Mixture Models, Robustness, and Sum of Squares Proofs*

Samuel B. Hopkins

Cornell University

USA

samhop@cs.cornell.edu

Jerry Li

MIT

USA

jerryzli@mit.edu

ABSTRACT

We use the Sum of Squares method to develop new efficient algorithms for learning well-separated mixtures of Gaussians and robust mean estimation, both in high dimensions, that substantially improve upon the statistical guarantees achieved by previous efficient algorithms. Our contributions are:

Mixture models with separated means: We study mixtures of k distributions in d dimensions, where the means of every pair of distributions are separated by at least k^ϵ . In the special case of spherical Gaussian mixtures, we give a $(dk)^{O(1/\epsilon^2)}$ -time algorithm that learns the means assuming separation at least k^ϵ , for any $\epsilon > 0$. This is the first algorithm to improve on greedy (“single-linkage”) and spectral clustering, breaking a long-standing barrier for efficient algorithms at separation $k^{1/4}$.

Robust estimation: When an unknown $(1 - \epsilon)$ -fraction of $X_1, \dots, X_n$ are chosen from a sub-Gaussian distribution with mean μ but the remaining points are chosen adversarially, we give an algorithm recovering μ to error $\epsilon^{1-1/t}$ in time $d^{O(t^2)}$, so long as sub-Gaussian-ness up to $O(t)$ moments can be certified by a Sum of Squares proof. This is the first polynomial-time algorithm with guarantees approaching the information-theoretic limit for non-Gaussian distributions. Previous algorithms could not achieve error better than $\epsilon^{1/2}$.

Both of these results are based on a unified technique. Inspired by recent algorithms of Diakonikolas et al. in robust statistics, we devise an SDP based on the Sum of Squares method for the following setting: given $X_1, \dots, X_n \in \mathbb{R}^d$ for large d and $n = \text{poly}(d)$ with the promise that a subset of $X_1, \dots, X_n$ were sampled from a probability distribution with bounded moments, recover some information about that distribution.

CCS CONCEPTS

• **Theory of computation** → **Unsupervised learning and clustering**; • **Mathematics of computing** → *Probability and statistics*; *Semidefinite programming*;

*The full version of this paper is available on the authors’ websites, as well as at <https://arxiv.org/abs/1711.07454>.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

STOC’18, June 25–29, 2018, Los Angeles, CA, USA

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-5559-9/18/06...\$15.00

<https://doi.org/10.1145/3188745.3188748>

KEYWORDS

Unsupervised learning, clustering, mixture models, mixture of Gaussians, robust statistics, high-dimensional statistics, sum of squares method, semidefinite programming

ACM Reference Format:

Samuel B. Hopkins and Jerry Li. 2018. Mixture Models, Robustness, and Sum of Squares Proofs. In *Proceedings of 50th Annual ACM SIGACT Symposium on the Theory of Computing (STOC’18)*. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3188745.3188748>

1 INTRODUCTION

We propose and analyze a family of new algorithms for some fundamental high-dimensional statistical estimation problems. In particular, we give new algorithms for the following problems.

- (1) **Learning Δ -separated mixture models:** Given n samples $X_1, \dots, X_n \in \mathbb{R}^d$ from a mixture of k probability distributions $\mathcal{D}_1, \dots, \mathcal{D}_k$ on $\mathbb{R}^d$ with means $\mu_1, \dots, \mu_k \in \mathbb{R}^d$ and covariances $\Sigma_1, \dots, \Sigma_k \leq \text{Id}$, where $\|\mu_i - \mu_j\| \geq \Delta$, estimate $\mu_1, \dots, \mu_k$.¹
- (2) **Robust mean estimation:** Given n vectors $X_1, \dots, X_n \in \mathbb{R}^d$, of which a $(1 - \epsilon)$ -fraction are samples from a probability distribution $\mathcal{D}$ with mean μ and covariance $\Sigma \leq \text{Id}$ and the remaining ϵ -fraction are arbitrary vectors (which may depend on the $(1 - \epsilon)n$ samples from $\mathcal{D}$), estimate μ .

Mixture models, and especially Gaussian mixture models (where $\mathcal{D}_1, \dots, \mathcal{D}_k$ are Gaussian distributions) have been studied since Pearson in 1894 [55]. Work in theoretical computer science dates at least to the pioneering algorithm of Dasgupta in 1999 [19], which has been followed by numerous other algorithms and lower bounds [1–3, 12, 14, 20–22, 29, 30, 33, 37, 40, 45, 48, 52, 58, 62, 67–69].

Robust estimation in the form we study here is a more recent transplant to theoretical computer science [17, 18, 23–25, 27, 46, 61], but statisticians have long sought outlier-robust estimators. Formal study of arbitrarily-bad/adversarially-chosen outliers originates in the 1960s with the advent of “breakdown points” in statistics [13, 32, 38, 39, 64].

Though outwardly rather different, mixture model learning and robust estimation share some underlying structure. An algorithm for either must identify or otherwise recover information about one or several *structured* subsets of a number of samples $X_1, \dots, X_n \in \mathbb{R}^d$. In the mixture model case, each collection of all the samples from each distribution $\mathcal{D}_i$ is a structured subset. In the robust estimation case there is just one structured subset: the $(1 - \epsilon)n$

¹A mixture model consists of probability distributions $\mathcal{D}_1, \dots, \mathcal{D}_k$ on $\mathbb{R}^d$ and mixing weights $\lambda_1, \dots, \lambda_k \geq 0$ with $\sum_{i \leq k} \lambda_i = 1$. The distribution $\mathcal{D}_i$ has mean μ_i . Each sample x_j is generated by first sampling a component $i \in [k]$ according to the weights λ , then sampling $x_j \sim \mathcal{D}_i$.

samples drawn from the distribution $\mathcal{D}$.² Our algorithms are based on new techniques for identifying such structured subsets of points in large data sets.

For mixture models, a special case of our main result yields the first progress in more than 15 years on efficiently clustering mixtures of separated spherical Gaussians. The question here is: if $\mathcal{D}_1, \dots, \mathcal{D}_k$ are all Gaussian with covariance identity, what is the minimum cluster separation Δ which allows for a polynomial-time algorithm to estimate $\mu_1, \dots, \mu_k$ from $\text{poly}(k, d)$ samples from the mixture model? When $k = d$, the guarantees of the previous best algorithms for this problem, which require $\Delta \geq O(k^{1/4})$, are captured by a simple greedy clustering algorithm, sometimes called *single-linkage clustering*: when $\Delta \geq O(k^{1/4})$, with high probability every pair of samples from the same cluster is closer in Euclidean distance than every pair of samples from differing clusters.

We break this single-linkage clustering barrier: for every $\gamma > 0$ we give a $\text{poly}(k, d)$ -time algorithm for this problem when $\Delta > k^\gamma$. Our results extend to any k and d . In this more general setting the previous-best algorithms combine spectral dimension reduction (by projecting the samples to the top eigenvectors of an empirical covariance matrix) with single-linkage clustering [67]. These algorithms require separation $\Delta \geq O(\min(d, k)^{1/4})$, while our algorithms continue to tolerate separation $\Delta > k^\gamma$ for any $\gamma > 0$.³

Our algorithm relies on novel use of higher moments (in fact, $O(1/\gamma)$ moments) of the underlying distributions $\mathcal{D}_i$. Our main technical contribution is a new algorithmic technique for finding either a structured subset of data points or the empirical mean of such a subset when the subset consists of independent samples from a distribution $\mathcal{D}$ which has bounded higher-order moments *and there is a simple certificate of this boundedness*. This technique leverages the Sum of Squares (SoS) hierarchy of semidefinite programs (SDPs), and in particular a powerful approach for designing SoS-based algorithms in machine learning settings, developed and used in [7–9, 31, 36, 49, 56].

This SoS approach to unsupervised learning rests on a notion of *simple identifiability proofs*: the main step in designing an algorithm using SoS to recover some parameters θ from samples $x_1, \dots, x_n \sim p(x | \theta)$ is to prove in a restricted proof system that θ is likely to be uniquely identifiable from $x_1, \dots, x_n$. We develop this thoroughly later on, but roughly speaking one may think of this as requiring the identifiability proof to use only simple inequalities, such as Cauchy-Schwarz and Hölder's inequality, applied to low-degree polynomials. The simple identifiability proofs we construct for both the mixture models and robust estimation settings are heavily inspired by the robust estimation algorithms of Diakonikolas et al. [23].

²The recent work [17] codifies this similarity by unifying both these problems into what they call a list-decodable learning setting.

³In the years since an algorithm obtaining $\Delta \geq O(\min(d, k)^{1/4})$ was achieved by [67] there has been progress in extending similar results for more general clustering settings. In fact, the algorithm of [67] already tolerates isotropic, log-concave distributions, and allows for each component to have a distinct variance $\sigma_i^2 \in \mathbb{R}$, with the separation condition becoming $\|\mu_i - \mu_j\| > \max(\sigma_i, \sigma_j) \min(d, k)^{1/4}$. Later works such as [1, 5, 45] continued to generalize these results to broader clustering settings. Most related to the present work are spectral algorithms which weaken log-concavity to a bounded-covariance assumption, at the cost of requiring separation $\Delta > \sqrt{k}$.

1.1 Results

Both of the problems we study have a long history; for now we just note some highlights and state our main results.

Mixture models. The problem of learning mixture models dates to Pearson in 1894, who invented the method of moments in order to separate a mixture of two Gaussians [55]. Mixture models have since become ubiquitous in data analysis across many disciplines [50, 63]. In recent years, computer scientists have devised many ingenious algorithms for learning mixture models as it became clear that classical statistical methods (e.g. maximum likelihood estimation) often suffer from computational intractability, especially when there are many mixture components or the components are high dimensional.

A highlight of this work is a series of algorithmic results when the components of the mixture model are Gaussian [3, 19, 20, 67]. Here the main question is: how small can the cluster separation Δ be such that there exists an algorithm to estimate $\mu_1, \dots, \mu_k$ from samples $x_1, \dots, x_n$ in $\text{poly}(k, d)$ time (hence also using $n = \text{poly}(k, d)$ samples)? Focusing for simplicity on spherical Gaussian components (i.e. with covariance equal to the identity matrix Id) and with number of components similar to the ambient dimension of the data (i.e. $k = d$) and uniform mixing weights (i.e. every cluster has roughly the same representation among the samples), the best result in previous work gives a $\text{poly}(k)$ -time algorithm when $\Delta \geq k^{1/4}$.

Separation $\Delta = k^{1/4}$ represents a natural algorithmic barrier: when $\Delta \geq k^{1/4}$, *every pair of samples from the same cluster are closer to each other in Euclidean distance than are every pair of samples from distinct clusters (with high probability)*, while this is no longer true if $\Delta < k^{1/4}$. Thus, when $\Delta \geq k^{1/4}$, a simple greedy algorithm correctly clusters the samples into their components (this algorithm is sometimes called *single-linkage clustering*). On the other hand, standard information-theoretic arguments show that the means remain approximately identifiable from $\text{poly}(k, d)$ samples when Δ is as small as $O(\sqrt{\log k})$, but these methods yield only exponential-time algorithms.⁴ Nonetheless, despite substantial attention, this $\Delta = k^{1/4}$ barrier representing the breakdown of single-linkage clustering has stood for nearly 20 years.

We prove the following main theorem, breaking the single-linkage clustering barrier.

THEOREM 1.1 (INFORMAL, SPECIAL CASE FOR UNIFORM MIXTURE OF SPHERICAL GAUSSIANS (SEE FULL VERSION)). *For every $\gamma > 0$ there is an algorithm with running time $(dk)^{O(1/\gamma^2)}$ using at most $n \leq k^{O(1)} d^{O(1/\gamma)}$ samples which, given samples $x_1, \dots, x_n$ from a uniform mixture of k spherical Gaussians $N(\mu_i, \text{Id})$ in d dimensions with means $\mu_1, \dots, \mu_k \in \mathbb{R}^d$ satisfying $\|\mu_i - \mu_j\| \geq k^\gamma$ for each $i \neq j$, returns estimators $\hat{\mu}_1, \dots, \hat{\mu}_k \in \mathbb{R}^d$ such that $\|\hat{\mu}_i - \mu_i\| \leq 1/\text{poly}(k)$ (with high probability).*

We pause here to make several remarks about this theorem. Our algorithm makes novel use of higher order moments of Gaussian (and sub-Gaussian) distributions. Most previous work for efficiently

⁴Recent and sophisticated arguments show that the means are identifiable (albeit inefficiently) with error depending only on the number of samples and not on the separation Δ even when $\Delta = O(\sqrt{\log k})$ [58].

learning well-separated mixtures either used only second-order moment information, and required separation $\Delta \geq \Omega(\sqrt{k})$, or made mild use of log-concavity to improve this to $k^{1/4}$, whereas we use $O(1/\gamma)$ moments.

The guarantees of our theorem hold well beyond the Gaussian setting; the theorem applies to any mixture model with k^γ separation and whose component distributions $\mathcal{D}_1, \dots, \mathcal{D}_k$ are what we term $O(1/\gamma)$ -explicitly bounded. We define this notion formally below, but roughly speaking, a t -explicitly bounded distribution $\mathcal{D}$ has t -th moments obeying a subgaussian-type bound—that is, for every unit vector $u \in \mathbb{R}^d$ one has $\mathbb{E}_{Y \sim \mathcal{D}} |Y \cdot u|^t \leq t^{t/2}$ —and there is a certain kind of *simple certificate* of this fact, namely a low-degree Sum of Squares proof. Among other things, this means the theorem also applies to mixtures of symmetric product distributions with bounded moments.

For mixtures of distributions with sufficiently-many bounded moments (such as Gaussians), our guarantees go even further. We show that using $d^{O(\log k)^2}$ time and $d^{O(\log k)}$ samples, we can recover the means to error $1/\text{poly}(k)$ even if the separation is only $C\sqrt{\log k}$ for some universal constant C . Strikingly, [58] show that any algorithm that can learn the means nontrivially given separation $o(\sqrt{\log k})$ must require super-polynomial samples and time. Our results show that just above this threshold, it is possible to learn with just quasipolynomially many samples and time.

Finally, throughout the paper we state error guarantees roughly in terms of obtaining $\hat{\mu}_i$ with $\|\hat{\mu}_i - \mu_i\| \leq 1/\text{poly}(k) \ll k^\gamma$, meaning that we get ℓ_2 error which is much less than the true separation. In the special case of spherical Gaussians, we note that we can use our algorithm as a warm-start to recent algorithms due to [58], and achieve error δ using $\text{poly}(m, k, 1/\delta)$ additional runtime and samples for some polynomial independent of γ .

Robust mean estimation. Estimators which are robust to outlying or corrupted samples have been studied in statistics at least since the 1960s [38, 65]. The model we consider in this paper is a slight generalization of Hübner's contamination model [38]. We are given $X_1, \dots, X_n$, originally drawn iid from some unknown distribution $\mathcal{D}$, but an adversary has changed an ε fraction of these points adversarially. We call such a set of points ε -corrupted.⁵ The goal of robust statistics is to recover statistics of $\mathcal{D}$ such as mean and covariance, given ε -corrupted samples from $\mathcal{D}$.

In classical robust statistics, the robust mean estimation problem is known as *robust estimation of location*, and robust covariance estimation is known as *robust estimation of scale*. Classical works consider a measure known as breakdown point, which is (informally) the fraction of samples that an adversary must corrupt before the estimator has no provable guarantees. They often design robust estimators for mean and covariance that achieve optimal error in many fundamental settings. For instance, given samples from a symmetric sub-Gaussian distribution in k dimensions such that an ε -fraction are arbitrarily corrupted, an estimator known as the Tukey median [65] achieves error $O(\varepsilon)$, which is information theoretically optimal. However, these estimators are all NP-hard to

compute [13, 39] and the best known algorithms require $\exp(d)$ time in general.

For a long time, all known computationally efficient robust statistics for the mean or covariance of a d -dimensional Gaussian had error degrading polynomially with the dimension.⁶ In recent work, [23, 46] gave efficient and robust estimators for these statistics which achieve substantially better error. In particular, [23] achieve error $O(\varepsilon\sqrt{\log 1/\varepsilon})$ for estimating the mean of a Gaussian with identity covariance, and error $O(\varepsilon \log^{3/2} 1/\varepsilon)$ for robustly estimating the mean of a Gaussian with unknown variance $\Sigma \leq I$.

Unfortunately, these results are somewhat tailored to Gaussian distributions, or require covariance very close to identity. For general sub-Gaussian distributions with unknown variance $\Sigma \leq I$, the best known efficient algorithms achieve only $O(\varepsilon^{1/2})$ error [24, 61]. We substantially improve this, under a slightly stronger condition than sub-Gaussianity. Recall that a distribution $\mathcal{D}$ with mean μ over $\mathbb{R}^d$ is sub-Gaussian if for every unit vector u and every $t \in \mathbb{N}$ even, the following moment bound holds:

$$\mathbb{E}_{X \sim \mathcal{D}} \langle u, X - \mu \rangle^t \leq t^{t/2}.$$

Informally stated, our algorithms will work under the condition that this moment bound can be certified by a low degree SoS proof, for all $s \leq t$. We call such distributions t -explicitly bounded (we are ignoring some parameters, see full version for a formal definition). This class captures many natural sub-Gaussian distributions, such as Gaussians, product distributions of sub-Gaussians, and rotations thereof (see full version). For such distributions, we show:

THEOREM 1.2 (INFORMAL, SEE FULL VERSION). *Fix $\varepsilon > 0$ sufficiently small and let $t \geq 4$. Let $\mathcal{D}$ be a $O(t)$ -explicitly bounded distribution over $\mathbb{R}^d$ with mean μ^* . There is an algorithm with sample complexity $d^{O(t)}(1/\varepsilon)^{O(1)}$ running time $(d^t \varepsilon)^{O(t)}$ such that given an ε -corrupted set of samples of sufficiently large size from $\mathcal{D}$, outputs μ so that with high probability $\|\mu - \mu^*\| \leq O(\varepsilon^{1-1/t})$.*

As with mixture models, we can push our statistical rates further, if we are willing to tolerate quasipolynomial runtime and sample complexity. In particular, we can obtain error $O(\varepsilon\sqrt{\log 1/\varepsilon})$ with $d^{O(\log 1/\varepsilon)}$ samples and $d^{O(\log 1/\varepsilon)^2}$ time.

1.2 Related Work

Mixture models. The literature on mixture models is vast so we cannot attempt a full survey here. The most directly related line of work to our results studies mixtures models under mean-separation conditions, and especially mixtures of Gaussians, where the number k of components of the mixture grows with the dimension d [3, 19, 20, 67]. The culmination of these works is the algorithm of Vempala and Wang, which used spectral dimension reduction to improve on the $d^{1/4}$ separation required by previous works to $k^{1/4}$ in ℓ_2 distance for $k \leq d$ spherical Gaussians in d dimensions. Concretely, they show the following:

THEOREM 1.3 ([67], INFORMAL). *There is a constant $C > 0$ and an algorithm with running time $\text{poly}(k, d)$ such that for every $\mu_1, \dots, \mu_k \in \mathbb{R}^d$ and $\sigma_1, \dots, \sigma_k > 0$, satisfying*

$$\|\mu_i - \mu_j\| > C \max(\sigma_i, \sigma_j) k^{1/4} \log^{1/4}(d)$$

⁶We remark that this was the state of affairs even for the Hübner contamination model.

⁵Hübner's contamination model essentially only allows the adversary to add corrupted points, but not remove uncorrupted points.

with high probability the algorithm produces estimates $\hat{\mu}_1, \dots, \hat{\mu}_k$ with $\|\mu_i - \hat{\mu}_i\| \leq 1/\text{poly}(k)$, given $\text{poly}(k, d)$ samples from a mixture $\frac{1}{k} \sum_{i \leq k} \mathcal{N}(\mu_i, \sigma_i I)$.

The theorem extends naturally to isotropic log-concave distributions; our main theorem generalizes to distributions with explicitly bounded moments. These families of distributions are not strictly comparable.

Other works have relaxed the requirement that the underlying distributions be Gaussian [1, 45]; to second-moment moment boundedness instead of to log-concavity; these algorithms typically tolerate separation of order $\sqrt{k}$ rather than $k^{1/4}$. Our work can be thought of as a generalization of these algorithms to use boundedness of higher moments. One recent work in this spirit uses SDPs to cluster mixture models under separation assumptions [51]; the authors show that a standard SDP relaxation of k -means achieves guarantees comparable to previously-known specially-tailored mixture model algorithms.

Information-theoretic sample complexity: Recent work of [58] considers the Gaussian mixtures problem in an information-theoretic setting: they show that there is some constant C so that if the means are pairwise separated by at least $C\sqrt{\log k}$, then the means can be recovered to arbitrary accuracy (given enough samples). They give an efficient algorithm which, warm-started with sufficiently-good estimates of the means, improves the accuracy to δ using $\text{poly}(1/\delta, d, k)$ additional samples. However, their algorithm for providing this warm start requires time exponential in the dimension d . Our algorithm requires somewhat larger separation but runs in polynomial time. Thus by combining the techniques in the spherical Gaussian setting we can estimate the means with ℓ_2 error δ in polynomial time using an extra $\text{poly}(1/\delta, d, k)$ samples, when the separation is at least k^γ , for any $\gamma > 0$.

Fixed number of Gaussians in many dimensions: Other works address parameter estimation for mixtures of $k \ll d$ Gaussians (generally $k = O(1)$ and d grows) under weak identifiability assumptions [12, 33, 40, 52]. In these works the only assumptions are that the component Gaussians are statistically distinguishable; the goal is to recover their parameters of the underlying Gaussians. It was shown in [52] that algorithms in this setting provably require $\exp(k)$ samples and running time. The question addressed in our paper is whether this lower bound is avoidable under stronger identifiability assumptions. A related line of work addresses proper learning of mixtures of Gaussians [21, 29, 48, 62], where the goal is to output a mixture of Gaussians which is close to the unknown mixture in total-variation distance, avoiding the $\exp(k)$ parameter-learning sample-complexity lower bound. These algorithms achieve $\text{poly}(k, d)$ sample complexity, but they all require $\exp(k)$ running time, and moreover, do not provide any guarantee that the parameters of the distributions output are close to those for the true mixture.

Tensor-decomposition methods: Another line of algorithms focus on settings where the means satisfy algebraic non-degeneracy conditions, which is the case for instance in smoothed analysis settings [2, 30, 37]. These algorithms are typically based on finding a rank-one decomposition of the empirical 3rd or 4th moment tensor of the mixture; they heavily use the special structure of these moments for Gaussian mixtures. One paper we highlight is [14],

which also uses much higher moments of the distribution. They show that in the smoothed analysis setting, the ℓ th moment tensor of the distribution has algebraic structure which can be algorithmically exploited to recover the means. Their main structural result holds only in the smoothed analysis setting, where samples from a mixture model on perturbed means are available.

In contrast, we do not assume any non-degeneracy conditions and use moment information only about the individual components rather than the full mixture, which always hold under separation conditions. Moreover, our algorithms do not need to know the exact structure of the 3rd or 4th moments. In general, clustering-based algorithms like ours seem more robust to modelling errors than algebraic or tensor-decomposition methods.

Expectation-maximization (EM): EM is the most popular algorithm for Gaussian mixtures in practice, but it is notoriously difficult to analyze theoretically. The works [6, 20, 22, 69] offer some theoretical guarantees for EM, but non-convergence results are a barrier to strong theoretical guarantees [68].

Robust statistics. The literature on robust estimation is too large to do justice to here. There has been a long line of work on making algorithms tolerant to error in supervised settings [41, 66], especially for learning halfspaces [4, 28, 42, 60], and for problems such as PCA [15, 16, 47, 70]. See [23] for a more detailed discussion on the relationship between these questions (and others) and the model we consider here.

We consider the classical statistical notion of robustness against corruption, introduced back in the 70's in seminal works of [32, 38, 64]. Even for the mean of a Gaussian distribution, essentially all classical robust estimators are hard in the worst case to compute ([13, 39]). However, a recent flurry of work ([17, 23, 27, 46, 61]) has given new, computationally efficient, nearly optimal robust estimators for the mean and covariance of a high dimensional Gaussian distribution. Given sufficiently-many samples from a sub-Gaussian distribution with identity covariance, where an ε -fraction are arbitrarily corrupted, these algorithms can output mean estimates which achieve error at most $O(\varepsilon\sqrt{\log 1/\varepsilon})$ in ℓ_2 , which is information-theoretically optimal up to the $\sqrt{\log 1/\varepsilon}$ factor. However, these mean estimation algorithms heavily rely on knowing that the covariance is equal (or very close) to the identity. When the distribution is a general sub-Gaussian distribution with unknown covariance, the best known error achieved by an efficient algorithm is $O(\varepsilon^{1/2})$ [24, 61]. Under a slightly stronger assumption, our algorithm is able to achieve $O(\varepsilon^{1-1/t})$ error in polynomial time, for arbitrarily large $t \in \mathbb{N}$, and error $O(\varepsilon\sqrt{\log 1/\varepsilon})$ in quasipolynomial time for distributions with $O(\log 1/\varepsilon)$ bounded moments.

SoS algorithms for unsupervised learning. SoS algorithms for unsupervised learning obtain the best known polynomial-time guarantees for many problems, including dictionary learning, tensor completion, and others [7–9, 31, 36, 49, 56]. While the running times of such algorithms are often large polynomials, due to the need to solve large SDPs, insights from the SoS algorithms have often been used in later works obtaining fast polynomial running times [34, 35, 59]. This lends hope that in light of our results there is a practical algorithm to learn mixture models under separation $k^{1/4-\varepsilon}$ for some $\varepsilon > 0$.

Concurrent work. Finally, we note that concurrent and independent works by several groups [26, 43, 44] have either obtained results or developed techniques similar to ours.

2 TECHNIQUES

In this section we give a high-level overview of the main ideas in our algorithms. First, we describe the proofs-to-algorithms methodology developed in recent work on SoS algorithms for unsupervised learning problems. Then we describe the core of our algorithms for mixture models and robust estimation: a simple proof of identifiability of the mean of a distribution $\mathcal{D}$ on $\mathbb{R}^d$ from samples $X_1, \dots, X_n$ when some fraction of the samples may not be from $\mathcal{D}$ at all.

2.1 Proofs to Algorithms for Machine Learning: the SoS Method

The Sum of Squares (SoS) hierarchy is a powerful tool in optimization, originally designed to approximately solve systems of polynomial equations via a hierarchy of increasingly strong but increasingly large semidefinite programming (SDP) relaxations (see [10] and the references therein). There has been much recent interest in using the SoS method to solve unsupervised learning problems in generative models [7, 8, 31, 36, 49, 56].

By now there is an established method for designing such SoS-based algorithms, which we employ in this paper. Consider a generic statistical estimation setting: there is a vector $\theta^* \in \mathbb{R}^k$ of parameters, and given some samples $x_1, \dots, x_n \in \mathbb{R}^d$ sampled iid according to $p(x | \theta^*)$, one wants to recover some $\hat{\theta}(x_1, \dots, x_n)$ such that $\|\theta^* - \hat{\theta}\| \leq \delta$ (for some appropriate norm $\|\cdot\|$ and $\delta \geq 0$). One says that θ^* is *identifiable* from $x_1, \dots, x_n$ if, for any θ with $\|\theta^* - \theta\| > \delta$, one has $\Pr(x_1, \dots, x_n | \theta) \ll \Pr(x_1, \dots, x_n | \theta^*)$. Often mathematical arguments for identifiability proceed via concentration of measure arguments culminating in a union bound over every possible θ with $\|\theta^* - \theta\| > \delta$. Though this would imply θ could be recovered via brute-force search, this type of argument generally has no implications for efficient algorithms.

The SoS proofs-to-algorithms method prescribes designing a simple proof of identifiability of θ from samples $x_1, \dots, x_n$. Here “simple” has a formal meaning: the proof should be captured by the low-degree SoS proof system. The SoS proof system can reason about equations and inequalities among low-degree polynomials. Briefly, if $p(y_1, \dots, y_m)$ and $q(y_1, \dots, y_m)$ are polynomials with real coefficients, and for every $y \in \mathbb{R}^m$ with $p(y) \geq 0$ it holds also that $q(y) \geq 0$, the SoS proof system can deduce that $p(y) \geq 0$ implies $q(y) \geq 0$ if there is a simple certificate of this implication: polynomials $r(y), s(y)$ which are sums-of-squares, such that $q(y) = r(y) \cdot p(y) + s(y)$. (Then r, s form an SoS proof that $p(y) \geq 0$ implies $q(y) \geq 0$.)

Remarkably, many useful polynomial inequalities have such certificates. For example, the usual proof of the Cauchy-Schwarz inequality $\langle y, z \rangle^2 \leq \|y\|^2 \|z\|^2$, where y, z are m -dimensional vectors, actually shows that the polynomial $\|y\|^2 \|z\|^2 - \langle y, z \rangle^2$ is a sum-of-squares in y and z . The simplicity of the certificate is measured by the degree of the polynomials r and s ; when these polynomials have small (usually constant) degree there is hope of transforming SoS proofs into polynomial-time algorithms. This transformation is possible because (under mild assumptions on p and q) the set

of low-degree SoS proofs is in fact captured by a polynomial-size semidefinite program.

Returning to unsupervised learning, the concentration/union-bound style of identifiability proofs described above are almost never captured by low-degree SoS proofs. Instead, the goal is to design

- (1) A system of constant-degree polynomial equations and inequalities

$$\mathcal{A} = \{p_1(\theta) = 0, \dots, p_m(\theta) = 0, q_1(\theta) \geq 0, \dots, q_m(\theta) \geq 0\},$$

where the polynomials p and q depend on the samples

$$x_1, \dots, x_n$$

such that with high probability θ^* satisfies all the equations and inequalities.

- (2) A low-degree SoS proof that $\mathcal{A}$ implies $\|\theta - \theta^*\| \leq \delta$ for some small δ and appropriate norm $\|\cdot\|$.

Clearly these imply that any solution θ of $\mathcal{A}$ also solves the unsupervised learning problem. It is in general NP-hard to find a solution to a system of low-degree polynomial equations and inequalities.

However, the SoS proof (2) means that such a search can be avoided. Instead, we will relax the set of solutions θ to $\mathcal{A}$ to a simple(er) convex set: the set of *pseudodistributions satisfying $\mathcal{A}$* . We define pseudodistributions formally later, for now saying only that they are the convex duals of SoS proofs which use the axioms $\mathcal{A}$. By this duality, the SoS proof (2) implies not only that any solution θ to $\mathcal{A}$ is a good choice of parameters but also that a good choice of parameters can be extracted from any pseudodistribution satisfying $\mathcal{A}$. (We are glossing over for now that this last step requires some SDP rounding algorithm, since we use only standard rounding algorithms in this paper.)

Thus, the final SoS algorithms from this method take the form: solve an SDP to find a pseudodistribution which satisfies $\mathcal{A}$ and round it to obtain an estimate $\hat{\theta}$ of θ^* . To analyze the algorithm, use the SoS proof (2) to prove that $\|\hat{\theta} - \theta^*\| \leq \delta$.

2.2 Hölder’s Inequality and Identifiability from Higher Moments

Now we discuss the core ideas in our simple SoS identifiability proofs. We have not yet formally defined SoS proofs, so our goal will just be to construct identifiability proofs which are (a) phrased in terms of inequalities of low-degree polynomials and (b) provable using only simple inequalities, like Cauchy-Schwarz and Hölder’s inequalities, leaving the formalities for later.

We consider an idealized version of situations we encounter in both the mixture model and robust estimation settings. Let $\mu^* \in \mathbb{R}^d$. Let $X_1, \dots, X_n \in \mathbb{R}^d$ have the guarantee that for some $T \subseteq [n]$ of size $|T| = \alpha n$, the vectors $\{X_i\}_{i \in T}$ are iid samples from $\mathcal{N}(\mu^*, \text{Id})$, a spherical Gaussian centered at μ^* ; for the other vectors we make no assumption. The goal is to estimate the mean μ^* .

The system $\mathcal{A}$ of polynomial equations and inequalities we employ will be designed so that a solution to $\mathcal{A}$ corresponds to a subset of samples $S \subseteq [n]$ of size $|S| = |T| = \alpha n$. We accomplish this by identifying S with its 0/1 indicator vector in $\mathbb{R}^n$ (this is standard). The inequalities in $\mathcal{A}$ will enforce the following crucial moment property on solutions: if $\mu = \frac{1}{|S|} \sum_{i \in S} X_i$ is the empirical mean of

samples in S and $t \in \mathbb{N}$, then

$$\frac{1}{|S|} \sum_{i \in S} \langle X_i - \mu, u \rangle^t \leq 2 \cdot t^{t/2} \cdot \|u\|^t \quad \text{for all } u \in \mathbb{R}^d. \quad (1)$$

This inequality says that every one-dimensional projection u of the samples in S , centered around their empirical mean, has a sub-Gaussian empirical t -th moment. (The factor 2 accounts for deviations in the t -th moments of the samples.) By standard concentration of measure, if $\alpha n \gg d^t$ the inequality holds for $S = T$. It turns out that this property can be enforced by polynomials of degree t . (Actually our final construction of $\mathcal{A}$ will need to use inequalities of matrix-valued polynomials but this can be safely ignored here.)

Intuitively, we would like to show that any S which satisfies $\mathcal{A}$ has empirical mean close to μ^* using a low-degree SoS proof. This is in fact true when $\alpha = 1 - \varepsilon$ for small ε , which is at the core of our robust estimation algorithm. However, in the mixture model setting, when $\alpha = 1/(\# \text{ of components})$, for each component j there is a subset $T_j \subseteq [n]$ of samples from component j which provides a valid solution $S = T_j$ to $\mathcal{A}$. The empirical mean of T_j is close to μ_j and hence not close to μ_i for any $i \neq j$.

We will prove something slightly weaker, which still demonstrates the main idea in our identifiability proof.

LEMMA 2.1. *With high probability, for every $S \subseteq [n]$ which satisfies (1), if $\mu = \frac{1}{|S|} \sum_{i \in S} X_i$ is the empirical mean of samples in S , then $\|\mu - \mu^*\| \leq 4t^{1/2} \cdot (|T|/|S \cap T|)^{1/t}$.*

Notice that a random $S \subseteq [n]$ of size αn will have $|S \cap T| \approx \alpha^2 n$. In this case the lemma would yield the bound $\|\mu - \mu^*\| \leq \frac{4t^{1/2}}{\alpha^{1/t}}$. Thinking of $\alpha \ll 1/t$, this bound improves exponentially as t grows. In the d -dimensional k -component mixture model setting, one has $1/\alpha = \text{poly}(k)$, and thus the bound becomes $\|\mu - \mu^*\| \leq 4t^{1/2} \cdot k^{O(1/t)}$. In a mixture model where components are separated by k^ε , such an estimate is nontrivial when $\|\mu - \mu^*\| \ll k^\varepsilon$, which requires $t = O(1/\varepsilon)$. This is the origin of the quantitative bounds in our mixture model algorithm.

We turn to the proof of Lemma 2.1. As we have already emphasized, the crucial point is that this proof will be accomplished using only simple inequalities, avoiding any union bound over all possible subsets S .

PROOF OF LEMMA 2.1. Let w_i be the 0/1 indicator of $i \in S$. To start the argument, we expand in terms of samples:

$$\begin{aligned} |S \cap T| \cdot \|\mu - \mu^*\|^2 &= \sum_{i \in T} w_i \|\mu - \mu^*\|^2 \\ &= \sum_{i \in T} w_i \langle \mu^* - \mu, \mu^* - \mu \rangle \\ &= \sum_{i \in T} w_i [\langle X_i - \mu, \mu^* - \mu \rangle + \langle \mu^* - X_i, \mu^* - \mu \rangle]. \end{aligned} \quad (2)$$

(3)

The key term to bound is the first one; the second amounts to a deviation term. By Hölder's inequality and for even t ,

$$\begin{aligned} \sum_{i \in T} w_i \langle X_i - \mu, \mu^* - \mu \rangle &\leq \left(\sum_{i \in T} w_i \right)^{\frac{t-1}{t}} \cdot \left(\sum_{i \in T} w_i \langle X_i - \mu, \mu^* - \mu \rangle^t \right)^{1/t} \\ &\leq \left(\sum_{i \in T} w_i \right)^{\frac{t-1}{t}} \cdot \left(\sum_{i \in [n]} w_i \langle X_i - \mu, \mu^* - \mu \rangle^t \right)^{1/t} \\ &\leq \left(\sum_{i \in T} w_i \right)^{\frac{t-1}{t}} \cdot 2t^{1/2} \cdot \|\mu^* - \mu\| \\ &= |S \cap T|^{\frac{t-1}{t}} \cdot 2t^{1/2} \cdot \|\mu^* - \mu\|. \end{aligned}$$

The second line follows by adding the samples from $[n] \setminus T$ to the sum; since t is even this only increases its value. The third line uses the moment inequality (1). The last line just uses the definition of w .

For the second, deviation term, we use Hölder's inequality again:

$$\begin{aligned} \sum_{i \in T} w_i \langle \mu^* - X_i, \mu^* - \mu \rangle &\leq \left(\sum_{i \in T} w_i \right)^{\frac{t-1}{t}} \cdot \left(\sum_{i \in T} \langle \mu^* - X_i, \mu^* - \mu \rangle^t \right)^{1/t}. \end{aligned}$$

The distribution of $\mu^* - X_i$ for $i \in T$ is $\mathcal{N}(0, \text{Id})$. By standard matrix concentration, if $|T| = \alpha n \gg d^t$,

$$\begin{aligned} \sum_{i \in T} \left[(X_i - \mu^*)^{\otimes t/2} \right] \left[(X_i - \mu^*)^{\otimes t/2} \right]^T &\leq 2|T| \mathbb{E}_{Y \sim \mathcal{N}(0, \text{Id})} \left(Y^{\otimes t/2} \right) \left(Y^{\otimes t/2} \right)^T \end{aligned}$$

with high probability and hence, using the quadratic form at $(\mu^* - \mu)^{\otimes t/2}$,

$$\sum_{i \in T} \langle \mu^* - X_i, \mu^* - \mu \rangle^t \leq 2|T| t^{t/2} \cdot \|\mu^* - \mu\|^t.$$

Putting these together and simplifying constants, we have obtained that with high probability,

$$|S \cap T| \cdot \|\mu - \mu^*\|^2 \leq 4t^{1/2} |T|^{1/t} \cdot |S \cap T|^{(t-1)/t} \cdot \|\mu - \mu^*\|$$

which simplifies to

$$|S \cap T|^{1/t} \cdot \|\mu - \mu^*\| \leq 4t^{1/2} |T|^{1/t}. \quad \square$$

2.3 From Identifiability to Algorithms

We now discuss how to use the ideas described above algorithmically for learning well-separated mixture models. The high level idea for robust estimation is similar. Given Lemma 2.1, a naive algorithm for learning mixture models would be the following: find a set of points T of size roughly n/k that satisfy the moment bounds described, and simply output their empirical mean. Since by a simple counting argument this set must have nontrivial overlap with the points from some mixture component, Lemma 2.1 guarantees that the empirical mean is close to mean of this component.

However, in general finding such a set of points is algorithmically difficult. In fact, it would suffice to find a distribution over such sets of points (since then one could simply sample from this distribution), however, this is just as computationally difficult. The critical insight is that because of the proof of Lemma 2.1 only uses facts about low degree polynomials, it suffices to find an object which is indistinguishable from such a distribution, considered as a functional on low-degree polynomials.

The natural object in this setting is a *pseudo-distribution*. Pseudo-distributions form a convex set, and for a set of low-degree polynomial equations and inequalities $\mathcal{A}$, it is possible to find a pseudo-distribution which is indistinguishable from a distribution over solutions to $\mathcal{A}$ (as such a functional) in polynomial time via semidefinite programming (under mild assumptions on $\mathcal{A}$). More specifically, the set of SoS proofs using axioms $\mathcal{A}$ is a semidefinite program (SDP), and the above pseudodistributions form the dual SDP. (We will make these ideas more precise in the next two sections.)

Our algorithm then proceeds via the following general framework: find an appropriate pseudodistribution via convex optimization, then leverage our low-degree sum of squares proofs to show that information about the true clusters can be extracted from this object by a standard SDP rounding procedure.

3 PRELIMINARIES

Throughout the paper we let d be the dimensionality of the data, and we will be interested in the regime where d is at least a large constant. We also let $\|v\|$ denote the ℓ_2 norm of a vector v , and $\|M\|_F$ to denote the Frobenius norm of a matrix M ; often we just write $\|M\|$. We will also give randomized algorithms for our problems that succeed with probability $1 - \text{poly}(1/k, 1/d)$; by standard techniques this probability can be boosted to $1 - \xi$ by increasing the sample and runtime complexity by a multiplicative $\log 1/\xi$.

We now formally define the class of distributions we will consider throughout this paper. At a high level, we will consider distributions which have bounded moments, for which there exists a low degree SoS proof of this moment bound. Formally:

Definition 3.1. Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean μ . For $c \geq 1, t \in \mathbb{N}$, we say that $\mathcal{D}$ is t -explicitly bounded with variance proxy σ if for every even $s \leq t$ there is a degree s SoS proof (see Section 3.1 for a formal definition) of

$$\vdash_s E_{Y \sim \mathcal{D}_k} \langle (Y - \mu), u \rangle^s \leq (\sigma s)^{s/2} \|u\|^s.$$

Equivalently, the polynomial $p(u) = (\sigma s)^{s/2} \|u\|^s - E_{Y \sim \mathcal{D}_k} \langle (Y - \mu), u \rangle^s$ should be a sum-of-squares. In our typical use case, $\sigma = 1$, we will omit it and call the distribution t -explicitly bounded.

Throughout this paper, since all of our problems are scale invariant, we will assume without loss of generality that $\sigma = 1$. This class of distributions captures a number of natural classes of distributions. Intuitively, if u were truly a vector in $\mathbb{R}^k$ (rather than a vector of indeterminates), then this exactly captures sub-Gaussian type moment. Our requirement is simply that these types of moment bounds not only hold, but also have a SoS proof.

We remark that our results also hold for somewhat more general settings. It is not particularly important that the s -th moment bound has a degree s proof; our techniques can tolerate degree $O(s)$ proofs.

Our techniques also generally apply for weaker moment bounds. For instance, our techniques naturally extend to explicitly bounded sub-exponential type distributions in the obvious way. We omit these details for simplicity.

As we show in the full version of this paper, this class still captures many interesting types of nice distributions, including Gaussians, product distributions with sub-Gaussian components, and rotations thereof. With this definition in mind, we can now formally state the problems we consider in this paper:

Learning well-separated mixture models. We first define the class of mixture models for which our algorithm works:

Definition 3.2 (t -explicitly bounded mixture model with separation Δ). Let $\mu_1, \dots, \mu_k \in \mathbb{R}^d$ satisfy $\|\mu_i - \mu_j\| > \Delta$ for every $i \neq j$, and let $\mathcal{D}_1, \dots, \mathcal{D}_k$ have means $\mu_1, \dots, \mu_k$, so that each $\mathcal{D}_i$ is t -explicitly bounded. Let $\lambda_1, \dots, \lambda_k \geq 0$ satisfy $\sum_{i \in [k]} \lambda_i = 1$. Together these define a mixture distribution on $\mathbb{R}^d$ by first sampling $i \sim \lambda$, then sampling $x \sim \mathcal{D}_i$.

The problem is then:

Problem 3.1. Let $\mathcal{D}$ be a t -explicitly bounded mixture model in $\mathbb{R}^d$ with separation Δ with k components. Given k, Δ , and n independent samples from $\mathcal{D}$, output $\hat{\mu}_1, \dots, \hat{\mu}_m$ so that with probability at least 0.99, there exists a permutation $\pi : [k] \rightarrow [k]$ so that $\|\mu_i - \hat{\mu}_{\pi(i)}\| \leq \delta$ for all $i = 1, \dots, k$.

Robust mean estimation. We consider the same basic model of corruption introduced in [23].

Definition 3.3 (ε -corruption). We say a set of samples $X_1, \dots, X_n$ is ε -corrupted from a distribution $\mathcal{D}$ if they are generated via the following process. First, n independent samples are drawn from $\mathcal{D}$. Then, an adversary changes εn of these points arbitrarily, and the altered set of points is then returned to us in an arbitrary order.

The problem we consider in this setting is the following:

Problem 3.2 (Robust mean estimation). Let $\mathcal{D}$ be an $O(t)$ -explicitly bounded distribution over $\mathbb{R}^d$ with mean μ . Given t, ε , and an ε -corrupted set of samples from $\mathcal{D}$, output $\hat{\mu}$ satisfying $\|\mu - \hat{\mu}\| \leq O(\varepsilon^{1-1/t})$.

3.1 The SoS Proof System

We refer the reader to [10, 54] and the references therein for a thorough exposition of the SoS algorithm and proof system; here we only define what we need.⁷

Let $x_1, \dots, x_n$ be indeterminates and $\mathcal{A}$ be the set of polynomial equations and inequalities $\{p_1(x) \geq 0, \dots, p_m(x) \geq 0, q_1(x) = 0, \dots, q_m(x) = 0\}$. We say that the statement $p(x) \geq 0$ has an SoS proof if there are polynomials $\{r_\alpha\}_{\alpha \subseteq [m]}$ (where α may be a multiset) and $\{s_i\}_{i \in [m]}$ such that

$$p(x) = \sum_{\alpha} r_{\alpha}(x) \cdot \prod_{i \in \alpha} p_i(x) + \sum_{i \in [m]} s_i(x) q_i(x)$$

and each polynomial $r_{\alpha}(x)$ is a sum of squares.

⁷Our definition of SoS proofs differs slightly from O'Donnell and Zhou's in that we allow proofs to use products of axioms.

If the polynomials $r_\alpha(x) \cdot \prod_{i \in \alpha} p_i(x)$ and $s_i(x)q_i(x)$ have degree at most d , we say the proof has degree at most d , and we write

$$\mathcal{A} \vdash_d p(x) \geq 0.$$

SoS proofs compose well, and we frequently use the following without comment.

FACT 3.3. *If $\mathcal{A} \vdash_d p(x) \geq 0$ and $\mathcal{A} \vdash_{d'} q(x) \geq 0$, then $\mathcal{A} \cup \mathcal{B} \vdash_{\max(d,d')} p(x) + q(x) \geq 0$ and $\mathcal{A} \cup \mathcal{B} \vdash_{dd'} p(x)q(x) \geq 0$.*

We turn to the dual objects to SoS proofs. A degree- d pseudoexpectation (for variety we sometimes say “pseudodistribution”) is a linear operator $\tilde{\mathbb{E}} : \mathbb{R}[x]_{\leq d} \rightarrow \mathbb{R}$, where $\mathbb{R}[x]_{\leq d}$ are the polynomials in indeterminates x with real coefficients, which satisfies the following

- (1) Normalization: $\tilde{\mathbb{E}}[1] = 1$
- (2) Positivity: $\tilde{\mathbb{E}}[p(x)^2] \geq 0$ for every p of degree at most $d/2$.

We say that a degree- d pseudoexpectation $\tilde{\mathbb{E}}$ satisfies inequalities and equalities $\{p_1(x) \geq 0, \dots, p_m(x) \geq 0, q_1(x) = 0, \dots, q_m(x) = 0\}$ at degree $r \leq d$ if

- (1) for every multiset $\alpha \subseteq [m]$ and SoS polynomial $s(x)$ such that the degree of $s(x) \prod_{i \in \alpha} p_i(x)$ is at most r , one has $\tilde{\mathbb{E}}[s(x) \prod_{i \in \alpha} p_i(x)] \geq 0$, and
- (2) for every $q_i(x)$ and every polynomial $s(x)$ such that the degree of $q_i(x)s(x) \leq r$, one has $\tilde{\mathbb{E}}[s(x)q_i(x)] = 0$.

The main fact relating pseudoexpectations and SoS proofs is:

FACT 3.4 (SOUNDNESS OF SoS PROOFS, INFORMAL). *If $\mathcal{A}$ is a set of equations and inequalities and $\mathcal{A} \vdash_\ell p(x) \geq 0$, and $\tilde{\mathbb{E}}$ is a degree $d > \ell$ pseudodistribution satisfying $\mathcal{A}$ at degree d , then $\tilde{\mathbb{E}}$ satisfies $\mathcal{A} \cup \{p \geq 0\}$ at degree $d - \ell$.⁸*

In the full version of this work we state and prove many basic SoS inequalities that we will require throughout the paper.

Gaussian distributions are explicitly bounded. In the full version of this work we show that product distributions (and rotations thereof) with bounded t -th moments are explicitly bounded.

LEMMA 3.5. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ so that $\mathcal{D}$ is a rotation of a product distribution $\mathcal{D}'$ where each coordinate X with mean μ of $\mathcal{D}$ satisfies*

$$\mathbb{E}[(X - \mu)^s] \leq 2^{-s} \left(\frac{s}{2}\right)^{s/2}$$

Then $\mathcal{D}$ is t -explicitly bounded (with variance proxy 1).

(The factors of $\frac{1}{2}$ can be removed for many distributions, including Gaussians.)

4 CAPTURING EMPIRICAL MOMENTS WITH POLYNOMIALS

To describe our algorithms we need to describe a system of polynomial equations and inequalities which capture the following problem: among $X_1, \dots, X_n \in \mathbb{R}^d$, find a subset of $S \subseteq [n]$ of size αn such that the empirical t -th moments obey a moment bound: $\frac{1}{\alpha n} \sum_{i \in S} \langle X_i, u \rangle^t \leq t^{t/2} \|u\|^t$ for every $u \in \mathbb{R}^d$.

Let $k, n \in \mathbb{N}$ and let $w = (w_1, \dots, w_n), \mu = (\mu_1, \dots, \mu_k)$ be indeterminates. Let

- (1) $X_1, \dots, X_n \in \mathbb{R}^d$
- (2) $\alpha \in [0, 1]$ be a number (the intention is $|S| = \alpha n$).
- (3) $t \in \mathbb{N}$ be a power of 2, the order of moments to control
- (4) $\mu_1, \dots, \mu_k \in \mathbb{R}^d$, which will eventually be the means of a k -component mixture model, or when $k = 1$, the true mean of the distribution whose mean we robustly estimate.
- (5) $\tau > 0$ be some error magnitude accounting for fluctuations in the sizes of clusters (which may be safely ignored at first reading).

Definition 4.1. Let $\mathcal{A}$ be the following system of equations and inequalities, depending on all the parameters above.

- (1) $w_i^2 = w_i$ for all $i \in [n]$ (enforcing that w is a 0/1 vector, which we interpret as the indicator vector of the set S).
- (2) $(1 - \tau)\alpha n \leq \sum_{i \in [n]} w_i \leq (1 + \tau)\alpha n$, enforcing that $|S| \approx \alpha n$ (we will always choose $\tau = o(1)$).
- (3) $\mu \cdot \sum_{i \in [n]} w_i = \sum_{i \in [n]} w_i X_i$, enforcing that μ is the empirical mean of the samples in S
- (4) $\sum_{i \in [n]} w_i \langle X_i - \mu, \mu - \mu_j \rangle^t \leq 2 \cdot t^{t/2} \sum_{i \in [n]} w_i \|\mu - \mu_j\|^t$ for every μ_j among $\mu_1, \dots, \mu_m$. This enforces that the t -th empirical moment of the samples in S is bounded in the direction $\mu - \mu_j$.

Notice that since we will eventually take μ_j 's to be unknown parameters we are trying to estimate, the algorithm cannot make use of $\mathcal{A}$ directly, since the last family of inequalities involve the μ_j 's. Later in this paper we exhibit a system of inequalities which requires the empirical t -th moments to obey a sub-Gaussian type bound in every direction, hence implying the inequalities here without requiring knowledge of the μ_j 's to write down. Formally, we will show:

LEMMA 4.1. *Let $\alpha \in [0, 1]$. Let $t \in \mathbb{N}$ be a power of 2, $t \geq 4$.⁹ Let $0.1 > \tau > 0$. Let $X_1, \dots, X_n \in \mathbb{R}^d$. Let $\mathcal{D}$ be a $10t$ -explicitly bounded distribution.*

There is a family $\hat{\mathcal{A}}$ of polynomial equations and inequalities of degree $O(t)$ on variables $w = (w_1, \dots, w_n), \mu = (\mu_1, \dots, \mu_k)$ and at most $n^{O(t)}$ other variables, whose coefficients depend on $\alpha, t, \tau, X_1, \dots, X_n$, such that

- (1) (Satisfiability) *If there $S \subseteq [n]$ of size at least $(\alpha - \tau)n$ so that $\{X_i\}_{i \in S}$ is an iid set of samples from $\mathcal{D}$, and $(1 - \tau)\alpha n \geq d^{100t}$, then for d large enough, with probability at least $1 - d^{-8}$, the system $\hat{\mathcal{A}}$ has a solution over $\mathbb{R}$ which takes w to be the 0/1 indicator vector of S .*
- (2) (Solvability) *For every $C \in \mathbb{N}$ there is an $n^{O(Ct)}$ -time algorithm which, when $\hat{\mathcal{A}}$ is satisfiable, returns a degree- Ct pseudodistribution which satisfies $\hat{\mathcal{A}}$ (up to additive error 2^{-n}).*
- (3) (Moment bounds for polynomials of μ) *Let $f(\mu)$ be a length- d vector of degree- ℓ polynomials in indeterminates $\mu = (\mu_1, \dots, \mu_k)$. $\hat{\mathcal{A}}$ implies the following inequality and the implication has a degree $t\ell$ SoS proof.*

$$\hat{\mathcal{A}} \vdash_{O(t\ell)} \frac{1}{\alpha n} \sum_{i \in [n]} w_i \langle X_i - \mu, f(\mu) \rangle^t \leq 2 \cdot t^{t/2} \|f(\mu)\|^t.$$

⁸See [11] for a full account of completeness and soundness of SoS.

⁹The condition $t \geq 4$ is merely for technical convenience.

- (4) (Booleanness) $\widehat{\mathcal{A}}$ includes the equations $w_i^2 = w_i$ for all $i \in [n]$.
- (5) (Size) $\widehat{\mathcal{A}}$ includes the inequalities $(1 - \tau)\alpha n \leq \sum w_i \leq (1 + \tau)\alpha n$.
- (6) (Empirical mean) $\widehat{\mathcal{A}}$ includes the equation $\mu \cdot \sum_{i \in [n]} w_i = \sum_{i \in [n]} w_i X_i$.

In particular this implies that $\widehat{\mathcal{A}} \vdash_{O(t)} \mathcal{A}$.

The proof of Lemma 4.1 can be found in the full version of this work.

Remark 4.1 (Numerical accuracy, semidefinite programming, and other monsters). We pause here to address issues of numerical accuracy. Our final algorithms use point 2 in Lemma 4.1 (itself implemented using semidefinite programming) to obtain a pseudodistribution $\tilde{\mathbb{E}}$ satisfying $\widehat{\mathcal{A}}$ approximately, up to error $\eta = 2^{-n}$ in the following sense: for every r a sum of squares and $f_1, \dots, f_\ell \in \mathcal{A}$ with $\deg[r \cdot \prod f_i] \leq Ct$, one has $\tilde{\mathbb{E}} r \cdot \prod_{i \in \mathcal{A}} f_i \geq -\eta \cdot \|r\|$, where $\|r\|$ is ℓ_2 norm of the coefficients of r . Our main analyses of this pseudodistribution employ the implication $\widehat{\mathcal{A}} \vdash \mathcal{B}$ for another family of inequalities $\mathcal{B}$ to conclude that if $\tilde{\mathbb{E}}$ satisfies $\mathcal{A}$ then it satisfies $\mathcal{B}$, then use the latter to analyze our rounding algorithms. Because all of the polynomials eventually involved in the SoS proof $\widehat{\mathcal{A}} \vdash \mathcal{B}$ have coefficients bounded by n^B for some large constant B , it may be inferred that if $\tilde{\mathbb{E}}$ approximately satisfies $\widehat{\mathcal{A}}$ in the sense above, it also approximately satisfies $\mathcal{B}$, with some error $\eta' \leq 2^{-\Omega(n)}$. The latter is a sufficient for all of our rounding algorithms.

Aside from mentioning at a couple key points why our SoS proofs have bounded coefficients, we henceforth ignore all numerical issues. For further discussion of numerical accuracy and well-conditionedness issues in SoS, see [11, 53, 57]

5 MIXTURE MODELS: ALGORITHM AND ANALYSIS

In this section we formally describe and analyze our algorithm for mixture models. We prove the following theorem.

THEOREM 5.1 (MAIN THEOREM ON MIXTURE MODELS). *For every large-enough $t \in \mathbb{N}$ there is an algorithm with the following guarantees. Let $\mu_1, \dots, \mu_k \in \mathbb{R}^d$, satisfy $\|\mu_i - \mu_j\| \geq \Delta$. Let $\mathcal{D}_1, \dots, \mathcal{D}_k$ be $10t$ -explicitly bounded, with means $\mu_1, \dots, \mu_k$. Let $\lambda_1, \dots, \lambda_k \geq 0$ satisfy $\sum \lambda_i = 1$. Given $n \geq (d^t k)^{O(1)} \cdot (\max_{i \in [m]} 1/\lambda_i)^{O(1)}$ samples from the mixture model given by $\lambda_1, \dots, \lambda_k, \mathcal{D}_1, \dots, \mathcal{D}_k$, the algorithm runs in time $n^{O(t)}$ and with high probability returns $\{\hat{\mu}_1, \dots, \hat{\mu}_k\}$ (not necessarily in that order) such that*

$$\|\mu_i - \hat{\mu}_i\| \leq \frac{2^{Ct} m^C t^{t/2}}{\Delta^{t-1}}$$

for some universal constant C .

In particular, we note two regimes: if $\Delta = k^\gamma$ for a constant $\gamma > 0$, choosing $t = O(1/\gamma)$ we get that the ℓ_2 error of our estimator is $\text{poly}(1/k)$ for any $O(1/\gamma)$ -explicitly bounded distribution, and our estimator requires only $(dk)^{O(1)}$ samples and time. This matches the guarantees of Theorem 1.1.

On the other hand, if $\Delta = C' \sqrt{\log k}$ (for some universal C') then taking $t = O(\log k)$ gives error

$$\|\mu_i - \hat{\mu}_i\| \leq k^{O(1)} \cdot \left(\frac{\sqrt{t}}{\Delta} \right)^t$$

which, for large-enough C' and t , can be made $1/\text{poly}(k)$. Thus for $\Delta = C' \sqrt{\log k}$ and any $O(\log k)$ -explicitly bounded distribution we obtain error $1/\text{poly}(k)$ with $d^{O(\log k)}$ samples and $d^{O(\log k)^2}$ time.

In this section we describe and analyze our algorithm. To avoid some technical work we analyze the uniform mixtures setting, with $\lambda_i = 1/m$. In the full version of this work we describe how to adapt the algorithm to the nonuniform mixture setting.

5.1 Algorithm and Main Analysis

We formally describe our mixture model algorithm now. We use the following lemma, which we prove in Section 5.6. The lemma says that given a matrix which is very close, in Frobenious norm, to the 0/1 indicator matrix of a partition of $[n]$ it is possible to approximately recover the partition. (The proof is standard.)

LEMMA 5.2 (SECOND MOMENT ROUNDING, FOLLOWS FROM THEOREM 5.11). *Let $n, m \in \mathbb{N}$ with $m \ll n$. There is a polynomial time algorithm `ROUNDSECONDMENTS` with the following guarantees. Suppose $S_1, \dots, S_m$ partition $[n]$ into m pieces, each of size $\frac{n}{2m} \leq |S_i| \leq \frac{2n}{m}$. Let $A \in \mathbb{R}^{n \times n}$ be the 0/1 indicator matrix for the partition S ; that is, $A_{ij} = 1$ if $i, j \in S_\ell$ for some ℓ and is 0 otherwise. Let $M \in \mathbb{R}^{n \times n}$ be a matrix with $\|A - M\|_F \leq \varepsilon n$. Given M , with probability at least $1 - \varepsilon^2 m^3$ the algorithm returns a partition $C_1, \dots, C_m$ of $[n]$ such that up to a global permutation of $[m]$, $C_i = T_i \cup B_i$, where $T_i \subseteq S_i$ and $|T_i| \geq |S_i| - \varepsilon^2 m^2 n$ and $|B_i| \leq \varepsilon^2 m^2 n$.*

Algorithm 1 Mixture Model Learning

- 1: **function** `LEARNMIXTUREMEANS`($t, X_1, \dots, X_n, \delta, \tau$)
 - 2: By semidefinite programming (see Lemma 4.1, item 2), find a pseudoexpectation of degree $O(t)$ which satisfies the structured subset polynomials from Lemma 4.1, with $\alpha = n/m$ such that $\|\tilde{\mathbb{E}} w w^\top\|_F$ is minimized among all such pseudoexpectations.
 - 3: Let $M \leftarrow m \cdot \tilde{\mathbb{E}} w w^\top$.
 - 4: Run the algorithm `ROUNDSECONDMENTS` on M to obtain a partition $C_1, \dots, C_m$ of $[n]$.
 - 5: Run the algorithm `ESTIMATEMEAN` from the full version of this paper on each cluster C_i , with $\varepsilon = 2^{Ct} t^{t/2} m^4 / \Delta^t$ for some universal constant C to obtain a list of mean estimates $\hat{\mu}_1, \dots, \hat{\mu}_m$.
 - 6: Output $\hat{\mu}_1, \dots, \hat{\mu}_m$.
 - 7: **end function**
-

Remark 5.1 (On the use of `ESTIMATEMEAN`). As described, `LEARNMIXTUREMEANS` has two phases: a clustering phase and a mean-estimation phase. The clustering phase is the heart of the algorithm; we will show that after running `ROUNDSECONDMENTS` the algorithm has obtained clusters $C_1, \dots, C_k$ which err from the ground-truth clustering on only a $\frac{2^{O(t)} t^{t/2} \text{poly}(k)}{\Delta^t}$ -fraction of points. To obtain estimates $\hat{\mu}_i$ of the underlying means from such a clustering, one simple option is to output the empirical mean of the clusters.

However, without additional pruning this risks introducing error in the mean estimates which grows with the ambient dimension d . By using the robust mean estimation algorithm instead to obtain mean estimates from the clusters we obtain errors in the mean estimates which depend only on the number of clusters k , the between-cluster separation Δ , and the number t of bounded moments.

Remark 5.2 (Running time). We observe that LEARNMIXTUREMEANS can be implemented in time $n^{O(t)}$. The main theorem requires $n \geq k^{O(1)}d^{O(t)}$, which means that the final running time of the algorithm is $(kd^t)^{O(t)}$.¹⁰

5.2 Proof of Main Theorem

In this section we prove our main theorem using the key lemmata; in the following sections we prove the lemmata.

Deterministic Conditions. We recall the setup. There are k mean vectors $\mu_1, \dots, \mu_k \in \mathbb{R}^d$, and corresponding distributions $\mathcal{D}_1, \dots, \mathcal{D}_k$ where $\mathcal{D}_j$ has mean μ_j . The distributions $\mathcal{D}_j$ are $10t$ -explicitly bounded for a choice of t which is a power of 2. Vectors $X_1, \dots, X_n \in \mathbb{R}^d$ are samples from a uniform mixture of $\mathcal{D}_1, \dots, \mathcal{D}_k$. We will prove that our algorithm succeeds under the following condition on the samples $X_1, \dots, X_n$.

- (D1) (Empirical moments) For every cluster $S_j = \{X_i : X_i \text{ is from } \mathcal{D}_j\}$, the system $\tilde{\mathcal{A}}$ from Lemma 4.1 with $\alpha = 1/m$ and $\tau = \Delta^{-t}$ has a solution which takes $w \in \{0, 1\}^n$ to be the 0/1 indicator vector of S_j .
- (D2) (Empirical means) Let $\bar{\mu}_j$ be the empirical mean of cluster S_j . The $\bar{\mu}_j$'s satisfy $\|\bar{\mu}_j - \mu_j\| \leq \Delta^{-t}$.

We note a few useful consequences of these conditions, especially (D1). First of all, it implies all clusters have almost the same size: $(1 - \Delta^{-t}) \cdot \frac{n}{k} \leq |S_j| \leq (1 + \Delta^{-t}) \cdot \frac{n}{k}$. Second, it implies that all clusters have explicitly bounded moments: for every S_j ,

$$\vdash_t \frac{k}{n} \sum_{i \in S_j} \langle X_i - \bar{\mu}_j, u \rangle^t \leq 2 \cdot t^{t/2} \cdot \|u\|^t.$$

Lemmas. The following key lemma captures our SoS identifiability proof for mixture models.

LEMMA 5.3. *Let $\mu_1, \dots, \mu_k, \mathcal{D}_1, \dots, \mathcal{D}_k$ be as in Theorem 5.1, with mean separation Δ . Suppose (D1), (D2) occur for samples $X_1, \dots, X_n$. Let $t \in \mathbb{N}$ be a power of two. Let $\tilde{\mathbb{E}}$ be a degree- $O(t)$ pseudoexpectation which satisfies $\mathcal{A}$ from Lemma 4.1 with $\alpha = 1/k$ and $\tau \leq \Delta^{-t}$. Then for every $j, \ell \in [k]$,*

$$\tilde{\mathbb{E}} \langle a_j, w \rangle \langle a_\ell, w \rangle \leq 2^{8t+8} \cdot t^{t/2} \cdot \frac{n^2}{k} \cdot \frac{1}{\Delta^t}.$$

¹⁰As discussed in Section 4, correctness of our algorithm at the level of numerical accuracy requires that the coefficients of every polynomial in the SoS program $\tilde{\mathcal{A}}$ (and every polynomial in the SoS proofs we use to analyze $\tilde{\mathcal{A}}$) are polynomially bounded. This may not be the case if some vectors $\mu_1, \dots, \mu_m$ have norms $\|\mu_i\| \geq d^{\omega(1)}$. This can be fixed by naively clustering the samples $X_1, \dots, X_n$ via single-linkage clustering, then running LEARNMIXTUREMEANS on each cluster. It is routine to show that the diameter of each cluster output by a naive clustering algorithm is at most $\text{poly}(d, k)$ under our assumptions, and that with high probability single-linkage clustering produces a clustering respecting the distributions $\mathcal{D}_j$. Hence, by centering each cluster before running LEARNMIXTUREMEANS we can assume that $\|\mu_i\| \leq \text{poly}(d, k)$ for every $i \leq d$.

The other main lemma shows that conditions (D1) and (D2) occur with high probability.

LEMMA 5.4 (CONCENTRATION FOR MIXTURE MODELS). *With notation as above, conditions (D1) and (D2) simultaneously occur with probability at least $1 - 1/d^{15}$ over samples $X_1, \dots, X_n$, so long as $n \geq d^{O(t)}k^{O(1)}$, for $\Delta \geq 1$.*

Lemma 5.4 follows from Lemma 4.1, for (D1), and standard concentration arguments for (D2). Now we can prove the main theorem.

PROOF OF THEOREM 5.1 (UNIFORM MIXTURES CASE). Suppose conditions (D1) and (D2) hold. Our goal will be to bound $\|M - A\|^2 \leq n \cdot \frac{2^{O(t)}t^{t/2}k^4}{\Delta^t}$, where A is the 0/1 indicator matrix for the ground truth partition $S_1, \dots, S_k$ of $X_1, \dots, X_n$ according to $\mathcal{D}_1, \dots, \mathcal{D}_k$. Then by Lemma 5.2, the rounding algorithm will return a partition $C_1, \dots, C_k$ of $[n]$ such that C_ℓ and S_ℓ differ by at most $n \frac{2^{O(t)}t^{t/2}k^{10}}{\Delta^t}$ points, with probability at least $1 - \frac{2^{O(t)}t^{t/2}k^{30}}{\Delta^t}$. By the guarantees of regarding the algorithm ESTIMATEMEAN (see full version of this work), with high probability the resulting error in the mean estimates $\hat{\mu}_i$ will satisfy

$$\|\mu_i - \hat{\mu}_i\| \leq \sqrt{t} \cdot \left(\frac{2^{O(t)}t^{t/2}k^{10}}{\Delta^t} \right)^{\frac{t-1}{t}} \leq \frac{2^{O(t)} \cdot t^{t/2} \cdot k^{10}}{\Delta^{t-1}}.$$

We turn to the bound on $\|M - A\|^2$. First we bound $\langle \tilde{\mathbb{E}} ww^\top, A \rangle$. Getting started,

$$\tilde{\mathbb{E}} \left(\sum_{i \in [k]} \langle w, a_i \rangle \right)^2 = \tilde{\mathbb{E}} \left(\sum_{i \in [n]} w_i \right)^2 \geq (1 - \Delta^{-t})^2 \cdot n^2/k^2.$$

By Lemma 5.3, choosing t later,

$$\sum_{i \neq j \in [k]} \tilde{\mathbb{E}} \langle a_i, w \rangle \langle a_j, w \rangle \leq n^2 2^{O(t)} t^{t/2} \cdot k \cdot \frac{1}{\Delta^t}.$$

Together, these imply

$$\tilde{\mathbb{E}} \sum_{i \in [k]} \langle w, a_i \rangle^2 \geq \frac{n^2}{k^2} \cdot \left[1 - \frac{2^{O(t)}t^{t/2}k^3}{\Delta^t} \right].$$

At the same time, $\|\tilde{\mathbb{E}} ww^\top\|_F \leq \frac{1}{k} \|A\|_F$ by minimality (since the uniform distribution over cluster indicators satisfies $\mathcal{A}$), and by routine calculation and assumption (D1), $\|A\|_F \leq \frac{n}{\sqrt{k}}(1 + O(\Delta^{-t}))$.

Together, we have obtained

$$\langle M, A \rangle \geq \left(1 - \frac{2^{O(t)}t^{t/2}k^3}{\Delta^t} \right) \cdot \|A\| \|M\|$$

which can be rearranged to give $\|M - A\|^2 \leq n \cdot \frac{2^{O(t)}t^{t/2}k^4}{\Delta^t}$. $\square$

5.3 Identifiability

In this section we prove Lemma 5.3. We use the following helpful lemmas. The first is in spirit an SoS version of Lemma 2.1.

LEMMA 5.5. *Let $\mu_1, \dots, \mu_k, \mathcal{D}_1, \dots, \mathcal{D}_k, t$ be as in Theorem 5.1. Let $\bar{\mu}_i$ be as in (D1). Suppose (D1) occurs for samples $X_1, \dots, X_n$. Let $\mathcal{A}$ be the system from Lemma 4.1, with $\alpha = 1/k$ and any τ . Then*

$$\mathcal{A} \vdash_{O(t)} \langle a_j, w \rangle^t \|\mu - \bar{\mu}_j\|^{2t} \leq 2^{t+2} t^{t/2} \cdot \frac{n}{k} \cdot \langle a_j, w \rangle^{t-1} \cdot \|\mu - \bar{\mu}_j\|^t.$$

The second lemma is an SoS triangle inequality, capturing the consequences of separation of the means. The proof is standard; see full version.

LEMMA 5.6. Let $a, b \in \mathbb{R}^k$ and $t \in \mathbb{N}$ be a power of 2. Let $\Delta = \|a - b\|$. Let $u = (u_1, \dots, u_k)$ be indeterminates. Then $\vdash_t \|a - u\|^t + \|b - u\|^t \geq 2^{-t} \cdot \Delta^t$.

The last lemma helps put the previous two together. Although we have phrased this lemma to concur with the mixture model setting, we note that the proof uses nothing about mixture models and consists only of generic manipulations of pseudodistributions.

LEMMA 5.7. Let $\mu_1, \dots, \mu_k, \mathcal{D}_1, \dots, \mathcal{D}_k, X_1, \dots, X_n$ be as in Theorem 5.1. Let a_j be the 0/1 indicator for the set of samples drawn from $\mathcal{D}_j$. Suppose $\tilde{\mathbb{E}}$ is a degree- $O(t)$ pseudodistribution which satisfies

$$\begin{aligned} \langle a_j, w \rangle &\leq n \\ \langle a_\ell, w \rangle &\leq n \\ \|\mu - \bar{\mu}_j\|^{2t} + \|\mu - \bar{\mu}_\ell\|^{2t} &\geq A \\ \langle a_j, w \rangle^t \|\mu - \bar{\mu}_j\|^{2t} &\leq Bn \langle a_j, w \rangle^{t-1} \|\mu - \bar{\mu}_j\|^t \\ \langle a_\ell, w \rangle^t \|\mu - \bar{\mu}_\ell\|^{2t} &\leq Bn \langle a_\ell, w \rangle^{t-1} \|\mu - \bar{\mu}_\ell\|^t \end{aligned}$$

for some scalars $A, B \geq 0$. Then

$$\tilde{\mathbb{E}} \langle a_j, w \rangle \langle a_\ell, w \rangle \leq \frac{2n^2 B}{\sqrt{A}}.$$

Now we have the tools to prove Lemma 5.3.

PROOF OF LEMMA 5.3. We will verify the conditions to apply Lemma 5.7. By Lemma 5.5, when (D1) holds, the pseudoexpectation $\tilde{\mathbb{E}}$ satisfies

$$\langle a_j, w \rangle^t \|\mu - \bar{\mu}_j\|^{2t} \leq Bn \langle a_j, w \rangle^{t-1} \|\mu - \bar{\mu}_j\|^t$$

for $B = 4(4t)^{t/2}/k$, and similarly with j, ℓ interposed. Similarly, by separation of the empirical means, $\tilde{\mathbb{E}}$ satisfies $\|\mu - \bar{\mu}_j\|^{2t} + \|\mu - \bar{\mu}_\ell\|^{2t} \geq A$ for $A = 2^{-2t} \Delta^{2t}$, recalling that the empirical means are pairwise separated by at least $\Delta - 2\Delta^{-t}$. Finally, clearly $\mathcal{A} \vdash_{O(1)} \langle a_j, w \rangle \leq n$ and similarly for $\langle a_\ell, w \rangle$. So applying Lemma 5.7 we get

$$\tilde{\mathbb{E}} \langle a_j, w \rangle \langle a_\ell, w \rangle \leq \frac{2n^2 B}{\sqrt{A}} \leq \frac{n^2 2^{2t+2} t^{t/2}}{k} \cdot \frac{1}{\Delta^t}. \quad \square$$

5.4 Proof of Lemma 5.5

In this subsection we prove Lemma 5.5. We use the following helpful lemmata. The first bounds error from samples selected from the wrong cluster using the moment inequality.

LEMMA 5.8. Let $j, \mathcal{A}, X_1, \dots, X_n, \mu_j, \bar{\mu}_j$ be as in Lemma 5.5. Then

$$\mathcal{A} \vdash_{O(t)} \left(\sum_{i \in S_j} w_i \langle \mu - X_i, \mu - \bar{\mu}_j \rangle \right)^t \leq 2t^{t/2} \cdot \langle a_j, w \rangle^{t-1} \|\mu - \bar{\mu}_j\|^t.$$

PROOF. The proof goes by Hölder's inequality followed by the moment inequality in $\mathcal{A}$. Carrying this out, by SoS Hölder's inequality (see full version) and evenness of t ,

$$\begin{aligned} \{w_i^2 = w_i\} \vdash_{O(t)} &\left(\sum_{i \in S_j} w_i \langle \mu - X_i, \mu - \bar{\mu}_j \rangle \right)^t \\ &\leq \left(\sum_{i \in S_j} w_i \right)^{t-1} \cdot \left(\sum_{i \in [n]} w_i \langle \mu - X_i, \mu - \bar{\mu}_j \rangle^t \right). \end{aligned}$$

Then, using the main inequality in $\mathcal{A}$,

$$\begin{aligned} \mathcal{A} \vdash_{O(t)} &\left(\sum_{i \in S_j} w_i \right)^{t-1} \cdot 2t^{t/2} \cdot \|\mu - \bar{\mu}_j\|^t \\ &= 2t^{t/2} \cdot \langle a_j, w \rangle^{t-1} \|\mu - \bar{\mu}_j\|^t \quad \square \end{aligned}$$

The second lemma bounds error from deviations in the empirical t -th moments of the samples from the j -th cluster.

LEMMA 5.9. Let $\mu_1, \dots, \mu_k, \mathcal{D}_1, \dots, \mathcal{D}_k$ be as in Theorem 5.1. Suppose condition (D1) holds for samples $X_1, \dots, X_n$. Let $w_1, \dots, w_n$ be indeterminates. Let $u = u_1, \dots, u_d$ be an indeterminate. Then for every $j \in [k]$,

$$\begin{aligned} \{w_i^2 = w_i\} \vdash_{O(t)} &\left(\sum_{i \in S_j} w_i \langle X_i - \bar{\mu}_j, u \rangle \right)^t \\ &\leq \langle a_j, w \rangle^{t-1} \cdot 2 \cdot \frac{n}{k} \cdot \|u\|^t. \end{aligned}$$

PROOF. The first step is Hölder's inequality again:

$$\begin{aligned} \{w_i^2 = w_i\} \vdash_{O(t)} &\left(\sum_{i \in S_j} w_i \langle X_i - \bar{\mu}_j, u \rangle \right)^t \\ &\leq \langle a_j, w \rangle^{t-1} \cdot \sum_{i \in S_j} \langle X_i - \bar{\mu}_j, u \rangle^t. \end{aligned}$$

Finally, condition (D1) yields

$$\begin{aligned} \{w_i^2 = w_i\} \vdash_{O(t)} &\left(\sum_{i \in S_j} w_i \langle X_i - \bar{\mu}_j, u \rangle \right)^t \\ &\leq \langle a_j, w \rangle^{t-1} \cdot 2 \cdot \frac{n}{k} \cdot \|u\|^t. \quad \square \end{aligned}$$

We can prove Lemma 5.5 by putting together Lemma 5.8 and Lemma 5.9.

PROOF OF LEMMA 5.5. Let $j \in [k]$ be a cluster and recall $a_j \in \{0, 1\}^n$ is the 0/1 indicator for the samples in cluster j . Let S_j be the samples in the j -th cluster, with empirical mean $\bar{\mu}_j$. We begin by writing $\langle a_j, w \rangle \|\mu - \bar{\mu}_j\|^2$ in terms of samples $X_1, \dots, X_n$.

$$\begin{aligned} \langle a_j, w \rangle \|\mu - \bar{\mu}_j\|^2 &= \sum_{i \in [n]} w_i \langle \mu - \bar{\mu}_j, \mu - \bar{\mu}_j \rangle \\ &= \sum_{i \in S_j} w_i \langle \mu - X_i, \mu - \bar{\mu}_j \rangle + \sum_{i \in [n]} w_i \langle X_i - \bar{\mu}_j, \mu - \bar{\mu}_j \rangle. \end{aligned}$$

Hence, using $(a + b)^t \leq 2^t(a^t + b^t)$, we obtain

$$\begin{aligned} \vdash_{O(t)} \langle a_j, w \rangle^t \|\mu - \bar{\mu}_j\|^{2t} &\leq 2^t \cdot \left(\sum_{i \in S_j} w_i \langle \mu - X_i, \mu - \bar{\mu}_j \rangle \right)^t \\ &\quad + 2^t \cdot \left(\sum_{i \in S_j} w_i \langle X_i - \bar{\mu}_j, \mu - \bar{\mu}_j \rangle \right)^t. \end{aligned}$$

Now using Lemma 5.8 and Lemma 5.9,

$$\mathcal{A} \vdash_{O(t)} \langle a_j, w \rangle^t \|\mu - \bar{\mu}_j\|^{2t} \leq 2^{t+2} t^{t/2} \cdot \frac{n}{k} \cdot \langle a_j, w \rangle^{t-1} \cdot \|\mu - \bar{\mu}_j\|^t$$

as desired. $\square$

5.5 Proof of Lemma 5.7

We prove Lemma 5.7. The proof only uses standard SoS and pseudodistribution tools. The main inequality we will use is the following version of Hölder's inequality.

FACT 5.10 (PSEUDOEXPECTATION HÖLDER'S, SEE LEMMA A.4 IN [7]). *Let p be a degree- ℓ polynomial. Let $t \in \mathbb{N}$ and let $\tilde{\mathbb{E}}$ be a degree- $O(t\ell)$ pseudoexpectation on indeterminates x . Then*

$$\tilde{\mathbb{E}} p(x)^{t-2} \leq \left(\tilde{\mathbb{E}} p(x)^t \right)^{\frac{t-2}{t}}.$$

Now we can prove Lemma 5.7.

PROOF OF LEMMA 5.7. We first establish the following inequality.

$$\tilde{\mathbb{E}} \langle a_j, w \rangle^t \langle a_\ell, w \rangle^t \|\mu - \bar{\mu}_j\|^{2t} \leq B^2 n^2 \cdot \tilde{\mathbb{E}} \langle a_j, w \rangle^{t-2} \langle a_\ell, w \rangle^t. \quad (4)$$

(The inequality will also hold by symmetry with j and ℓ exchanged.) This we do as follows:

$$\begin{aligned} \tilde{\mathbb{E}} \langle a_j, w \rangle^t \langle a_\ell, w \rangle^t \|\mu - \bar{\mu}_j\|^{2t} &\leq Bn \tilde{\mathbb{E}} \langle a_j, w \rangle^{t-1} \langle a_\ell, w \rangle^t \|\mu - \bar{\mu}_j\|^t \\ &\leq Bn \left(\tilde{\mathbb{E}} \langle a_j, w \rangle^{t-2} \langle a_\ell, w \rangle^t \right)^{1/2} \cdot \left(\tilde{\mathbb{E}} \langle a_j, w \rangle^t \langle a_\ell, w \rangle^t \|\mu - \bar{\mu}_j\|^{2t} \right)^{1/2} \end{aligned}$$

where the first line is by assumption on $\tilde{\mathbb{E}}$ and the second is by pseudoexpectation Cauchy-Schwarz. Rearranging gives the inequality (4).

Now we use this to bound $\tilde{\mathbb{E}} \langle a_j, w \rangle^t \langle a_\ell, w \rangle^t$. By hypothesis,

$$\begin{aligned} &\tilde{\mathbb{E}} \langle a_j, w \rangle^t \langle a_\ell, w \rangle^t \\ &\leq \frac{1}{A} \tilde{\mathbb{E}} \langle a_j, w \rangle^t \langle a_\ell, w \rangle^t (\|\mu - \bar{\mu}_j\|^{2t} + \|\mu - \bar{\mu}_\ell\|^{2t}), \end{aligned}$$

which, followed by (4) gives

$$\begin{aligned} &\tilde{\mathbb{E}} \langle a_j, w \rangle^t \langle a_\ell, w \rangle^t \\ &\leq \frac{1}{A} \cdot B^2 n^2 \cdot \tilde{\mathbb{E}} \left[\langle a_j, w \rangle^{t-2} \langle a_\ell, w \rangle^t + \langle a_\ell, w \rangle^{t-2} \langle a_j, w \rangle^t \right]. \end{aligned}$$

Using $\langle a_j, w \rangle, \langle a_\ell, w \rangle \leq n$, we obtain

$$\tilde{\mathbb{E}} \langle a_j, w \rangle^t \langle a_\ell, w \rangle^t \leq \frac{2}{A} \cdot B^2 n^4 \cdot \tilde{\mathbb{E}} \langle a_j, w \rangle^{t-2} \langle a_\ell, w \rangle^{t-2}.$$

Finally, using Fact 5.10, the right side is at most

$$2B^2 n^4 / A \cdot \left(\tilde{\mathbb{E}} \langle a_j, w \rangle^t \langle a_\ell, w \rangle^t \right)^{(t-2)/t},$$

so cancelling terms we get

$$\left(\tilde{\mathbb{E}} \langle a_j, w \rangle^t \langle a_\ell, w \rangle^t \right)^{2/t} \leq \frac{2B^2 n^4}{A}.$$

Raising both sides to the $t/2$ power gives

$$\tilde{\mathbb{E}} \langle a_j, w \rangle^t \langle a_\ell, w \rangle^t \leq \frac{2^{t/2} B^t n^{2t}}{A^{t/2}},$$

and finally using Cauchy-Schwarz,

$$\tilde{\mathbb{E}} \langle a_j, w \rangle \langle a_\ell, w \rangle \leq \left(\tilde{\mathbb{E}} \langle a_j, w \rangle^t \langle a_\ell, w \rangle^t \right)^{1/t} \leq \frac{2n^2 B}{\sqrt{A}}. \quad \square$$

5.6 Rounding

In this section we state and analyze our second-moment round algorithm. As have discussed already, our SoS proofs in the mixture model setting are quite strong, meaning that the rounding algorithm is relatively naive.

The setting in this section is as follows. Let $n, m \in \mathbb{N}$ with $m \ll n$. There is a ground-truth partition of $[n]$ into m parts $S_1, \dots, S_m$ such that $|S_i| = (1 \pm \delta) \frac{n}{m}$. Let $A \in \mathbb{R}^{n \times n}$ be the 0/1 indicator matrix for this partition, so $A_{ij} = 1$ if $i, j \in S_\ell$ for some ℓ and is 0 otherwise. Let $M \in \mathbb{R}^{n \times n}$ be a matrix such that $\|M - A\| \leq \varepsilon n$, where $\|\cdot\|$ is the Frobenius norm. The algorithm takes M and outputs a partition $C_1, \dots, C_m$ of $[m]$ which makes few errors compared to $S_1, \dots, S_m$.

Algorithm 2 Rounding the second moment of $\tilde{\mathbb{E}}[ww^T]$

```

1: function ROUNDSECONDMOMENTS( $M \in \mathbb{R}^{n \times n}, E \in \mathbb{R}$ )
2:   Let  $S = [n]$ 
3:   Let  $v_1, \dots, v_n$  be the rows of  $M$ 
4:   for  $\ell = 1, \dots, m$  do
5:     Choose  $i \in S$  uniformly at random
6:     Let
7:        $C_\ell = \left\{ i' \in S : \|v_i - v_{i'}\|_2 \leq 2 \frac{n^{1/2}}{E} \right\}$ 
8:   end for
9:   return The clusters  $C_1, \dots, C_m$ .
10: end function
```

We will prove the following theorem.

THEOREM 5.11. *With notation as before Algorithm 2 with $E = m$, with probability at least $1 - \varepsilon^2 m^3$ Algorithm 2 returns a partition $C_1, \dots, C_m$ of $[n]$ such that (up to a permutation of $[m]$), $C_\ell = T_\ell \cup B_\ell$, where $T_\ell \subseteq S_\ell$ has size $|T_\ell| \geq |S_\ell| - \varepsilon^2 mn$ and $|B_\ell| \leq \varepsilon^2 mn$.*

To get started analyzing the algorithm, we need a definition.

Definition 5.1. For cluster S_j , let $a_j \in \mathbb{R}^n$ be its 0/1 indicator vector. If $i \in S_j$, we say it is E -good if $\|v_i - a_j\|_2 \leq \sqrt{n/E}$, and otherwise E -bad, where v_i is the i -th row of M . Let $I_g \subseteq [n]$ denote the set of E -good indices and I_b denote the set of E -bad indices. (We will choose E later.) For any $j = 1, \dots, k$, let $I_{g,j} = I_g \cap S_j$ denote the set of good indices from cluster j .

We have:

LEMMA 5.12. *Suppose E as in ROUNDSECONDMOMENTS satisfies $E \geq m/8$. Suppose that in iterations $1, \dots, m$, ROUNDSECONDMOMENTS has chosen only good vectors. Then, there exists a permutation $\pi : [m] \rightarrow [m]$ so that $C_\ell = I_{g,\pi(\ell)} \cup B_\ell$, where $B_\ell \subseteq I_b$ for all ℓ .*

PROOF. We proceed inductively. We first prove the base case. WLOG assume that the algorithm picks v_1 , and that v_1 is good, and is from component j . Then, for all $i \in I_{g,j}$, by the triangle inequality we have $\|v_i - v_1\|_2 \leq 2\frac{n^{1/2}}{B}$, and so $I_{g,j} \subseteq C_1$. Moreover, if $i \in I_{g,j'}$ for some $j' \neq j$, we have

$$\|v_i - v_1\|_2 \geq \|a'_j - a_j\|_2 - 2\frac{n^{1/2}}{E^{1/2}} \geq \frac{n^{1/2}}{\sqrt{m}} - 2\frac{n^{1/2}}{E^{1/2}} > 2\frac{n^{1/2}}{E^{1/2}},$$

and so in this case $i \notin C_1$. Hence $C_1 = I_{g,j} \cup B_1$ for some $B_1 \subseteq I_b$.

Inductively, suppose that if the algorithm chooses good indices in iterations $1, \dots, a-1$, then there exist distinct $j_1, \dots, j_{a-1}$ so that $C_\ell = I_{g,j_\ell} \cup B_\ell$ for $B_\ell \subseteq I_b$. We seek to prove that if the algorithm chooses a good index in iteration a , then $C_a = I_{g,j_a} \cup B_a$ for some $j_a \notin \{j_1, \dots, j_{a-1}\}$ and $B_a \subseteq I_b$. Clearly by induction this proves the Lemma. WLOG assume that the algorithm chooses v_1 in iteration a . Since by assumption 1 is good, and we have removed I_{g_ℓ} for $\ell = 1, \dots, a-1$, then $1 \in I_{g,j_a}$ for some $j_a \notin \{j_1, \dots, j_{a-1}\}$. Then, the conclusion follows from the same calculation as in the base case. $\square$

LEMMA 5.13. *There are at most $\varepsilon^2 En$ indices which are E -bad; i.e. $|I_b| \leq \varepsilon^2 En$.*

PROOF. We have

$$\begin{aligned} \varepsilon^2 n^2 &\geq \left\| M - \sum_{i \leq m} a_i a_i^\top \right\|_F^2 \geq \sum_j \sum_{i \in S_j \text{ bad}} \|v_i - a_j\|_2^2 \\ &\geq \frac{n}{E} |I_b|, \end{aligned}$$

from which the claim follows by simplifying. $\square$

This in turns implies:

LEMMA 5.14. *With probability at least $1 - \varepsilon^2 m^3$, the algorithm ROUNDSECONDMOMENTS chooses good indices in all k iterations.*

PROOF. By Lemma 5.13, in the first iteration the probability that a bad vector is chosen is at most $\varepsilon^2 E$. Conditioned on the event that in iterations $1, \dots, a$ the algorithm has chosen good vectors, then by Lemma 5.12, there is at least one j_a so that no points in I_{g,j_a} have been removed. Thus at least $(1 - \delta)n/m$ vectors remain, and in total there are at most $\varepsilon^2 En$ bad vectors, by Lemma 5.13. So, the probability of choosing a bad vector is at most $\varepsilon^2 Em$. Therefore, by the chain rule of conditional expectation and our assumption, the probability we never choose a bad vector is at least

$$(1 - \varepsilon^2 Em)^m$$

Choosing $E = m$ this is $(1 - \varepsilon^2 m^2)^m \geq 1 - \varepsilon^2 m^3$. as claimed. $\square$

Now Theorem 5.11 follows from putting together the lemmas.

ACKNOWLEDGEMENTS

The authors would like to thank David Steurer, Daniel Freund, Gautam Kamath, Pablo Parillo, and especially Aravindan Vijayaraghavan for some helpful conversations, and anonymous reviewers of this manuscript for their helpful remarks.

REFERENCES

- [1] Dimitris Achlioptas and Frank McSherry. 2005. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*. Springer, 458–469.
- [2] Joseph Anderson, Mikhail Belkin, Navin Goyal, Luis Rademacher, and James R. Voss. 2014. The More, the Merrier: the Blessing of Dimensionality for Learning Large Gaussian Mixtures. In *COLT (JMLR Workshop and Conference Proceedings)*, Vol. 35. JMLR.org, 1135–1164.
- [3] Sanjeev Arora and Ravi Kannan. 2005. Learning mixtures of separated nonspherical Gaussians. *Ann. Appl. Probab.* 15, 1A (2005), 69–92. <https://doi.org/10.1214/105051604000000512>
- [4] P. Awasthi, M. F. Balcan, and P. M. Long. 2014. The power of localization for efficiently learning linear separators with noise. In *STOC*. 449–458.
- [5] Pranjal Awasthi and Or Sheffet. 2012. Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*. Springer, 37–49.
- [6] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. 2014. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *CoRR abs/1408.2156* (2014).
- [7] Boaz Barak, Jonathan A. Kelner, and David Steurer. 2014. Rounding sum-of-squares relaxations. In *STOC*. ACM, 31–40.
- [8] Boaz Barak, Jonathan A. Kelner, and David Steurer. 2015. Dictionary Learning and Tensor Decomposition via the Sum-of-Squares Method. In *STOC*. ACM, 143–151.
- [9] Boaz Barak and Ankur Moitra. 2016. Noisy Tensor Completion via the Sum-of-Squares Hierarchy. In *COLT (JMLR Workshop and Conference Proceedings)*, Vol. 49. JMLR.org, 417–445.
- [10] Boaz Barak and David Steurer. 2014. Sum-of-squares proofs and the quest toward optimal algorithms. *CoRR abs/1404.5236* (2014).
- [11] Boaz Barak and David Steurer. 2017. The sos algorithm over general domains. <http://www.sumofsquares.org/public/lec-definitions-general.html>. (2017). [Online; accessed 11-1-2017].
- [12] Mikhail Belkin and Kaushik Sinha. 2010. Polynomial Learning of Distribution Families. In *FOCS*. IEEE Computer Society, 103–112.
- [13] T. Bernholt. 2006. *Robust Estimators are Hard to Compute*. Technical Report. University of Dortmund, Germany.
- [14] Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. 2014. Smoothed analysis of tensor decompositions. In *STOC*. ACM, 594–603.
- [15] S. C. Brubaker. 2009. Robust PCA and clustering in noisy mixtures. In *SODA 2009*. 1078–1087.
- [16] E. J. Candès, X. Li, Y. Ma, and J. Wright. 2011. Robust principal component analysis? *J. ACM* 58, 3 (2011), 11.
- [17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. 2016. Learning from Untrusted Data. *CoRR abs/1611.02315* (2016).
- [18] Yeshwanth Cherapanamjeri, Prateek Jain, and Praneeth Netrapalli. 2017. Thresholding based Efficient Outlier Robust PCA. In *COLT*.
- [19] Sanjoy Dasgupta. 1999. Learning mixtures of Gaussians. In *Foundations of computer science, 1999. 40th annual symposium on*. IEEE, 634–644.
- [20] Sanjoy Dasgupta and Leonard Schulman. 2007. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learning Research* 8, Feb (2007), 203–226.
- [21] Constantinos Daskalakis and Gautam Kamath. 2014. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Conference on Learning Theory*.
- [22] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. 2017. Ten steps of EM suffice for mixtures of two Gaussians. *Conference on Learning Theory* (2017).
- [23] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. 2016. Robust Estimators in High Dimensions without the Computational Intractability. In *FOCS*. IEEE Computer Society, 655–664.
- [24] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. 2017. Being robust (in high dimensions) can be practical. In *ICML*.
- [25] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. 2017. Robustly Learning a Gaussian: Getting Optimal Error, Efficiently. In *Symposium on Discrete Algorithms*.
- [26] Ilias Diakonikolas, Daniel Kane, and Alistair Stewart. 2017. personal communication. (2017).
- [27] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. 2016. Statistical Query Lower Bounds for Robust Estimation of High-dimensional Gaussians and Gaussian Mixtures. *arXiv preprint arXiv:1611.03473* (2016).
- [28] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. 2017. Learning Geometric Concepts with Nasty Noise. *arXiv preprint arXiv:1707.01242* (2017).
- [29] Jon Feldman, Rocco A. Servedio, and Ryan O'Donnell. 2006. PAC Learning Axis-Aligned Mixtures of Gaussians with No Separation Assumption. In *COLT (Lecture Notes in Computer Science)*, Vol. 4005. Springer, 20–34.
- [30] Rong Ge, Qingqing Huang, and Sham M. Kakade. 2015. Learning mixtures of Gaussians in high dimensions [extended abstract]. In *STOC'15—Proceedings of*

- the 2015 ACM Symposium on Theory of Computing. ACM, New York, 761–770.
- [31] Rong Ge and Tengyu Ma. 2015. Decomposing overcomplete 3rd order tensors using sum-of-squares algorithms. *arXiv preprint arXiv:1504.05287* (2015).
 - [32] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. 1986. *Robust statistics. The approach based on influence functions*. Wiley New York.
 - [33] Moritz Hardt and Eric Price. 2015. Tight Bounds for Learning a Mixture of Two Gaussians. In *STOC*. ACM, 753–760.
 - [34] Samuel B Hopkins, Pravesh Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. 2017. The power of sum-of-squares for detecting hidden structures. *Symposium on Foundations of Computer Science* (2017).
 - [35] Samuel B. Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. 2016. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *STOC*. ACM, 178–191.
 - [36] Samuel B. Hopkins, Jonathan Shi, and David Steurer. 2015. Tensor principal component analysis via sum-of-square proofs. In *COLT (JMLR Workshop and Conference Proceedings)*, Vol. 40. JMLR.org, 956–1006.
 - [37] Daniel Hsu and Sham M. Kakade. 2013. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *ITCS'13—Proceedings of the 2013 ACM Conference on Innovations in Theoretical Computer Science*. ACM, New York, 11–19.
 - [38] Peter J Huber. 1964. Robust estimation of a location parameter. *The Annals of Mathematical Statistics* 35, 1 (1964), 73–101.
 - [39] D. S. Johnson and F. P. Preparata. 1978. The densest hemisphere problem. *Theoretical Computer Science* 6 (1978), 93–107.
 - [40] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. 2010. Efficiently learning mixtures of two Gaussians. In *STOC*. ACM, 553–562.
 - [41] M. J. Kearns and M. Li. 1993. Learning in the Presence of Malicious Errors. 22, 4 (1993), 807–837.
 - [42] A. Klivans, P. Long, and R. Servedio. 2009. Learning Halfspaces with Malicious Noise. (2009).
 - [43] Pravesh Kothari and Jacob Steinhardt. 2017. Better Clustering via Relaxed Tensor Norms. personal communication. (2017).
 - [44] Pravesh Kothari and David Steurer. 2017. Outlier-robust moment estimation via sum-of-squares. personal communication. (2017).
 - [45] Amit Kumar and Ravindran Kannan. 2010. Clustering with Spectral Norm and the k-Means Algorithm. In *FOCS*. IEEE Computer Society, 299–308.
 - [46] Kevin A. Lai, Anup B. Rao, and Santosh Vempala. 2016. Agnostic Estimation of Mean and Covariance. In *FOCS*. IEEE Computer Society, 665–674.
 - [47] G. Lerman, M. B. McCoy, J. A. Tropp, and T. Zhang. 2012. Robust computation of linear models, or How to find a needle in a haystack. *CoRR* abs/1202.4044 (2012). <http://arxiv.org/abs/1202.4044>
 - [48] Jerry Li and Ludwig Schmidt. 2017. Robust and Proper Learning for Mixtures of Gaussians via Systems of Polynomial Inequalities. In *Conference on Learning Theory*.
 - [49] Tengyu Ma, Jonathan Shi, and David Steurer. 2016. Polynomial-Time Tensor Decompositions with Sum-of-Squares. In *FOCS*. IEEE Computer Society, 438–446.
 - [50] Geoffrey McLachlan and David Peel. 2004. *Finite mixture models*. John Wiley & Sons.
 - [51] Dustin G Mixon, Soledad Villar, and Rachel Ward. 2017. Clustering subgaussian mixtures by semidefinite programming. *Information and Inference: A Journal of the IMA* (2017), iax001.
 - [52] Ankur Moitra and Gregory Valiant. 2010. Settling the Polynomial Learnability of Mixtures of Gaussians. In *FOCS*. IEEE Computer Society, 93–102.
 - [53] Ryan O'Donnell. 2017. SOS is not obviously automatizable, even approximately. (2017).
 - [54] Ryan O'Donnell and Yuan Zhou. 2013. Approximability and proof complexity. In *SODA*. SIAM, 1537–1556.
 - [55] Karl Pearson. 1894. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*. A 185 (1894), 71–110.
 - [56] Aaron Potechin and David Steurer. 2017. Exact tensor completion with sum-of-squares. *CoRR* abs/1702.06237 (2017).
 - [57] Prasad Raghavendra and Benjamin Weitz. 2017. On the Bit Complexity of Sum-of-Squares Proofs. *CoRR* abs/1702.05139 (2017).
 - [58] Oded Regev and Aravindan Vijayaraghavan. 2017. On learning mixtures of well-separated Gaussians. In *Symposium on Foundations of Computer Science*.
 - [59] Tselil Schramm and David Steurer. 2017. Fast and robust tensor decomposition with applications to dictionary learning. *Conference on Learning Theory* (2017).
 - [60] R. Servedio. 2003. Smooth boosting and learning with malicious noise. *JMLR* 4 (2003), 633–648.
 - [61] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. 2017. Resilience: A criterion for learning in the presence of arbitrary outliers. *ITCS*.
 - [62] Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. 2014. Near-optimal-sample estimators for spherical gaussian mixtures. In *Advances in Neural Information Processing Systems*. 1395–1403.
 - [63] D Michael Titterton, Adrian FM Smith, and Udi E Makov. 1985. *Statistical analysis of finite mixture distributions*. Wiley.
 - [64] J.W. Tukey. 1975. Mathematics and picturing of data. In *Proceedings of ICM*, Vol. 6. 523–531.
 - [65] John W Tukey. 1975. Mathematics and the picturing of data. In *Proceedings of the international congress of mathematicians*, Vol. 2. 523–531.
 - [66] Leslie G. Valiant. 1985. Learning Disjunction of Conjunctions. In *IJCAI*. Morgan Kaufmann, 560–566.
 - [67] Santosh Vempala and Grant Wang. 2002. A Spectral Algorithm for Learning Mixtures of Distributions. In *FOCS*. IEEE Computer Society, 113.
 - [68] CF Jeff Wu. 1983. On the convergence properties of the EM algorithm. *The Annals of statistics* (1983), 95–103.
 - [69] Ji Xu, Daniel J Hsu, and Arian Maleki. 2016. Global analysis of expectation maximization for mixtures of two gaussians. In *Advances in Neural Information Processing Systems*. 2676–2684.
 - [70] T. Zhang and G. Lerman. 2014. A Novel M-estimator for Robust PCA. *J. Mach. Learn. Res.* 15, 1 (Jan. 2014), 749–808.