covariance) Gaussian, our technique implies an information-theoretic lower bound that scales *linearly* in the dimension. This lower bound matches the sample complexity of the corresponding robust learning problem and separates the sample complexity of robust testing from standard (non-robust) testing. This separation is surprising because such a gap 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 [56] in the context of supervised learning as a natural restriction of the PAC model [71]. Subsequently, the SQ model has been extensively studied in a plethora of contexts (see, e.g., [39] and references therein).

A recent line of work [41], [43], [42], [38] developed a framework of SQ algorithms for 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 computational complexity of SQ algorithms via the notion of *Statistical Query dimension*. This complexity measure was introduced in [10] for PAC learning of Boolean functions and was recently generalized to the unsupervised setting [41], [38]. A lower bound on the SQ dimension of a learning problem provides an unconditional lower bound on the 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 [21], [4], [72], [2], [54], [11] provides computationally efficient algorithms for recovering the parameters of a GMM under separability assumptions. Subsequently, efficient parameter learning algorithms have been obtained [65], [6], [46] under minimal information-theoretic separation assumptions. The related problems of density estimation and proper learning have also been extensively studied [36], [69], [26], [65], [46], [63]. 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 any assumptions. (In the full version, we describe a simple SQ algorithm for this learning problem with sample complexity $\mathrm{poly}(n, k, 1/\epsilon)$, albeit exponential running time). Given that there is no information-theoretic barrier for learnability in this setting, the following question arises: *Is there a* $\mathrm{poly}(n, k, 1/\epsilon)$ time *algorithm for density estimation (or proper learning) of n-dimensional k-GMMs?* This question has been raised as an open problem in a number of settings (see, e.g., [64], [28] and references therein).

For parameter learning, the situation is somewhat subtle: In full generality, the sample complexity is of the form $\mathrm{poly}(n) \cdot (1/\gamma)^{\Omega(k)}$, where the parameter $\gamma > 0$ quantifies the "separation" between the components. Even in one-dimension, a sample complexity lower bound of $(1/\gamma)^{\Omega(k)}$ is known [65], [46][1]. The corresponding "hard" instances [65], [46] consist of GMMs whose components have large overlap, so many samples are required to distinguish between them. *Is this the only obstacle towards a* $\mathrm{poly}(n, k)$ *time*

[1]To circumvent the information-theoretic bottleneck of parameter learning, a related line of work has studied parameter learning in a smoothed setting [48], [9], [3], [44].

parameter learning algorithm? Specifically, suppose that we are given an instance of the problem with the additional promise that the components are "nearly non-overlapping" – so that $\mathrm{poly}(n,k)$ samples suffice for the parameter learning problem as well. (In the full version, we show that when the total variation distance between any pair of components in the given mixture is close to 1, parameter learning reduces to proper learning; hence, there is a $\mathrm{poly}(n,k)$-sample parameter learning (SQ) algorithm that runs in exponential time.) Is there a $\mathrm{poly}(n,k)$ *time* parameter learning algorithm for such instances?

In summary, the sample complexity of both versions of the learning problem is $\mathrm{poly}(n)f(k)$. On the other hand, the running time of all known algorithms for either version scales as $n^{g(k)}$, where $g(k) \geq k$. This runtime is super-polynomial in the sample complexity of the problem for super-constant values of k and is tight for these algorithms, even for GMMs with almost non-overlapping components. The preceding discussion is summarized in the following:

Question I.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 and realistic setting of *robust* (or agnostic) learning – when our assumption about the model is *approximately* true – turns out to be significantly more challenging. Specifically, until recently, even the most 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 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 [29] 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. Note that the minimum variation distance, $d_{\mathrm{TV}}(H,G)$, information-theoretically achievable under these assumptions is $\Theta(\epsilon)$, and we would like to obtain a polynomial-time algorithm with this guarantee.

Agnostically learning a single high-dimensional Gaussian is arguably *the* prototypical problem in robust statistics [50], [45], [49]. Early work in this field [70], [34] studied the sample complexity of robust estimation. Specifically, for the case of an unknown mean and known covariance Gaussian, the Tukey median [70] achieves $O(\epsilon)$-error with $O(n/\epsilon^2)$ samples. Since $\Omega(n/\epsilon^2)$ samples are information-theoretically necessary – even without noise – the robustness requirement does not change the sample complexity of the 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 [29], [60] made a first step in designing robust polynomial-time estimators for this problem. The results of [29] apply in the standard agnostic model; [60] works in a weaker model – known as Huber's contamination model [50] – where the noisy distribution D is of the form $(1-\epsilon)G + \epsilon N$, where N is an unknown "noise" distribution. For the problem of robustly estimating an unknown mean Gaussian $N(\mu, I)$, [60] obtains an error guarantee of $O(\epsilon\sqrt{\log n})$, while [29] obtains error $O(\epsilon\sqrt{\log(1/\epsilon)})$, independent of the dimension[2].

A natural and important open problem, put forth by these works [29], [60], is the following:

Question I.2. *Is there a $\mathrm{poly}(n/\epsilon)$- time agnostic learning algorithm, with error $O(\epsilon)$, for an n-dimensional Gaussian?*

Statistical–Computational Tradeoffs: A statistical–computational tradeoff refers to the phenomenon that there is an inherent gap between the information-theoretic sample complexity of a learning problem and its computational sample complexity, i.e, the minimum sample complexity attainable by any polynomial time algorithm for the problem. The prototypical example is the estimation of a covariance matrix under sparsity constraints (sparse PCA) [51], [13], [12], where a nearly-quadratic gap between information-theoretic and computational sample complexity has been established (see [8], [74]) – assuming the computational hardness of the planted clique problem.

For a number of high-dimensional learning problems (including the problem of robustly learning a Gaussian under the total variation distance), it is known that the robustness requirement does not change the information-theoretic sample complexity of the problem. On the other hand, it is an intriguing possibility that injecting noise into a high-dimensional learning problem may change its computational sample complexity.

Question I.3. *Does robustness create inherent statistical–computational tradeoffs for natural high-dimensional estimation problems?*

In this work, we consider two natural instantiations of the

[2]The algorithm of [60] can be extended to work in the standard agnostic model at the expense of an increased error guarantee of $O(\epsilon\sqrt{\log n \log(1/\epsilon)})$.

above general question: (i) robust estimation of the covariance matrix in spectral norm, and (ii) robust sparse mean estimation. We give basic background for these problems in the following paragraphs.

For (i), suppose we have sample access to a (zero-mean) n-dimensional unknown-covariance Gaussian, and we want to estimate the covariance matrix *with respect to the spectral norm*. It is known (see, e.g., [73]) that $O(n/\epsilon^2)$ samples suffice so that the empirical covariance is within spectral error at most ϵ from the true covariance; and this bound is information-theoretically optimal, to constant factors, for any estimator. For simplicity, let us assume that the desired accuracy is a small positive constant, e.g., $\epsilon = 1/10$. Now suppose that we observe samples from a corrupted Gaussian in Huber's contamination model (the weaker adversarial model) where the noise rate $\delta \ll 1/10$. First, it is not hard to see that the injection of noise does not change the information-theoretic sample complexity of the problem: there exist (computationally inefficient) robust estimators (see, e.g., [17]) that use $O(n)$ samples. (There is a straightforward SQ algorithm for this problem as well that uses $O(n)$ samples, but again runs in exponential time.) On the other hand, if we are willing to use $\tilde{O}(n^2)$ samples, a polynomial-time robust estimator with constant spectral error guarantee is known [29], [30][3]. The immediate question that follows is this:

> *Is there a computationally efficient robust covariance estimator in spectral error that uses a strongly sub-quadratic sample size, i.e., $O(n^{2-c})$ for a constant $0 < c < 1$?*

For (ii), suppose we want to estimate the mean $\mu \in \mathbb{R}^n$ of an identity covariance Gaussian up to ℓ_2-distance ϵ, under the additional promise that μ is k-sparse, and suppose that $k \ll n^{1/2}$. It is well-known that the information-theoretic sample complexity of this problem is $O(k \log n/\epsilon^2)$, and the truncated empirical mean achieves the optimal bound. For simplicity, let us assume that $\epsilon = 1/10$. Now suppose that we observe samples from a corrupted sparse mean Gaussian (in Huber's contamination model), where the noise rate $\delta \ll 1/10$. As in the setting of the previous paragraph, the injection of noise does not change the information-theoretic sample complexity of the problem: there exist a (computationally inefficient) robust SQ algorithm for this problem (see [62]) that use $O(k \log n)$ samples. Two recent works [62], [35] gave polynomial time robust algorithms for robust sparse mean estimation with sample complexity $\tilde{O}(k^2 \log n)$. In summary, in the absence of robustness, the information-theoretically optimal sample bound is known to be achievable by a computationally efficient algorithm. In contrast, in the presence of robustness, there is a quadratic gap between the information-theoretic optimum and the sample complexity of known polynomial-time algorithms. The immediate question is whether this gap is inherent:

> *Is there a computationally efficient robust k-sparse mean estimator that uses a strongly sub-quadratic sample size , i.e., $O(k^{2-c})$ for a constant $0 < c < 1$?*

It is conjectured in [62] that a quadratic gap is in fact inherent for efficient algorithms.

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* [61]: Given samples from a distribution in a given family, we want to distinguish between a null hypothesis and an alternative hypothesis. Starting with [5], 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 has been to obtain testers with sub-learning sample complexity.

In this paper, we study 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 [47], which is unfortunately not defined when the sample size is smaller than the dimension [77]. More recently, testers that succeed in the sub-linear regime have been developed [68]. In the full version, 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 D on $\mathbb{R}^n$ which is promised to be either (a) a standard Gaussian $N(0, I)$, or (b) a δ-noisy version of $N(\mu, I)$, where $\mu \in \mathbb{R}^n$ satisfies $\|\mu\|_2 \geq \epsilon$, and the noise rate δ satisfies $\delta \ll \epsilon$. The *robust* hypothesis testing problem is to distinguish, with high constant probability, 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.

[3]We note that the robust covariance estimators of [29], [30] provide error guarantees under the Mahalanobis distance, which is stronger than the spectral norm. Under the stronger metric, $\Omega(n^2)$ samples are information-theoretically required even without noise.

Robust hypothesis testing is of fundamental importance and has been extensively studied in robust statistics [49], [45], [76]. 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 natural testers fail 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. The following question arises:

Question I.4. *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 D on $\mathbb{R}^n$, we want to distinguish between the case that $D = N(0, I)$, or D is a 2-mixture of identity covariance Gaussians. This is a natural high-dimensional testing problem that we believe merits investigation in its own right. The obvious open question here is whether there exists a tester for this problem with *sub-learning* sample complexity.

B. Our Results

The main contribution of this paper is a general technique to prove lower bounds for a range of high-dimensional estimation problems involving Gaussian distributions. We use analytic and probabilistic ideas to construct explicit families of hard instances for the estimation problems described in Section I-A. Using our technique, we prove super-polynomial Statistical Query (SQ) lower bounds that answer Questions I.1 and I.2 in the negative for the class of SQ algorithms. We also show that the observed quadratic statistical–computational gap for robust sparse mean estimation and robust spectral covariance estimation is inherent for SQ algorithms. 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 I.4 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.

SQ Lower Bound for Learning GMMs: 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:

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

Theorem I.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 $\Theta(k \cdot \log n)$ (the standard cover-based algorithm that achieves this sample upper bound is SQ). In contrast, any SQ learning algorithm for this family of instances requires runtime at least $n^{\Omega(k)}$.

At a conceptual level, Theorem I.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 identifies a common barrier of the strongest known algorithmic approaches for this learning problem, and provides a rigorous explanation why a long line of algorithmic research on this front either relied on strong separation assumptions or resulted in runtimes of the form $n^{\Omega(k)}$.

SQ Lower Bound for Robustly Learning a Gaussian: Our second main result concerns the agnostic learning of a single n-dimensional Gaussian. We prove two SQ lower bounds with qualitatively similar guarantees for different versions of this problem. Our first lower bound is for the problem of agnostically learning a Gaussian with unknown mean and identity covariance. Roughly speaking, we show that any SQ algorithm that solves this learning problem to accuracy $O(\epsilon)$ requires complexity $n^{\Omega(\log^{1/4}(1/\epsilon))}$. We show:

Theorem I.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 $2^{n^{\Omega(1)}} \geq n^{\Omega(\log(1/\epsilon)^{c/2})}$ queries to an SQ oracle of precision $n^{-\Omega(\log(1/\epsilon)^{c/2})}$.*

Some comments are in order. First, Theorem I.2 shows a *super-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 an unknown mean Gaussian to within error $O(\epsilon)$. Second, the robust learning algorithm of [29] runs in $\mathrm{poly}(n, 1/\epsilon)$ time, can be implemented in the SQ model, and achieves error $O(\epsilon\sqrt{\log(1/\epsilon)})$. As a corollary of Theorem I.2, we

obtain that the $O(\epsilon\sqrt{\log(1/\epsilon)})$ error guarantee of the [29] algorithm is best possible among all polynomial-time SQ algorithms.

Roughly speaking, Theorem I.2 shows that any SQ algorithm that solves the (unknown mean Gaussian) robust learning problem to accuracy $O(\epsilon)$ needs to have running time at least $n^{\Omega(\log^{1/4}(1/\epsilon))}$, i.e., *quasi-polynomial* in $1/\epsilon$. It is natural to ask whether this quasi-polynomial lower bound can be improved to, say, exponential, e.g., $n^{\Omega(1/\epsilon)}$. We show that the lower bound of Theorem I.2 is qualitatively tight. We design 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 complexity. Specifically, we show:

Theorem I.3 (SQ Algorithm for Robust Learning of Unknown Mean Gaussian)**.** *Let D be a distribution on $\mathbb{R}^n$ such that $d_{\mathrm{TV}}(D, N(\mu, I)) \leq \epsilon$ for some $\mu \in \mathbb{R}^n$. There is an SQ algorithm that uses $O_\epsilon(n^{O(\sqrt{\log(1/\epsilon)})})$ SQ's 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 I.2 and I.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 distributions and $\epsilon_1 + \epsilon_2 \leq \epsilon$. A weaker error model, known as *Huber's contamination model* in the statistics literature [50], [45], [49], 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 the adversary is only allowed to add corrupted points. We note that the lower bound of Theorem I.2 does not apply in Huber's contamination model. This holds for a reason: Concurrent work [31] gives a $\mathrm{poly}(n/\epsilon)$ time algorithm with $O(\epsilon)$ error for robustly learning $N(\mu, I)$ in Huber's model. Hence, as a corollary, we establish a computational separation between these two models of corruptions. We provide an intuitive justification in Section I-C.

Our second super-polynomial SQ lower bound is for the problem of robustly learning a zero-mean unknown covariance Gaussian with respect to the *spectral norm*. Specifically, we show:

Theorem I.4 (SQ Lower Bound for Robust Learning of Unknown Covariance Gaussian)**.** *Let $\epsilon > 0$, $0 < c \leq 1$, and $n \geq \mathrm{poly}(\log(1/\epsilon))$. Any SQ algorithm that, given access to an ϵ-corrupted n-dimensional Gaussian $N(0, \Sigma)$, with $I/2 \preceq \Sigma \preceq 2I$, returns $\widetilde{\Sigma}$ with $\|\widetilde{\Sigma} - \Sigma\|_2 \leq O(\epsilon \log(1/\epsilon)^{1-c})$, requires at least $2^{n^{\Omega(1)}} \geq n^{\Omega(\log(1/\epsilon)^{c/4})}$ queries to an SQ oracle of precision $n^{-\Omega(\log(1/\epsilon)^{c/4})}$.*

Similarly, Theorem I.4 shows a *super-polynomial gap* between the information-theoretic sample complexity and the complexity of any SQ algorithm for this problem. As mentioned in the introduction, $O(n/\epsilon^2)$ samples information-theoretically suffice to agnostically learn the covariance to within spectral error $O(\epsilon)$. Second, the robust learning algorithm of [29] runs in $\mathrm{poly}(n, 1/\epsilon)$ time, can be implemented in the SQ model, and achieves error $O(\epsilon \log(1/\epsilon))$ in Mahalanobis distance (hence, also in spectral norm). Again, the immediate corollary is that the $O(\epsilon \log(1/\epsilon))$ error guarantee of the [29] algorithm is best possible among all polynomial-time SQ algorithms. The lower bound of Theorem I.4 does not apply in Huber's contamination model. This holds for a reason: [31] gives a $\mathrm{poly}(n) \cdot 2^{\mathrm{poly}\log(1/\epsilon)}$ time algorithm with $O(\epsilon)$ error in Huber's model.

Statistical–Computational Tradeoffs for SQ algorithms: Our next SQ lower bounds establish nearly quadratic statistical–computational tradeoffs for robust spectral covariance estimation and robust sparse mean estimation. We note that both these lower bounds also hold in Huber's contamination model. For the former problem, we show:

Theorem I.5. *Let $0 < c < 1/6$, and n sufficiently large. Any SQ algorithm that, given access to an ϵ-corrupted $N(0, \Sigma)$, where $\epsilon \leq c/\ln(n)$ for $\|\Sigma\|_2 \leq \mathrm{poly}(n/\epsilon)$, and returns $\tilde{\Sigma}$ with $\tilde{\Sigma}/2 \preceq \Sigma \preceq 2\tilde{\Sigma}$, requires at least $2^{\Omega(n^{c/3})}$ queries to an SQ oracle of precision $\gamma = O(n)^{-(1-5c/2)}$.*

We note that, in order to simulate a single query of the above precision, we need to draw $\Omega(1/\gamma^2) = \Omega(n^{2-5c})$ samples from our distribution. Roughly speaking, Theorem I.5 shows that if an SQ algorithm uses less than this many samples, then it needs to run in $2^{\Omega(n^{c/3})}$ time. This suggests a nearly-quadratic statistical-computational tradeoff for this problem.

For robust sparse mean estimation we show:

Theorem I.6. *Fix any $0 < c < 1$ and let $n \geq 8k^2$. Any SQ algorithm that, given access to an ϵ-corrupted $N(\mu, I)$, where $\epsilon = k^{-c/4}$, and $\mu \in \mathbb{R}^n$ is promised to be k-sparse with $\|\mu\|_2 = 1$, and outputs a hypothesis vector $\widehat{\mu}$ satisfying $\|\widehat{\mu} - \mu\|_2 \leq 1/2$, requires at least $n^{\Omega(ck^c)}$ queries to an SQ oracle of precision $\gamma = O(k)^{3c/2-1}$.*

Similarly, to simulate a single query of the above precision, we need to draw $\Omega(1/\gamma^2) = \Omega(k^{2-3c})$ samples from our distribution. Hence, any SQ algorithm that uses this many samples requires runtime at least $n^{\Omega(ck^c)}$. This suggests a nearly-quadratic statistical-computational tradeoff for this problem.

Sample Complexity Lower Bounds for High-Dimensional Testing: 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 in Huber's contamination model, we show:

Theorem I.7 (Sample Complexity Lower Bound for Robust Testing of Unknown Mean Gaussian)**.** *Fix $\epsilon > 0$. Any algorithm with sample access to a distribution D on $\mathbb{R}^n$ which satisfies either (a) $D = N(0, I)$ or (b) D is a δ-noisy $N(\mu, I)$, and $\|\mu\|_2 \geq \epsilon$, and distinguishes between the two cases with probability $2/3$ requires (i) $\Omega(n)$ samples if $\delta = \epsilon/100$, (ii) $\Omega(n^{1-c})$ samples if $\delta = \epsilon/n^{c/4}$, for any constant $0 < c < 1$.*

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 I.7 is that robustness makes the Gaussian mean testing problem *information-theoretically* harder. In particular, the sample complexity of robust testing is essentially the same as that of the corresponding learning problem. Theorem I.7 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 spherical GMM is close to a Gaussian:

Theorem I.8 (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) $D = (1/2)N(\mu_1, I) + (1/2)N(\mu_2, I)$ such that $d_{\mathrm{TV}}(D, N(0, I)) \geq \epsilon$, and distinguishes between the two cases with probability at least $2/3$ requires $\Omega(n/\epsilon^2)$ samples.*

Similarly, the sample lower bound of Theorem I.8 is optimal, up to constant factors, and coincides with the sample complexity of learning the underlying distribution.

C. Our Approach and Techniques

In this section, we provide a detailed outline of our approach and techniques. The structure of this section is as follows: We start by describing our Generic Lower Bound Construction, followed by our main applications to the problems of Learning GMMs and Robustly Learning an Unknown Gaussian. We continue with our applications to statistical–computational tradeoffs. We then explain how our generic technique can be used to obtain our Sample Complexity Testing Lower Bounds, which rely on essentially the same hard instances as our SQ lower bounds. We conclude with a sketch of our new (SQ) Algorithm for Robustly Learning an Unknown Mean Gaussian to optimal accuracy.

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. 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, we are able to show that any SQ algorithm with insufficient precision needs many queries in order to learn an unknown distribution from $\mathcal{D}$.

To prove our generic SQ lower bound, we need to bound from below the SQ-dimension of our hard family of distributions $\mathcal{D}$. Roughly speaking, the SQ-dimension of a distribution family corresponds to the number of *nearly uncorrelated* distributions (with respect to some fixed distribution) in the family. 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.

More concretely, our hard families of distributions are constructed as follows: Given a distribution A on the real-line, we define a family of high-dimensional distributions $\mathbf{P}_v(x)$, for $v \in \mathbb{S}_n$ 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 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: (1) 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}-order moment tensors. Since these tensors can be approximated in roughly n^m queries (and time), the aforementioned lower bound construction would necessarily fail unless the low-order moments of A match the corresponding low-order moments of G. We show that, aside from a few mild technical conditions, 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, to extract useful information from these higher degree tensors, one needs to approximate essentially all of the n^{m+1} many such tensor entries. (2) 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 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. To prove this fact, we make essential use of the Gaussian (Ornstein–Uhlenbeck) noise operator and its properties (see, e.g., [66]). We explain this connection in the following paragraph.

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 show that this quantity is *close* to a Gaussian, when $v \cdot v'$ is *close* to 0.

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 retains 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, 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 $H_i G$ with eigenvalue t^i. 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.

So far, we have provided a proof sketch of the following statement: When two unit vectors v, v' are *nearly* orthogonal, then the distributions $\mathbf{P}_v, \mathbf{P}_{v'}$ are *nearly* uncorrelated. Since, for $0 < c < 1/2$, we can pack $2^{\Omega(n^c)}$ unit vectors v onto the sphere so that their pairwise inner products are at most $n^{c-1/2}$, we obtain an SQ-dimension lower bound of our hard family. In particular, to learn the distribution $\mathbf{P}_v$, for unknown v, any SQ algorithm requires either $2^{\Omega(n^c)}$ queries or queries of accuracy better than $O(n)^{(m+1)(c-1/2)}$. This completes the proof sketch of our generic construction.

In our main applications, we construct one-dimensional distributions A satisfying the necessary moment-matching conditions for m taken to be *super-constant*, thus obtaining *super-polynomial* SQ lower bounds. For our quadratic statistical–computational tradeoffs, we match a constant number of moments. In the following paragraphs, we explain how we apply our framework to bound the SQ dimension for: (i) learning k-GMMs to constant accuracy, (ii) robustly learning an ϵ-corrupted Gaussian to accuracy $O(\epsilon)$, and (iii) robustly estimating a Gaussian covariance within constant spectral error and robustly estimating a sparse Gaussian mean to constant ℓ_2-error. In all cases, we construct a distribution A on the real-line that satisfies the necessary moment-matching conditions such that the family $\mathcal{D} = \{\mathbf{P}_v \mid v \in \mathbb{S}_n\}$ belongs in the appropriate class, e.g., is a k-GMM for (i), an ϵ-corrupted Gaussian for (ii), etc.

SQ Lower Bound for Learning k-GMMs: 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. For technical reasons, we require that the chi-squared divergence of A to $N(0,1)$ is bounded from above by an appropriate quantity. The Gaussian components, A_i, have the same variance and appropriately bounded means. We can also guarantee that the components A_i are almost non-overlapping. This implies that the corresponding high-dimensional distributions $\mathbf{P}_v, \mathbf{P}'_v$ will be at total variation distance close to 1 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 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 by adding a zero-mean skinny Gaussian to an appropriately rescaled version of B. Additional technical work 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 Brubaker and Vempala [11]. For the case of $k = 2$, 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 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 [11], 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 $\Omega(1/\sqrt{k})$. We emphasize once more that our hard family of instances is learnable with $O(k \log n)$ samples – both for density estimation and parameter estimation. On the other hand, any SQ learning algorithm for the family requires $n^{\Omega(k)}$ time.

SQ Lower Bounds for Robustly Learning Unknown Gaussian: In 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 corrupted points added by the adversary. The approach of [29] does not do anything to address subtractive noise, but shows that this type of noise can incur "small" error, e.g., at most $O(\epsilon\sqrt{\log(1/\epsilon)})$ for the case of unknown mean. For additive noise, [29] uses an iterative spectral algorithm to filter out outliers.

For concreteness, let us consider the case of robustly learning $N(\mu, I)$. Intuitively, achieving error $O(\epsilon)$ in the agnostic 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 $m = \Omega(\log^{1/4}(1/\epsilon))$ moments and is $\epsilon/100$-close in total variation distance to $G' = N(\epsilon, 1)$. 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 polynomials*, we can prove that its L_1 and L_∞ norms within the effective support of G' are much smaller than ϵ. This result is then used to bound from above the distance of A from G', which gives our SQ lower bound.

We use a similar technique to prove our SQ lower bound for robust covariance estimation in spectral norm. Specifically, we construct a distribution A that agrees with $N(0,1)$ on the first $m = \Omega(\log(1/\epsilon))$ moments and is $\epsilon/100$-close in total variation distance to $G' = N(0, (1-\delta)^2)$, for some $\delta = O(\epsilon)$. We similarly take A to be the Gaussian G' outside its effective support, while in the effective support we add an appropriate degree-m univariate polynomial p satisfying the appropriate moment conditions. The analysis proceeds similarly as above.

Statistical–Computational Tradeoffs for SQ algorithms: For robust covariance estimation in spectral norm, our one-dimensional distribution is selected to be $A = (1-\epsilon)N(0,\sigma) + \epsilon N_1$, where N_1 is a mixture of 2 unit-variance Gaussians with opposite means. By selecting σ appropriately, we can have A match the first 3 moments of $N(0,1)$. For robust sparse mean estimation, it suffices to take $A = (1-\delta)N(\epsilon, 1) + \delta N_1$, where N_1 is a unit-variance Gaussian selected so that $\mathbf{E}[A] = 0$. An important aspect of both these constructions is that the chi-squared distance $\chi^2(A, N(0,1))$ needs to be as small as possible. Indeed, since we only match a small number of moments, our bound on $\chi^2(A, N(0,1))$ crucially affects the accuracy of our SQ queries.

Sample Complexity Testing Lower Bounds: Our sample complexity lower bounds follow from standard information-theoretic arguments, and rely on the same lower bound instances 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 $G \sim N(0, I)$ 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

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

Note that the inner integral was bounded from above by roughly $(1 + (v \cdot v')^m)$. A careful analysis of the distribution of the angle between two random unit vectors allows us to show that, unless $N = \Omega(n)$, the chi-squared divergence is close to 1, and thus that this testing problem is impossible.

Algorithm for Robustly Learning Unknown Mean Gaussian: We give an SQ algorithm with $O(\epsilon)$-error for robustly learning an unknown mean Gaussian, showing that our corresponding SQ lower bound is qualitatively tight. Our algorithm builds on the filter technique of [29], 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 $\Omega(\epsilon)$-far 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}-order moment tensors for $t \leq k = \Omega(\log^{1/4}(1/\epsilon))$. Note that this already requires n^k SQ queries.

The first thing we will need to show is that k moments *suffice*, 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 show that $k = O(\log^{1/2}(1/\epsilon))$

suffices. Specifically, we prove a one-dimensional moment-matching lemma establishing the following: If an ϵ-noisy one-dimensional Gaussian approximately matches a reference Gaussian G in all of its first k moments, where $k = \Theta(\log^{1/2}(1/\epsilon))$ (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 statement 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 basic filter algorithm from [29], 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 robustly 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.

D. Related Work

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., [67], [61]) 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 for 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 [57], [21], [4], [72], [11], [53], [65], [6], [24], [25], [14], [23], [15], [16], [1], [27], [22], [33], [32] 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 I-A.

With respect to computational lower bounds for unsupervised estimation problems, the most relevant references are the works [41], [43], [55] 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 [37] 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 [37] that relied on the obvious combinatorial structure of the discrete setting.

A related line of work gives statistical-computational tradeoffs for sparse PCA [7], [8], [75], based on various computational hardness assumptions. These results are of similar flavor as our statistical–computational tradeoffs for SQ algorithms (Theorems I.5 and I.6). An important difference between these tradeoffs and the super-polynomial SQ lower bounds we prove in this paper (Theorems I.1, I.2, and I.4) is that the aforementioned sparse problems are known to be tractable if we increase the sample size by a quadratic factor beyond the information-theoretic limit. In contrast, our main SQ lower bound results establish a *super-polynomial* gap between the information-theoretic limit and the computational 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., [52], [58], [40], [59], [20], [19] for the problems of learning halfspaces and learning intersections thereof.

E. Discussion and Future Directions

The main contribution of this paper is a technique that gives essentially tight SQ lower bounds for a number of fundamental high-dimensional learning problems, including 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 corollary, we provide a rigorous explanation of the observed (super-polynomial) gap between the sample complexity of these problems and the runtime of the best known algorithms.

Our work naturally raises a number of interesting future directions. A natural open problem is to extend our lower bound technique to broader families of high-dimensional distributions. More concretely, is there a $k^{\omega(1)}\mathrm{poly}(n)$ SQ lower bound for learning k-mixtures of n-dimensional *spherical* Gaussians? Note that our $n^{\Omega(k)}$ 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 [69], albeit with exponential dependence on the number k of components. More broadly, can we extend our techniques to other families of structured high-dimensional distributions (e.g., mixtures of other distribution families)?

ACKNOWLEDGMENT

I.D. was supported by NSF Award CCF-1652862 (CAREER) and a Sloan Research Fellowship. D.K. was supported by NSF Award CCF-1553288 (CAREER) and a Sloan Research Fellowship.

REFERENCES

[1] J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt. Sample-optimal density estimation in nearly-linear time. In *SODA 2017*, pages 1278–1289, 2017.

[2] D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *COLT*, pages 458–469, 2005.

[3] J. Anderson, M. Belkin, N. Goyal, L. Rademacher, and J. R. Voss. The more, the merrier: the blessing of dimensionality for learning large gaussian mixtures. In *COLT 2014*, pages 1135–1164, 2014.

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

[5] 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.

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

[7] Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *COLT 2013*, pages 1046–1066, 2013.

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

[9] A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Symposium on Theory of Computing, STOC 2014*, pages 594–603, 2014.

[10] 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 *STOC*, pages 253–262, 1994.

[11] 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.

[12] T. Cai, Z. Ma, and Y. Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, 161(3):781–815, 2015.

[13] T. T. Cai, Z. Ma, and Y. Wu. Sparse pca: Optimal rates and adaptive estimation. *Ann. Statist.*, 41(6):3074–3110, 12 2013.

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

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

[16] 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.

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

[18] C.-T. Chu, S. K. Kim, Y. A. Lin, Y. Yu, G. Bradski, A. Y. Ng, and K. Olukotun. Map-reduce for machine learning on multicore. In *NIPS*, pages 281–288, 2006.

[19] A. Daniely. Complexity theoretic limitations on learning halfspaces. In *STOC 2016*, pages 105–117, 2016.

[20] A. Daniely, N. Linial, and S. S.-Shwartz. From average case complexity to improper learning complexity. In *STOC 2014*, pages 441–448, 2014.

[21] S. Dasgupta. Learning mixtures of Gaussians. In *FOCS*, pages 634–644, 1999.

[22] 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.

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

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

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

[26] C. Daskalakis and G. Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *COLT 2014*, pages 1183–1213, 2014.

[27] A. De, I. Diakonikolas, and R. Servedio. Learning from satisfying assignments. In *SODA 2015*, pages 478–497, 2015.

[28] I. Diakonikolas. Learning structured distributions. In P. Bühlmann, P. Drineas, M. Kane, and M. van Der Laan, editors, *Handbook of Big Data*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, chapter 15, pages 267–284. Taylor & Francis, 2016.

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

[30] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Being robust (in high dimensions) can be practical. *CoRR*, abs/1703.00893, 2017.

[31] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. *CoRR*, abs/1704.03866, 2017.

[32] 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.

[33] I. Diakonikolas, D. M. Kane, and A. Stewart. Optimal learning via the fourier transform for sums of independent integer random variables. In *Proceedings of COLT 2016*, pages 831–849, 2016.

[34] 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.

[35] S. Du, S. Balakrishnan, and A. Singh. Computationally efficient robust estimation of sparse functionals. *CoRR*, abs/1702.07709, 2017.

[36] J. Feldman, R. O'Donnell, and R. Servedio. PAC learning mixtures of Gaussians with no separation assumption. In *COLT 2006*, pages 20–34, 2006.

[37] 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.

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

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

[40] V. Feldman, P. Gopalan, S. Khot, and A. Ponnuswami. New results for learning noisy parities and halfspaces. In *FOCS*, pages 563–576, 2006.

[41] 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.

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

[43] V. Feldman, W. Perkins, and S. Vempala. On the complexity of random satisfiability problems with planted solutions. In *STOC 2015*, pages 77–86, 2015.

[44] R. Ge, Q. Huang, and S. M. Kakade. Learning mixtures of gaussians in high dimensions. In *STOC 2015*, pages 761–770, 2015.

[45] 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.

[46] M. Hardt and E. Price. Tight bounds for learning a mixture of two gaussians. In *STOC 2015*, pages 753–760, 2015.

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

[48] D. Hsu and S. M. Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *ITCS '13*, pages 11–20, 2013.

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

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

[51] I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

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

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

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

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

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

[57] 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.

[58] A. Klivans and A. Sherstov. Cryptographic hardness for learning intersections of halfspaces. In *FOCS*, pages 553–562, 2006.

[59] A. R. Klivans and P. Kothari. Embedding hard learning problems into gaussian space. In *APPROX/RANDOM 2014*, pages 793–809, 2014.

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

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

[62] J. Li. Robust sparse estimation tasks in high dimensions. *CoRR*, abs/1702.05860, 2017.

[63] J. Li and L. Schmidt. A nearly optimal and agnostic algorithm for properly learning a mixture of k gaussians, for any constant k. In *COLT*, 2017.

[64] A. Moitra. *Algorithmic aspects of machine learning*. 2014.

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

[66] R. O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, New York, NY, USA, 2014.

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

[68] 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.

[69] A. T. Suresh, A. Orlitsky, J. Acharya, and A. Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In *NIPS*, pages 1395–1403, 2014.

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

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

[72] S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. In *FOCS*, pages 113–122, 2002.

[73] R. Vershynin. How close is the sample covariance matrix totheactual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012.

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

[75] 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.

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

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