2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)

Outlier-Robust Clustering of Gaussians and Other Non-Spherical Mixtures

Ainesh Bakshi
CMU
abakshi@cs.cmu.edu

Ilias Diakonikolas
UW Madison
ilias@cs.wisc.edu

Samuel B. Hopkins
UC Berkeley
hopkins@berkeley.edu

Daniel Kane
UC San Diego
dakane@ucsd.edu

Sushrut Karmalkar
UT Austin
sushrutk@cs.utexas.edu

Pravesh K. Kothari
CMU
praveshk@cs.cmu.edu

***Abstract*—We give the first outlier-robust efficient algorithm for clustering a mixture of k statistically separated d-dimensional Gaussians (k-GMMs). Concretely, our algorithm takes input an ϵ-corrupted sample from a k-GMM and outputs an approximate clustering that misclassifies at most $k^{O(k)}(\epsilon+\eta)$ fraction of the points whenever every pair of mixture components are separated by $1-\exp(-poly(k/\eta))$ in total variation distance. This is the statistically weakest possible notion of separation and allows, for e.g., clustering of mixtures with components with the same mean with covariances differing in a single unknown direction or separated in Frobenius distance. The running time of our algorithm is $d^{poly(k/\eta)}$. Such results were not known prior to our work, even for $k=2$.**

More generally, our algorithms succeed for mixtures of any distribution that satisfies two well-studied analytic assumptions - sum-of-squares certifiable *hypercontractivity* and *anti-concentration*. As an immediate corollary, they extend to clustering mixtures of arbitrary affine transforms of the uniform distribution on the d-dimensional unit sphere. Even the information theoretic clusterability of separated distributions satisfying our analytic assumptions was not known and is likely to be of independent interest.

Our algorithms build on the recent flurry of work relying on *certifiable anti-concentration* first introduced in [1], [2]. Our techniques expand the sum-of-squares toolkit to show robust *certifiability* of TV-separated Gaussian clusters in data. This involves giving a low-degree *sum-of-squares proof* of statements that relate parameter (i.e. mean and covariances) distance to total variation distance by relying only on hypercontractivity and anti-concentration.

***Keywords*-Robust statistics; Gaussian Mixture Models; Sum of Squares Method.**

Remark .1 (Merging Bakshi and Kothari [3] and Diakonikolas, Hopkins, Kane and Karmalkar [4]). This extended abstract serves as a brief overview of concurrent and independent works Bakshi and Kothari [3] and Diakonikolas, Hopkins, Kane and Karmalkar [4]. We point the reader to the respective full versions of these papers, both available on arXiv,[1] for a comprehensive treatment of the results. In this overview, we mainly follow the presentation of Bakshi and Kothari; we note differences to the results and techniques of Diakonikolas et al. when necessary.

I. Introduction

A flurry of recent work has focused on designing outlier-robust efficient algorithms for statistical estimation for basic tasks such as estimating mean, covariance [5]–[13], moment tensors [8] of distributions, regression [14]–[19], and clustering of spherical mixtures [14], [20], [21]. This progress (see [22] for a recent survey) has come via fundamentally new algorithmic techniques such as agnostic filtering [6] and robust-learning frameworks based on the sum-of-squares method in both the strong contamination [8], [20], [21] and list-decodable learning models [2], [18], [19], [23].

In this paper, we extend this line of work by studying outlier-robust *clustering* of mixtures of distributions that exhibit mean or covariance separation. As a corollary, we obtain a poly-time outlier-robust algorithm for clustering mixtures of k-Gaussians (k-GMMs) when each pair of components is separated in total variation (TV)[2] distance. This is the information-theoretically weakest notion of separation, allows components of same mean but variances differing in an unknown direction[3] or covariances separated in *relative* Frobenius distance (see Fig 1) and includes well-studied problems such as *mixed linear regression* and *subspace clustering* as special cases.

The Gaussian Mixture Model has been the subject of a century-old line of research beginning with Pearson [24]. A k-GMM $\sum_{r\leq k} p_r \mathcal{N}(\mu(r),\Sigma(r))$ is a probability distribution sampled by choosing a component $r\sim[k]$ with probability p_r and outputting a sample from the Gaussian distribution with mean $\mu(r)$ and covariance $\Sigma(r)$. In the k-GMM learning problem, the goal is to output an approximate *clustering* of the input sample or estimate the parameters (the mean and covariances) of the components. Progress on provable

[1] For [3] by Bakshi and Kothari, see https://arxiv.org/abs/2005.02970, and for [4] by Diakonikolas, Hopkins, Kane, and Karmalkar, see https://arxiv.org/abs/2005.06417.

[2] The TV distance between distributions with PDFs p,q is defined as $\frac{1}{2}\int_{-\infty}^{\infty}|p(x)-q(x)|dx$.

[3] As an interesting example, consider the case of subspace clustering: mixture of standard Gaussians restricted to unknown distinct subspaces. The components have a TV distance of 1 regardless of how close the subspaces are and thus satisfy our assumptions.

2575-8454/20/$31.00 ©2020 IEEE
DOI 10.1109/FOCS46700.2020.00023

algorithms for learning k-GMMs began with the influential work of Dasgupta [25] followed up by [26]–[29] yielding clustering algorithms that succeed under various separation assumptions. These assumptions, however, do not capture natural separated instances of Gaussians (e.g., see (b) or (c) in Fig 1). A more general approach [30]–[32] circumvents clustering altogether by giving an efficient algorithm (time $\sim d^{poly(k)}$) for parameter estimation without any separation assumptions.

In this work, we focus on the important special case of this problem where the mixture components are "separated". Various notions of separation have been used in the literature. Here we focus on the following definition: We say that a k-mixture of Gaussians is *separated* if the *overlap* between any pair of components P, Q (i.e., $1 - \text{tv}(P, Q)$, where $\text{tv}(P, Q)$ is the total variation distance between P and Q) is a small constant — independent of the dimension. We note that this is qualitatively the weakest possible separation assumption under which accurate clustering of the components is information-theoretically possible — even without outliers.

The preceding discussion motivates the following question, whose resolution is the main result of this work:

Question I.1. Is there a $poly(d)$-time robust learning algorithm for a mixture of any constant number of (or even two) arbitrary *separated* Gaussians on $\mathbb{R}^d$?

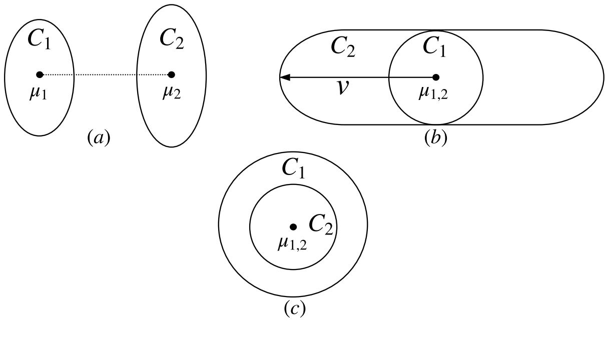

Figure 1. (a) Mean Separation (b) Spectral Separation (c) Relative Frobenius Separation

A. Our Results

Our main result is an efficient algorithm for outlier-robust clustering of k-GMMs whenever every pair of components of the mixture are separated in total variation distance. Formally, our algorithms work in the *strong contamination* model studied in the bulk of the prior works on robust estimation where an adversary changes an arbitrary, potentially adversarially chosen ϵ-fraction of the input sample before passing it on to the algorithm.

Theorem I.2 (Main Result, Outlier-Robust Clustering of k-GMMs, [3])**.** *Fix $\eta, \epsilon > 0$. Let $\mathcal{D}_r = \mathcal{N}(\mu(r), \Sigma(r))$ for $r \leq k$ be k-Gaussians such that $d_{TV}(\mathcal{D}_r, \mathcal{D}_{r'}) \geq 1 - \exp(-poly(k/\eta))$ whenever $r \neq r'$. Then, there exists an algorithm that takes input an ϵ-corruption Y of a sample $X = C_1 \cup C_2 \cup \ldots \cup C_k$ of size n, with equal sized clusters C_i drawn i.i.d. from $\mathcal{D}_i$ for each $r \leq k$, and with probability ≥ 0.99, outputs an approximate clustering $Y = \hat{C}_1 \cup \hat{C}_2 \cup \ldots \cup \hat{C}_k$ satisfying $\min_{i \leq k} \frac{|\hat{C}_i \cap C_i|}{|C_i|} \geq 1 - O(k^{2k})(\epsilon + \eta)$. The algorithm succeeds whenever $n \geq d^{O(poly(k/\eta))}$ and runs in time $n^{O(poly(k/\eta)))}$.*

We can use off-the-shelf robust estimators for mean and covariance of Gaussians([6]) in order to get statistically optimal estimates of the mean and covariances of the target k-GMM.

Corollary I.3 (Parameter Recovery from Clustering, [3])**.** *In the setting of Theorem I.2, with the same running time, sample complexity and success probability, our algorithm can output $\{\hat{\mu}(r), \hat{\Sigma}(r)\}_{r \leq k}$ such that for some permutation $\pi : [k] \to [k]$, $d_{TV}(\mathcal{N}(\mu(r), \Sigma(r)), \mathcal{N}(\hat{\mu}(\pi(r)), \hat{\Sigma}(\pi(r))) \leq \tilde{O}(k^{2k}(\epsilon + \eta))$.*

Remark I.4 (Differences to Diakonikolas et al [4]). The above Theorem I.2 and Corollary I.3, along with the corresponding algorithms and analyses, appear in this form in [3]. Diakonikolas et al. [4] obtain similar results with the following differences. The algorithm of [4] requires $d^{F(k)}/poly(\epsilon)$ samples and runs in time $n^{F(k)}$, for $F(k)$ at most an exponential tower of height $poly(k)$, and outputs a Gaussian mixture model whose TV distance to $\frac{1}{k}\sum_{r=1}^{k} \mathcal{N}(\mu(r), \Sigma(r))$ is at most $\tilde{O}(\epsilon)$, so long as $\epsilon \leq 1/F(k)$ and $\min_{r \neq r'} d_{TV}(\mathcal{N}(\mu(r), \Sigma(r)), \mathcal{N}(\mu(r'), \Sigma(r'))) \leq 1/F(k)$.

By combining the approaches of [3], [4], we can obtain the following improvement to the sample complexity of Corollary I.3: for every $\epsilon \leq k^{-O(k)}$ there is an algorithm requiring $n \geq d^{k^{O(k)}}/poly(\epsilon)$ ϵ-corrupted samples from a k-GMM with pairwise separation $d_{TV}(\mathcal{N}(\mu(r), \Sigma(r)), \mathcal{N}(\mu(r'), \Sigma(r'))) \geq 1 - 2^{-k^{O(k)}}$ and running in time $n^{k^{O(k)}}$ which returns a hypothesis k-GMM which is accurate to total variation distance $\tilde{O}(\epsilon)$. It is a fascinating open problem to understand whether the doubly-exponential sample complexity and running time can be brought down to single-exponential.

These are the first outlier-robust algorithms that work for clustering k-GMMs under information-theoretically optimal separation assumptions. Such results were not known even for $k = 2$. To discuss the bottlenecks in prior works, it is helpful to use following consequence of two Gaussians with means $\mu(1), \mu(2)$ and covariances $\Sigma(1), \Sigma(2)$ being at a TV distance $\geq 1 - \exp(-O(\Delta^2))$ in terms of the distance between their parameters:

Definition I.5 (Δ-Separated Mixture Model)**.** An equi-weighted mixture $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ with parameters $\{\mu(i), \Sigma(i)\}_{i\le k}$ is Δ-separated if for every pair of distinct components i, j, one of the following three conditions hold ($\Sigma^{\dagger/2}$ is the square root of pseudo-inverse of Σ):

1) **Mean-Separation:** $\exists v \in \mathbb{R}^d$ such that $\langle \mu(i) - \mu(j), v\rangle^2 > \Delta^2 v^\top (\Sigma(i) + \Sigma(j))v$,
2) **Spectral-Separation:** $\exists v \in \mathbb{R}^d$ such that $v^\top \Sigma(i) v > \Delta v^\top \Sigma(j) v$,
3) **Relative-Frobenius Separation:**[4] $\Sigma(i)$ and $\Sigma(j)$ have the same range space and

$$\left\|\Sigma(i)^{\dagger/2}\Sigma(j)\Sigma(i)^{\dagger/2} - I\right\|_F^2 > \Delta^2 \left\|\Sigma(i)^{\dagger/2}\Sigma(j)^{1/2}\right\|_{op}^4$$

The key bottleneck for known algorithms was handling separation in cases 2 and 3 above. Further, we note that no algorithm prior to our work (such as [30], [33]) could handle rank deficient covariances, even non-robustly. In particular, prior work obtains parameter recovery in non-affine invariant norms. However, two non-overlapping subspaces can be arbitrarily close in any such norm, while still being TV distance 1 apart.

Beyond Gaussians. Our results apply more generally to mixture models where each component distribution $\mathcal{D}$ satisfies two natural and well-studied analytic conditions: *hypercontractivity* of degree 2 polynomials and *anti-concentration* of all directional marginals. Our algorithmic results hold for distributions (e.g. Gaussians and affine transforms of uniform distribution on the unit sphere) that admit efficiently verifiable analogs of these properties.

Definition I.6 (Certifiable Hypercontractivity)**.** An isotropic distribution $\mathcal{D}$ on $\mathbb{R}^d$ is said to be h-certifiably C-hypercontractive if there's a degree h sum-of-squares proof of the following unconstrained polynomial inequality in $d \times d$ matrix-valued indeterminate Q:

$$\left\{\mathbb{E}_{x\sim\mathcal{D}}\left(x^\top Q x\right)^h \le (Ch)^h \left(\mathbb{E}_{x\sim\mathcal{D}}\left(x^\top Q x\right)^2\right)^{h/2}\right\}$$

A set of points $X \subseteq \mathbb{R}^d$ is said to be C-certifiably hypercontractive if the uniform distribution on X is h-certifiably C-hypercontractive.

Hypercontractivity is an important notion in high-dimensional probability and analysis on product spaces [34]. Kauers, O'Donnell, Tan and Zhou [35] showed certifiable hypercontractivity of Gaussians and more generally product distributions with subgaussian marginals. Certifiable hypercontractivity strictly generalizes the better known *certifiable subgaussianity* property (studied first in [8]) that controls higher moments of linear polynomials.

Certifiable anti-concentration.: In contrast to subgaussianity, anti-concentration forces *lower-bounds* of the form $\Pr[\langle x, v\rangle^2 \ge \delta \|v\|_2^2] \ge \delta'$ for all directions v. Certifiable anti-concentration was recently introduced in independent works of Karmalkar, Klivans and Kothari [18] and Raghavendra and Yau [19] and later used [2], [36] for the related problems of list-decodable linear regression and subspace recovery[5].

Following Karmalkar, Klivans and Kothari [18], we formulate certifiable anti-concentration via a univariate, even polynomial $p_{\delta,\Sigma}$ that uniformly approximates the 0-1 core-indicator $\mathbf{1}(\langle x, v\rangle^2 \ge \delta v^\top \Sigma v)$ over a large enough interval around 0. Let $q_{\delta,\Sigma}(x, v)$ be a multivariate (in v) polynomial defined by $q_{\delta,\Sigma}(x,v) = \left(v^\top \Sigma v\right)^{2s} p_{\delta,\Sigma}\left(\frac{\langle x,v\rangle}{\sqrt{v^\top \Sigma v}}\right)$.Since $p_{\delta,\Sigma}$ is an even polynomial, $q_{\delta,\Sigma}$ is a polynomial in v.

Definition I.7 (Certifiable Anti-Concentration)**.** An mean 0 distribution D with covariance Σ is $2s$-certifiably $(\delta, C\delta)$-anti-concentrated if for $q_{\delta,\Sigma}(x, v)$ defined above, there exists a degree $2s$ sum-of-squares proof of the following two unconstrained polynomial inequalities in indeterminate v:

$$\begin{aligned}&\left\{\langle x, v\rangle^{2s} + \delta^{2s} q_{\delta,\Sigma}(x,v)^2 \ge \delta^{2s}\left(v^\top \Sigma v\right)^{2s}\right\}, \\ &\left\{\mathbb{E}_{x\sim D}\, q_{\delta,\Sigma}(x,v)^2 \le C\delta\left(v^\top \Sigma v\right)^{2s}\right\}\end{aligned} \tag{1}$$

An isotropic subset $X \subseteq \mathbb{R}^d$ is $2s$-certifiably $(\delta, C\delta)$-anti-concentrated if the uniform distribution on X is $2s$-certifiably $(\delta, C\delta)$-anti-concentrated.

Remark I.8. For natural examples, $s(\delta) \le 1/\delta^c$ for some fixed constant c. For e.g., $s(\delta) = O(\frac{1}{\delta^2})$ for standard Gaussian distribution and the uniform distribution on the unit sphere (see [18] and [36]). To simplify notation, we will assume $s(\delta) \le poly(1/\delta)$ in the statement of our results.

Our general result gives an outlier-robust clustering algorithm for separated mixtures of *reasonable* distributions, i.e., one that satisfies both certifiable hypercontractivity and anti-concentration. Even the information-theoretic (and without outliers, i.e., $\epsilon = 0$) clusterability of such distributions was not known prior to our work.

Theorem I.9 (Outlier-Robust Clustering of Separated Mixtures)**.** *Fix $\eta > 0, \epsilon > 0$. Let $\mathcal{D}_r$ be $s(\delta)$-certifiably δ-anti-concentrated distributions for all $\delta > 0$ and has h-certifiably C-hypercontractive degree 2 polynomials for all h such that the mixture of $\mathcal{D}_r$ is Δ-separated. Then, there exists an algorithm that takes input an ϵ-corruption Y of a sample $X = C_1 \cup C_2 \cup \ldots C_k$ of size n, with true clusters C_i drawn i.i.d. $\mathcal{D}_r$ for each $r \le k$, and outputs an approximate clustering $Y = \hat{C}_1 \cup \hat{C}_2 \cup \ldots \cup \hat{C}_k$ satisfying*

[4] Unlike the other two distances, relative Frobenius distance is meaningful only for high-dimensional Gaussians. As an illustrative example, consider two 0 mean Gaussians with covariances $\Sigma_1 = I$ and $\Sigma_2 = (1 + \Theta(1/\sqrt{d}))I$. Then, for large enough d, the parameters are separated in relative Frobenius distance but not spectral or mean distance.

[5] List-decodable versions of these problems generalize the "mixture" variants - mixed linear regression and subspace clustering - that are easily seen to be special cases of mixtures of k-Gaussians with TV separation 1.

$\min_{i\leq k} \frac{|\hat{C}_i \cap C_i|}{|C_i|} \geq 1 - O(k^{2k})(\epsilon+\eta)$. *The algorithm succeeds with probability at least* 0.99 *over the draw of the original sample* X *whenever* $n \geq d^{O(poly(k/\eta))}$ *and runs in time* $n^{O(poly(k/\eta)))}$ *whenever* $\Delta \geq poly(k/\eta)^k$.

Robust Covariance Estimation. We give an outlier-robust algorithm for covariance estimation for all certifiably hypercontractive distributions.

Theorem I.10 (Robust Parameter Covariance Estimation for Certifiably Hypercontractive Distributions)**.** *Fix an* $\epsilon > 0$ *small enough fixed constant so that* $Ct\epsilon^{1-4/t} \ll 1$[6]. *For every even* $t \in \mathbb{N}$, *there's an algorithm that takes input* Y *be an* ϵ*-corruption of a sample* X *of size* n *from a reasonable distribution with unknown mean* μ_* *and covariance* Σ_* *respectively and outputs an estimate* $\hat{\mu}$ *and* $\hat{\Sigma}$ *satisfying:*

1) $\left\|\Sigma^{-1/2}(\mu_* - \hat{\mu})\right\|_2 \leq O(Ct)^{1/2}\epsilon^{1-1/t}$,
2) $(1-\eta)\Sigma_* \preceq \hat{\Sigma} \preceq (1+\eta)\Sigma_*$ *for* $\eta \leq O(Ck)\epsilon^{1-2/t}$, *and,*
3) $\left\|\Sigma_*^{-1/2}\hat{\Sigma}\Sigma_*^{-1/2} - I\right\|_F \leq (Ct)O(\epsilon^{1-1/t})$.

In particular, letting $t = O(\log(1/\epsilon))$ *results in the error bounds of* $\tilde{O}(\epsilon)$ *in all the three inequalities above.*

The first two guarantees above were shown by Kothari and Steurer [8] for all certifiably subgaussian distributions. [8] also observed (see last paragraph of page 6 for a counter example) that it is provably impossible to obtain dimension-independent error bounds in relative Frobenius distance assuming only certifiable subgaussianity (see Definition 5.1 in [8]). We prove that under the stronger assumption of certifiable *hypercontractivity*, we can indeed obtain dimension-independent, information-theoretically optimal (for e.g. for Gaussians) error guarantees in relative Frobenius error. Prior works either obtained the weaker spectral error guarantee (that incurs a loss of $\sqrt{d}$ factor when translating into relative Frobenius distance) [5], [8] or worked only for Gaussians [6][7].

Combining this theorem with our clustering results above yields:

Corollary I.11 (Parameter Recovery from Clustering, General Case)**.** *In the setting of either Theorem I.9, there's an algorithm with same bounds on running time and sample complexity, that with probability at least* 0.99, *outputs* $\{\hat{\mu}(r), \hat{\Sigma}(r)\}_{r\leq k}$ *such that for some permutation* $\pi : [k] \to [k]$, *for every* i, $\hat{\mu}(\pi(i)), \hat{\Sigma}(\pi(i))$ *is* Δ*-close to* μ, Σ *in the three distances defined in Definition I.5 for* $\Delta = \tilde{O}(k^{O(k)}(\epsilon+\eta))$.

Applications. Two important special cases of mixtures of separated reasonable distributions are noiseless *mixed linear regression* where we are given samples generated as $y = \langle \ell, x\rangle$ where x is drawn from $\mathcal{N}(0, I_d)$ and ℓ is chosen uniformly from an unknown list $(\ell_1, \ell_2, \ldots, \ell_k)$ and *subspace clustering* - where the input is a mixture of isotropic Gaussians restricted to a k unknown subspaces. There's extensive work [37]–[47] on both these problems in signal processing and machine learning with recent push [45], [48] in TCS to obtain efficient algorithms with provable guarantees. Both these cases are immediately seen as mixtures with pairwise separation of ∞ (for Gaussians, this is equivalent to TV separation of 1). Thus, we immediately obtain efficient outlier-robust algorithms for these problems.

II. Related Work

In this subsection, we provide a detailed summary of the most relevant prior work.

The known non-robust parameter estimation algorithms for k-mixtures of arbitrary Gaussians [33], [49] work by reducing the problem to a collection of univariate problems by taking many random one-dimensional projections and piecing together the information obtained from all these projections. The univariate problem is solved using the method of moments. Unfortunately, the accuracy required for each univariate problem for this approach to work is inverse polynomial in the dimension d, which is information-theoretically impossible to achieve in the presence of even a sub-constant fraction of outliers. In summary, this approach is highly non-robust.

In the robust setting, significant progress has been made for mixtures of *spherical* Gaussians. The work of [50] gave a robust density estimation algorithm for for a mixture of (any constant number of) spherical Gaussians. More recently, [51]–[53] obtained efficient robust parameter estimation algorithm for mixtures of spherical Gaussians under near-optimal separation assumptions. Our SoS-based clustering framework is identical to that of [51]. Our contribution lies in our construction of a low-degree identifiability proof for the clusters that can handle arbitrary separated Gaussians.

It should be emphasized that the identifiability proofs in [51], [52] are (essentially) constrained to the spherical setting and in particular cannot even handle (non-robust) parameter estimation of two hyperplane separated Gaussians. This is due to their definition of a good cluster that only imposes upper bounds on the low-degree moments of the clusters.

We note that [54] gave an SQ lower bound, which provides evidence that an exponential dependence on k is required for the sample complexity and runtime of our problem, even for the hyperplane separated case without outliers.

III. Overview

In this section, we given an informal overview of our approach and main ideas. All of our conceptual ideas appear

[6] This notation means that we needed $Ct\epsilon^{1-2/t}$ to be at most c_0 for some absolute constant $c_0 > 0$

[7] We note that the algorithm of [6] for Gaussian distributions works in fixed polynomial time to obtain $\tilde{O}(\epsilon)$ error-estimate of the covariance in relative Frobenius distance whereas our algorithm works more generally for all certifiably hypercontractive distributions but runs in time $d^{O(log^2(1/\epsilon))}$.

in obtaining a clustering algorithm in the non-robust (without outliers) setting. So we will restrict ourselves to this setting for most of this section.

Formally, our results hold for Δ-separated (in the sense of Definition I.5) mixtures of all *reasonable* distributions defined below.

Definition III.1 (Reasonable Distributions)**.** An isotropic (i.e. mean 0 and I-covariance) distribution $\mathcal{D}$ on $\mathbb{R}^d$ is *reasonable* if it satisfies the following two properties:

1) *Certifiable Anti-Concentration Under* 4*-wise Convolutions*: The distribution of $x \pm y \pm z \pm w$ for independent copies $x, y, z, w \sim \mathcal{D}$ is certifiably $(\delta, C\delta)$ anti-concentrated for all $\delta > 0$.
2) *Certifiable Hypercontractivity Under* 4*-wise Convolutions*: The distribution of $x \pm y \pm z \pm w$ for independent $x, y, z, w \sim \mathcal{D}$ has certifiably hypercontractive degree 2 polynomials.

Observe that if $\mathcal{D}$ has h-certifiably C-hypercontractive degree 2 polynomials then it is also h-certifiably C-subgaussian. For any $\mu, \Sigma \succ 0$, we denote $\mathcal{D}(\mu, \Sigma)$ to be the distribution of the random variable $\Sigma^{1/2}x + \mu$ where $x \sim \mathcal{D}$.

In Section 8 of [3], we prove that Gaussian distributions and affine transforms of uniform distribution on the unit sphere are reasonable distributions.

Setup. The input to our algorithm is a sample X of size n from an equi-weighted mixture of $\{\mathcal{D}(\mu(r), \Sigma(r))\}_{r \leq k}$ for some reasonable distribution $\mathcal{D}$. Let $X = C_1 \cup C_2 \cup \ldots C_k$ be the partition of X into true clusters unknown to the algorithm. We follow the high-level approach of using low-degree sum-of-squares proofs of *certifiability*[8] to design efficient algorithms.

The two key parts of our proofs are 1) giving a low-degree sum-of-squares proof of certifiability of approximate clusters and 2) a recursive clustering based on rounding pseudo-distributions. We discuss the high-level ideas behind both these pieces below.

Certifying Purported Clusters. In this approach, we ignore the algorithmic issues and focus simply on the issue of how to *certify* that a given subset $\hat{C} \subseteq X$ - described by an associated set of indicator variables $w_1, w_2, \ldots, w_n$ of the samples included in $\hat{C}$ - is (close to) a true cluster C_r for some $r \leq k$. Let $w(C_r) = \frac{|\hat{C} \cap C_r|}{|C_r|}$ for every r.

By standard concentration arguments, for n large enough, the uniform distribution on C_i for each i is itself reasonable - that is, it satisfies the conditions in Def III.1. Further, the parameters of each C_r are close to the true parameters $\mu(r), \Sigma(r)$. Instead of introducing new notation, we will simply assume that $\mu(r), \Sigma(r)$ are the mean and covariances of C_r (instead of the distribution that generates C_r). This slight abuse of notation doesn't meaningfully change our results or techniques.

Finally, another simple but useful observation is that for distributions that are uniform on subsets of $A, B \subseteq X$ of size n/k, the total variation distance equals $1 - (k/n)|A \cap B|$. In particular, large TV distance corresponds to small intersection and vice-versa.

The only properties we know of the true clusters is that they are of size n/k and that uniform distributions on them are reasonable distributions. Thus, the natural checks we can perform on $\hat{C}$ is to simply verify the properties of being certifiably hypercontractive and anti-concentrated.

Since we check only the properties that a true cluster C_i would satisfy, it's clear that the true clusters should pass our checks. Thus, we can focus on proving *soundness* of our test: if $\hat{C}$ passes the checks we made, then it must be close to one of the true clusters C_is. The key "bad case" for us to rule out is when $w(C_r)$ and $w(C_{r'})$ are both large for some $r \neq r'$. In that case, the set $\hat{C}$ indicated by w cannot be close to any single cluster C_i.

Indeed, bulk of our analysis goes into showing that for every $r \neq r'$, $w(C_r)w(C_{r'})$ must be small whenever w passes our checks above. This immediately implies that $w(C_r)$ and $w(C_{r'})$ cannot simultaneously be large. We call such results *simultaneous intersection bounds* because they control the simultaneous intersection of $\hat{C}$ with C_r and $C_{r'}$.

A. Enter TV vs Parameter Distance Lemmas

When $w(C_r)$ and $w(C_{r'})$ are simultaneously larger than, say, η, the uniform distribution on $\hat{C}$ is $1 - \eta$ close in TV distance to both C_r and $C_{r'}$. On the other hand, since C_r and $C_{r'}$ have Δ-separated parameters, the parameters of the uniform distribution on $\hat{C}$ must be far from that of at least one of C_r and $C_{r'}$ - say, C_r WLOG (follows from a triangle-like inequality that is easy to prove for the notion of parameter distance in Def I.5). In that case, we have a reasonable distribution (uniform distribution on $\hat{C}$) that is close to another reasonable distribution (uniform on C_r) in TV distance but their parameters are far from each other! We will prove that this is not possible because of the following

Reasonable distributions close in TV distance have close parameters.

It is important to observe that such a statement is false even for subgaussian distributions - indeed, moment upper bounds (such as those that follow from subgaussianity) are simply not enough to give *any* bound on the parameter distance of TV-close pairs at all. We note that anti-concentration

[8] We find the term *certifiability* more accurate than the usual "identifiability" in this context. Formally, certifiability refers to checking that a purported solution is "good" while identifiability relates to a sample containing information about a certain parameter we desire to estimate. Certifiability implies identifiability - it gives a test that we can check for all possible candidate solutions with the guarantee that only true solutions will pass the checks.

(and the consequent moment lower bound) is crucial to prove such a statement.

There's a lot of work in statistics that proves such statements for natural families of distributions such as Gaussians (see for e.g. [55]). In fact, all works that design outlier-robust estimation algorithms in the strong contamination model implicitly prove such a statement. This connection is made explicit in the work on robust moment estimation [8]. Our setting, however, differs from these works because we deal with the regime where the TV distance is close to 1 (in contrast to the setting where TV distance is close to 0 in the above works) outlier-robust estimation.

For the special case of Gaussians, proving such a statement even for the regime where TV distance happens to be ~ 1 turns out to be elementary However such a proof, because it uses the PDF of the distribution heavily is unlikely to be expressible in low-degree sum-of-squares proof system - a key necessity for our algorithmic application.

But perhaps even more importantly, the proof for the Gaussian case above is opaque and doesn't reveal what properties of the distribution come into play for such a statement to be true. We show that the statement above holds for all hyper-contractive and anti-concentrated distributions. As a result, we obtain both, an argument that applies to more general class of distributions and a proof translatable (with some effort) into low-degree sum-of-squares proof system.

Proof Idea: Proving TV vs Parameter Distance Bounds via Variance Mismatch. We will prove the TV vs parameter distance relationships for reasonable distributions by giving a low-degree sum-of-squares proof of the statement in the contrapositive form. In this form, the result informally says that if $\hat{C}$ (indicated by w) that defines a reasonable distribution cannot simultaneously have large intersections with two well-separated, reasonable distributions C_r and $C_{r'}$. That is, the product $w(C_r)w(C_{r'})$ must be small.

To prove such a statement, we deal with each of the three ways (see Def I.5) $C_r, C_{r'}$ can be separated one by one. In each of these cases, we will find a degree 2 polynomial in $x \sim \hat{C}$ (the purported cluster) that simultaneously has high variance if $w(C_r)$ and $w(C_{r'})$ are both large (since C_r and $C_{r'}$ are separated). On the other hand, we will also show that for certifiably hyper-contractive $\hat{C}$, the polynomial above cannot have too large a variance. Taken together, these two statement yield a bound on the product $w(C_r)w(C_{r'})$.

In the following, we discuss the ideas that go into proving such statements for each of the three kinds of parameter separation. We will also briefly discuss two basic additions to sum-of-squares toolkit that allow us to translate this proof into the low-degree SoS proof system.

It turns out that the "hardest" case to deal is that of spectral separation and we begin our exposition with it.

B. Simultaneous intersection bounds from spectral separation

For the purpose of this discussion, assume that the means $\mu(r) = \mu(r') = 0$. Since C_r and $C_{r'}$ are spectrally separated, there exists a unit vector v such that $\Delta_{\text{spec}} v^\top \Sigma(r) v \leq v^\top \Sigma(r') v$. We will use the polynomial $\langle x, v\rangle^2$ for this v as our "mismatch" marker as discussed above.

The key idea of the proof is to show that if $w(C_r)$ and $w(C_{r'})$ are simultaneously large, then, because of the stark difference in the behavior of C_r and $C_{r'}$ in direction v, the degree 2 polynomial $\langle x, v\rangle^2$ for $x \sim \hat{C}$ must have a large variance. We will prove this statement by using anti-concentration of C_r and $C_{r'}$. On the other hand, we will show that since $\hat{C}$ is also anti-concentrated, $\langle x, v\rangle^2$ for $x \sim \hat{C}$ cannot have *too* large a variance. Stringing together these bounds should, in principle, give us upper bound on $w(C_r)w(C_{r'})$.

While we manage to prove both the statements above via low-degree SoS proofs, putting them together turns out to be involved. It's easy to do this via a "real-world" argument. However, such a proof relies on case analysis that doesn't appear easy to SoSize. This is where we incur a dependence on the spread parameter κ. We explain these steps in more detail next.

Lower-Bound on the variance. We start by considering (the reason will become clear in a moment) the random variable $z - z'$ where $z, z' \sim \hat{C}$ are independent uniform draws. Then, it's easy to compute that $z - z'$ has mean 0 and covariance $2\Sigma(w)$. Thus, in order to lower bound $v^\top \Sigma(w) v$, we can consider the polynomial $\mathbb{E}_{z,z' \sim \hat{C}} \langle z - z', v\rangle^2$.

Here's the simple but important observation (and our reason for looking at $z - z'$). With probability $w(C_r)$, $z \in C_r$ and with probability $w(C_{r'})$, $z' \in C_{r'}$. Thus, $w(C_r)w(C_{r'})$ fraction of samples $z - z'$ from $\hat{C}$ are differences of independent samples from C_r and $C_{r'}$.

Let's now understand the distribution of differences of independent samples from C_r and $C_{r'}$. The covariance of this distribution is $\Sigma(r) + \Sigma(r')$. Further, since each of C_r and $C_{r'}$ are anti-concentrated, so is the convolution obtained by taking differences of independent samples from C_r and $C_{r'}$. Thus, $z - z'$ takes a value $\leq \delta \sqrt{v^\top \left(\Sigma(r) + \Sigma(r')\right) v}$ with probability at most $\sim \delta$. Thus, the contribution of $z - z'$ to $v^\top \Sigma(w) v$, when it's larger than the above bound, should be at least $\geq \left(w(C_r)w(C_{r'}) - \delta\right) \delta^2 v^\top \left(\Sigma(r) + \Sigma(r')\right) v \geq \delta^2 v^\top \Sigma(r') v$.

Upper bound on variance. The main idea is to again rely on anti-concentration - but this time of $\hat{C}$ which is enforced by our constraint system $\mathcal{A}$. Now, we know that with $w(C_r)$ probability, $\hat{C}$ outputs a point from C_r. Since these points are in C_r, their contribution to the variance of $\hat{C}$ cannot be larger than $v^\top \Sigma(r) v$. On the other hand, since $\hat{C}$ is anti-concentrated, the contribution to the variance of $\hat{C}$

from points shared with C_r must be comparable to that of $\hat{C}$ if $w(C_r)$ is large. Stringing together these observations allows us to conclude that when $w(C_r)$ is large, $v^\top \Sigma(w) v$ must be comparable to $v^\top \Sigma(r) v$.

Combining Upper and Lower Bounds: Real Life vs SoS, dependence on κ. Observe that the first claim above showed a lower bound on $v^\top \Sigma(w) v$ in terms of $v^\top \Sigma(r') v$ when $w(C_r) w(C_{r'})$ is large. The second claim shows an upper bound on $v^\top \Sigma(w) v$ (when $w(C_r)$ is large) in terms of $v^\top \Sigma(r) v$. Combining this with the spectral separation condition $\Delta_{\text{spec}} v^\top \Sigma(r) v \leq v^\top \Sigma(r') v$ should immediately yield a bound on $w(C_r) w(C_{r'})$.

This argument indeed can be done easily in "real-world" and complete the proof of TV to parameter distance bounds. However, the proof involves a case-analysis based on when $w(C_r) > \delta$ vs $w(C_r) \leq \delta$ separately. This is unfortunately not possible to capture in low-degree SoS as is.

A natural strategy to do this in SoS requires, in addition, a "rough" bound on $v^\top \Sigma(w) v$. We obtain this bound by again relying on anti-concentration of $\hat{C}$. This rough bound essentially allows us to bound $v^\top \Sigma(w) v$ by (some multiple of) the maximum of $v^\top \Sigma(r) v$ as r ranges over all the k clusters.

The case of $k = 2$ vs $k > 2$. For the case of $k = 2$, the rough bound above depends only on the clusters we are dealing with (since there are only two of them) and leads to a proof without any dependence on κ. For the case of $k > 2$, however, the rough bound depends on $v^\top \Sigma(i) v$ for clusters C_i for $i \notin \{r, r'\}$ - the set we are currently dealing with and, in principle, could be arbitrarily large. We use our assumption on the *spread* of the mixture to control $v^\top \Sigma(i) v$ for all such $i \notin \{r, r'\}$.

Using uniform approximators for thresholds over $[0, 1]$. A naive argument implementing the above reasoning loses a polynomial factor in κ in the exponent. We lessen the blow by a technical trick using uniform approximator thresholds over the unit interval. We construct such polynomial by relying on standard tools from approximation theory. These polynomials allow us to capture the conditional reasoning in the real-world proof above with a low-loss -leading to a logarithmic dependence on the SoS degree on κ.

C. *Intersection Bounds from Relative Frobenius Separation*

Obtaining intersection bounds from mean separation turns out to be relatively straight forward and uses ideas similar to the ones discussed in the spectral separation case above. So we move on to the case of Relative Frobenius separation here. For the sake of exposition here, we assume $\mu(r), \mu(r') = 0$ as before and set $\Sigma(r') = I$. Then, relative Frobenius separation guarantees us that $\|\Sigma(r) - I\|_F^2 \geq \Delta_{cov}^2$.

Let's understand what happens to $\mathbb{E}_{\hat{C}} Q(x)$ - the expectation of this polynomial over the purported cluster $\hat{C}$ if it has a large intersection with both C_r and $C_{r'}$.

Lower Bound on the Variance of Q. Consider the polynomial $Q(x) = x^\top Q x$ for $Q = \Sigma(r) - I$. By direct computation, the expectation of this polynomial on C_r equals $\|\Sigma(r) - I\|_F^2 + \text{Tr}(\Sigma(r) - I)$. While the expectation on $C_{r'}$ equals $\text{Tr}(\Sigma(r) - I)$.

Using *hypercontractivity* of degree 2 polynomials over C_r and $C_{r'}$, we show that the variance of the polynomial $Q(x)$ on C_r and $C_{r'}$ is $\ll \Delta_{cov}^2$. Thus, on $\hat{C}$, for a $w(C_r)$ fraction of points $Q(x)$ would be $\approx \|\Sigma(r) - I\|_F^2 + \text{Tr}(\Sigma(r) - I)$ while for a $w(C_{r'})$ fraction of points, $Q(x)$ would be $\approx \text{Tr}(\Sigma(r) - I$. The difference in these values is $|\mathbb{E}_{x \sim C_r} Q(x) - \mathbb{E}_{x \sim C_{r'}} Q(x)| = \|\Sigma(r) - I\|_F^2 \geq \Delta_{cov}^2$. Thus, if $w(C_r) w(C_{r'})$ is large, $Q(x)$ must have a variance comparable to $w(C_r) w(C_{r'}) \Delta_{cov}^2$ on $\hat{C}$. Thus, we expect that if $\hat{C}$ picks a significant mass from both C_r and $C_{r'}$, then, $Q(x)$ must have a large variance on $\hat{C}$.

Upper Bound on the Variance of Q via SoSizing Contraction. In contrast to the the case of mean separation where we relied on anti-concentration of $\hat{C}$, we prove an upper bound on the variance of Q by relying on hypercontractivity of degree 2 polynomials of $\hat{C}$. A key step in this proof relies on *SoSizing* a basic matrix inequality: For all $d \times d$ matrices A, B, $\|AB\|_F^2 \leq \|A\|_{op}^2 \|B\|_F^2$, which follows from sub-multiplicativity of the Frobenius norm.

D. *Outlier-Robust Variant*

Making the algorithm in the discussion above outlier-robust is relatively straightforward. Observe that in this case, we do not get access to the original sample X as above. Instead, we get an ϵ-corruption of X, say Y as input. Our goal is to give a clustering of Y that corresponds to the clustering X with at most $O(k\epsilon)$ points misclassified in any given cluster. Observe that this is the information-theoretically the best possible result we can expect since all the ϵn outliers could end up corrupting a single chosen true cluster.

Our key idea here is to introduce a new collection of variables X' that "guess" the original sample that generated Y. We add the constraint that X and Y intersect in $(1 - \epsilon)$-fraction of the points to capture the only property of X that we know.

We then use a version of the system of constraints $\mathcal{A}$ with X replaced by X'. Let $C_1', C_2', \ldots, C_k'$ be the clusters induced by taking the points with the same indices as in C_i from X'. Note that in this case, X' and C_i's are indeterminates in our constraint system. Our proof from the previous section generalizes with only a few changes to yield simultaneous intersection bounds on $w'(C_r') w'(C_{r'}')$. The intersection bounds with Y then follow by noting a (degree 2 SoS proof of) $|C_i' \cap C_i| \geq (1 - 2k\epsilon)|C_i|$.

E. Recursive Clustering Algorithm

Simple rounding with larger running time. Given our certifiability proofs that prove upper bounds on simultaneous intersection of $\hat{C}$ with true clusters, one can immediately obtain an algorithm for clustering mixtures of reasonable distributions that runs in time $n^{O(s(poly(\eta/k))\log(\kappa))}$. These algorithms work by computing a pseudo-distribution $\tilde{\zeta}$ on w (the indicator of samples in $\hat{C}$) and rounding it. For the purpose of this overview, it is helpful to think of pseudo-distributions as giving us access to low-degree moments of a distribution on w that satisfies the checks that we made (certifiable anti-concentration and hypercontractivity) in our certifiability proofs above. A pseudo-distribution of degree t in n variables can be computed in time $n^{O(t)}$ via semidefinite programming and satisfies all inequalities that can be derived from our checks (constraint system) via low-degree SoS proofs.

Our rounding algorithm is simple and is the same as the one described in Section 4.3 of the monograph [56] that gives a simpler proof of the recently obtained algorithm for clustering spherical mixtures [20], [21]. We use the simultaneous intersection bounds to derive that the second moment matrix $\tilde{\mathbb{E}}_{\tilde{\zeta}}[ww^\top]$ of w (indicating $\hat{C}$) is approximately block diagonal, with approximate clusters as blocks. This allows us to iteratively *peel off* approximate clusters greedily. To establish this block diagonal structure our proof requires the pseudo-distribution to have a degree that scales with $\log \kappa$ where κ is the spread of the mixture.

Spread-independent recursive rounding. We then give a more sophisticated rounding with a running time that does not depend on the spread κ. The conceptual idea behind this rounding is based on two curious facts that we establish:

1) *Simple rounding has non-trivial information at constant degrees*. This first fact shows that when we run the simple rounding with a pseudo-distribution $\tilde{\zeta}$ of degree that *does not* grow with $\log \kappa$, we can still prove that $\tilde{\mathbb{E}}_{\tilde{\zeta}}[ww^\top]$ has a *partial block diagonal* structure. This structure allows us to prove that for the clustering $\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_k$ output by our simple rounding above, there exists a (non-trivial) partition $S \cup L = [k]$ such that both $\cup_{i \in S} \hat{C}_i$ and $\cup_{j \in T} \hat{C}_j$ are essentially unions of the true clusters.
   The proof relies on two facts: 1) if no pair of components of the input mixture are spectrally separated, then, the spread κ is small so our simple rounding already works. 2) Even when there's a pair of components that are spectrally separated, the SoS degree required in our simultaneous intersection bounds can be much smaller than κ. Concretely, our analysis yields a degree that scales with $\frac{v^\top \Sigma(i) v}{v^\top \Sigma(r') v}$ that we loosely upper bound by $\kappa = \max_{i,j} \frac{v^\top \Sigma(i) v}{v^\top \Sigma(j) v}$. If $v^\top \Sigma(r') v$ is comparable to $\max_{i \leq k} v^\top \Sigma(i) v$, then, the SoS degree of the proof is much smaller than κ. We use this observation to show that there's a $S \subseteq [k]$ and a $O(1)$ degree SoS proof that bounds the simultaneous intersection of $\hat{C}$ with true clusters C_i and C_j whenever $i \in S$ and $j \notin S$. This is enough to obtain a *partial cluster recovery* guarantee. Thus, $\cup_{i \in S} \hat{C}_i$ can be treated as a mixture of ($< k$) components along with a small fraction of outliers and we can recurse. Of course, we do not know S, so our algorithm tries all the 2^k possible choices and recursively tries to cluster them.
2) *Verifying clusters requires only constant-degree pseudo-distributions*. In order to run the recursive clustering algorithm suggested above, we need a subroutine that can efficiently verify that a given purported cluster is close to a true one. While we cannot show that degree $O(s(poly(\eta/k)))$ pseudo-distributions are enough to *find* a clustering, we will prove that they are enough to *verify* a purported clustering. Concretely, given a purported cluster $\hat{C}$, we show that there's a pseudo-distribution of constant degree (independent of κ) consistent with verification constraints iff $\hat{C}$ is close to a true cluster.
   The "completeness" of the verification algorithm is easy to prove. The meat of the analysis is proving soundness - i.e. if a purported cluster $\hat{C}$ has an appreciable intersection with two different true clusters, then the verification algorithm must output reject.
   A priori, such a result can appear a bit confusing - after all, we just spent most of this overview arguing SoS proofs of degree that grow with κ for verifying purported clusters. The key technical difference (quite curious from a proof complexity perspective) is that in the setting of verification, we are trying to derive a contradiction from the assumption that the intersection bounds are simultaneously large for two distinct true clusters. While in the simultaneous intersection bounds, the goal is similar statement but stated in terms of the *contrapositive*.

F. Covariance Estimation in Relative Frobenius Error

The tools we develop allow us to get an additional application - an outlier-robust algorithm to compute the covariance of a distribution with optimal *relative Frobenius error*. Prior works of Lai Rao and Vempala [5] and Kothari and Steurer [8] gave guarantees for covariance estimation in spectral distance (which implies only dimension dependent bounds on the relative Frobenius error) or worked only for Gaussian distributions [6]. We show an optimal $\tilde{O}(\epsilon)$ (independent of the dimension) error guarantee on relative Frobenius error in the presence of an ϵ-fraction adversarial outliers whenever the target distribution is certifiably hypercontractive. Our algorithm is same as the one used in [8] but our analysis relies on certifiable hypercontractivity along with the SoS contraction lemma discussed above.

As a corollary of this result, we can take an accurate clustering output by our clustering algorithms for reasonable distributions and use our covariance estimation algorithm here to get statistically optimal estimates of mean and covariance in the distances presented in Definition I.5 thus obtaining outlier-robust parameter estimation algorithms from our outlier-robust clustering algorithm.

IV. Algorithm

Our constraint system $\mathcal{A}$ uses polynomial inequalities to describe a subset $\hat{C}$ of size αn of the input sample X. We impose constraints on $\hat{C}$ so that the uniform distribution on $\hat{C}$ satisfies certifiable anti-concentration and hypercontractivity of degree-2 polynomials. We intend the true clusters $C_1, C_2, \ldots, C_r$ to be the only solutions for $\hat{C}$. Proving that this statement holds and that it has a low-degree SoS proof is the bulk of our technical work in this section.

We describe the specific formulation next. Throughout this section, we use the notation $Q(x)$ to denote $x^\top Qx$ for $d \times d$ matrix valued indeterminate Q. For ease of exposition, we break our constraint system $\mathcal{A}$ into natural categories $\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_5$. Our constraint system relies on parameter τ, δ and an appropriate setting can be found in [3].

For our argument, we will need access to the square root of the indeterminate Σ. So we introduce the constraint system $\mathcal{A}_1$ with an extra matrix valued indeterminate Π (with auxiliary matrix-valued indeterminate U) that satisfies the polynomial equality constraints corresponding to Π being the square root of Σ. Note that the first constraint is equivalent to $\Pi \succeq 0$ in "ordinary math". Square-Root Constraints:

$$\mathcal{A}_1 = \left\{ \begin{array}{c} \Pi = UU^\top \\ \Pi^2 = \Sigma \end{array} \right\} \tag{2}$$

Next, we formulate intersection constraints that identify the subset $\hat{C}$ of size αn. Subset Constraints:

$$\mathcal{A}_2 = \left\{ \begin{array}{lr} \forall i \in [n] & w_i^2 = w_i \\ & \sum_{i \in [n]} w_i = \dfrac{n}{k} \end{array} \right\} \tag{3}$$

Next, we enforce that $\hat{C}$ must have mean μ and covariance Σ, where both μ and Σ are indeterminates. Parameter Constraints:

$$\mathcal{A}_3 = \left\{ \begin{array}{r} \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} w_i x_i = \mu \\ \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} w_i (x_i - \mu)(x_i - \mu)^\top = \Sigma \end{array} \right\} \tag{4}$$

Finally, we enforce certifiable anti-concentration at two slightly different parameter regimes (characterized by $\tau \leq \delta$) along with the hypercontractivity of $\hat{C}$. Certifiable Anti-Concentration :

$$\mathcal{A}_4 = \left\{ \begin{array}{r} \dfrac{k^2}{n^2} \displaystyle\sum_{i,j=1}^{n} \quad w_i w_j q_{\delta,2\Sigma}^2 \left((x_i - x_j), v \right) \\ \leq 2^{s(\delta)} C \delta \left(v^\top \Sigma v \right)^{s(\delta)} \\ \dfrac{k^2}{n^2} \displaystyle\sum_{i,j=1}^{n} \quad w_i w_j q_{\tau,2\Sigma}^2 \left((x_i - x_j), v \right) \\ \leq 2^{s(\tau)} C \tau \left(v^\top \Sigma v \right)^{s(\tau)} \end{array} \right\} \tag{5}$$

Certifiable Hypercontractivity :

$$\mathcal{A}_5 = \left\{ \begin{array}{r} \forall j \leq 2s, \quad \dfrac{k^2}{n^2} \displaystyle\sum_{i,j \leq n} w_i w_j Q(x_i - x_j)^{2j} \\ \leq (Cj)^{2j} \left\| \Pi Q \Pi \right\|_F^{2j} \end{array} \right\} \tag{6}$$

We are now ready to describe our algorithm. Our algorithm follows the same outline as the simplified proof for clustering spherical mixtures presented in [56] (Chapter 4.3). The idea is to find a pseudo-distribution $\tilde{\zeta}$ that minimizes the objective $\left\| \tilde{\mathbb{E}}_{\tilde{\zeta}}[w] \right\|_2$ and is consistent with the constraint system $\mathcal{A}$.

It is simple to round the resulting solution to true clusters: our analysis yields that the matrix $\tilde{\mathbb{E}}_{\tilde{\zeta}}[ww^\top]$ is approximately block diagonal with the blocks approximately corresponding to the true clusters $C_1, C_2, \ldots, C_k$. We can then recover a cluster by a repeatedly greedily selecting n/k largest entries in a random row, removing those columns off and repeating. We describe this algorithm below.

```
**Algorithm IV.1** (Clustering General Mixtures).
Given: A sample $X$ of size $n$ with true clusters
    $C_1, C_2, \ldots, C_k$ of size $n/k$ each.
Output: A partition of $X$ into an approximately correct clusters $\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_k$.
Operation:
    1) Find a pseudo-distribution $\tilde{\zeta}$ satisfying $\mathcal{A}$ minimizing $\left\| \tilde{\mathbb{E}}[w] \right\|_2^2$.
    2) For $M = \tilde{\mathbb{E}}_{w \sim \tilde{\zeta}}[ww^\top]$, repeat for $1 \leq \ell \leq k$:
        a) Choose a uniformly random row $i$ of $M$.
        b) Let $\hat{C}_\ell$ be the set of points indexed by the largest $\frac{n}{k}$ entries in the $i$th row of $M$.
        c) Remove the rows and columns with indices in $\hat{C}_\ell$.
```

We defer the description of the recursive algorithm to obtain the spread independent bound and the analysis to the full versions of this paper [4], [36].

Acknowledgment

AB and PK thank Boaz Barak, Ryan O'Donnell, Venkat Guruswami, Rajesh Jayaram, Gautam Kamath, Roie Levin, Jerry Li, Pedro Paredes, Nicolas Resch and David Woodruff for illuminating discussions related to this project, and Misha Ivkov for pointing out several typos. ID and SK thank the Simons Institute, especially the program on Foundations of Data Science, for hosting some of the authors while much of this work was completed

References

[1] S. Karmalkar, A. Klivans, and P. Kothari, "List-decodable linear regression," in *Advances in Neural Information Processing Systems*, 2019, pp. 7425–7434.

[2] P. Raghavendra and M. Yau, "List decodable learning via sum of squares," in *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2020, pp. 161–180.

[3] A. Bakshi and P. Kothari, "Outlier-robust clustering of non-spherical mixtures," *arXiv preprint arXiv:2005.02970*, 2020.

[4] I. Diakonikolas, S. B. Hopkins, D. Kane, and S. Karmalkar, "Robustly learning any clusterable mixture of gaussians," *arXiv preprint arXiv:2005.06417*, 2020.

[5] K. A. Lai, A. B. Rao, and S. Vempala, "Agnostic estimation of mean and covariance," in *FOCS*. IEEE Computer Society, 2016, pp. 665–674.

[6] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, "Robust estimators in high dimensions without the computational intractability," in *FOCS*. IEEE Computer Society, 2016, pp. 655–664.

[7] M. Charikar, J. Steinhardt, and G. Valiant, "Learning from untrusted data," in *STOC*. ACM, 2017, pp. 47–60.

[8] P. K. Kothari and D. Steurer, "Outlier-robust moment-estimation via sum-of-squares," *CoRR*, vol. abs/1711.11581, 2017. [Online]. Available: http://arxiv.org/abs/1711.11581

[9] J. Steinhardt, M. Charikar, and G. Valiant, "Resilience: A criterion for learning in the presence of arbitrary outliers," *CoRR*, vol. abs/1703.04940, 2017.

[10] Y. Cheng, I. Diakonikolas, and R. Ge, "High-dimensional robust mean estimation in nearly-linear time," in *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, T. M. Chan, Ed. SIAM, 2019, pp. 2755–2771. [Online]. Available: https://doi.org/10.1137/1.9781611975482.171

[11] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, "Being robust (in high dimensions) can be practical," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 2017, pp. 999–1008.

[12] ——, "Robustly learning a gaussian: Getting optimal error, efficiently," in *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, A. Czumaj, Ed. SIAM, 2018, pp. 2683–2702. [Online]. Available: https://doi.org/10.1137/1.9781611975031.171

[13] Y. Cheng, I. Diakonikolas, R. Ge, and D. P. Woodruff, "Faster algorithms for high-dimensional robust covariance estimation," in *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, ser. Proceedings of Machine Learning Research, A. Beygelzimer and D. Hsu, Eds., vol. 99. PMLR, 2019, pp. 727–757. [Online]. Available: http://proceedings.mlr.press/v99/cheng19a.html

[14] I. Diakonikolas, D. M. Kane, and A. Stewart, "Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures," in *FOCS*. IEEE Computer Society, 2017, pp. 73–84.

[15] A. Klivans, P. K. Kothari, and R. Meka, "Efficient algorithms for outlier-robust regression," *arXiv preprint arXiv:1803.03241*, 2018.

[16] I. Diakonikolas, G. Kamath, D. Kane, J. Li, J. Steinhardt, and A. Stewart, "Sever: A robust meta-algorithm for stochastic optimization," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 1596–1606. [Online]. Available: http://proceedings.mlr.press/v97/diakonikolas19a.html

[17] A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar, "Robust estimation via robust gradient estimation," *CoRR*, vol. abs/1802.06485, 2018. [Online]. Available: http://arxiv.org/abs/1802.06485

[18] S. Karmalkar, A. R. Klivans, and P. K. Kothari, "List-decodable linear regression," *CoRR*, vol. abs/1905.05679, 2019. [Online]. Available: http://arxiv.org/abs/1905.05679

[19] P. Raghavendra and M. Yau, "List decodable learning via sum of squares," *CoRR*, vol. abs/1905.04660, 2019. [Online]. Available: http://arxiv.org/abs/1905.04660

[20] P. K. Kothari and J. Steinhardt, "Better agnostic clustering via relaxed tensor norms," 2017.

[21] S. B. Hopkins and J. Li, "Mixture models, robustness, and sum of squares proofs," 2017.

[22] I. Diakonikolas and D. M. Kane, "Recent advances in algorithmic high-dimensional robust statistics," *arXiv preprint arXiv:1911.05911*, 2019.

[23] E. Ben-Sasson, "Size space tradeoffs for resolution," in *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, 2002, pp. 457–464.

[24] K. Pearson, "Contributions to the mathematical theory of evolution," *Philosophical Transactions of the Royal Society of London. A*, vol. 185, pp. 71–110, 1894.

[25] S. Dasgupta, "Learning mixtures of gaussians," in *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*. IEEE, 1999, pp. 634–644.

[26] S. Arora and R. Kannan, "Learning mixtures of arbitrary gaussians," in *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, 2001, pp. 247–257.

[27] S. Vempala and G. Wang, "A spectral algorithm for learning mixture models," *Journal of Computer and System Sciences*, vol. 68, no. 4, pp. 841–860, 2004.

[28] S. C. Brubaker and S. S. Vempala, "Isotropic pca and affine-invariant clustering," in *Building Bridges*. Springer, 2008, pp. 241–281.

[29] S. C. Brubaker, "Robust pca and clustering in noisy mixtures," in *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 2009, pp. 1078–1087.

[30] A. T. Kalai, A. Moitra, and G. Valiant, "Efficiently learning mixtures of two gaussians," in *STOC*. ACM, 2010, pp. 553–562.

[31] A. Moitra and G. Valiant, "Settling the polynomial learnability of mixtures of gaussians," in *FOCS*. IEEE Computer Society, 2010, pp. 93–102.

[32] M. Belkin and K. Sinha, "Polynomial learning of distribution families," *SIAM J. Comput.*, vol. 44, no. 4, pp. 889–911, 2015.

[33] A. Moitra and G. Valiant, "Settling the polynomial learnability of mixtures of Gaussians," in *FOCS*, 2010, pp. 93–102.

[34] R. O'Donnell, *Analysis of Boolean functions*. Cambridge University Press, New York, 2014. [Online]. Available: http://dx.doi.org/10.1017/CBO9781139814782

[35] M. Kauers, R. O'Donnell, L. Tan, and Y. Zhou, "Hypercontractive inequalities via sos, and the frankl-rödl graph," in *SODA*. SIAM, 2014, pp. 1644–1658.

[36] A. Bakshi and P. Kothari, "List-decodable subspace recovery via sum-of-squares," *arXiv preprint arXiv:2002.05139*, 2020.

[37] R. D. De Veaux, "Mixtures of linear regressions," *Comput. Statist. Data Anal.*, vol. 8, no. 3, pp. 227–245, 1989. [Online]. Available: https://doi.org/10.1016/0167-9473(89)90043-1

[38] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the em algorithm," *Neural Computation*, vol. 6, no. 2, pp. 181–214, 1994.

[39] S. Faria and G. Soromenho, "Fitting mixtures of linear regressions," *J. Stat. Comput. Simul.*, vol. 80, no. 1-2, pp. 201–225, 2010. [Online]. Available: https://doi.org/10.1080/00949650802590261

[40] X. Yi, C. Caramanis, and S. Sanghavi, "Alternating Minimization for Mixed Linear Regression," *arXiv e-prints*, p. arXiv:1310.3745, Oct 2013.

[41] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the EM algorithm: From population to sample-based analysis," *CoRR*, vol. abs/1408.2156, 2014.

[42] Y. Chen, X. Yi, and C. Caramanis, "A convex formulation for mixed regression with two components: Minimax optimal rates," in *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, 2014, pp. 560–604. [Online]. Available: http://jmlr.org/proceedings/papers/v35/chen14.html

[43] K. Zhong, P. Jain, and I. S. Dhillon, "Mixed linear regression with multiple components," in *NIPS*, 2016, pp. 2190–2198.

[44] H. Sedghi, M. Janzamin, and A. Anandkumar, "Provable tensor methods for learning mixtures of generalized linear models," in *AISTATS*, ser. JMLR Workshop and Conference Proceedings, vol. 51. JMLR.org, 2016, pp. 1223–1231.

[45] Y. Li and Y. Liang, "Learning mixtures of linear regressions with nearly optimal complexity," in *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, 2018, pp. 1125–1144. [Online]. Available: http://proceedings.mlr.press/v75/li18b.html

[46] R. Vidal, "Subspace clustering," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 52–68, Mar. 2011.

[47] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: a review," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 90–105, Jun. 2004.

[48] S. Chen, J. Li, and Z. Song, "Learning mixtures of linear regressions in subexponential time via fourier moments," *CoRR*, vol. abs/1912.07629, 2019. [Online]. Available: http://arxiv.org/abs/1912.07629

[49] A. T. Kalai, A. Moitra, and G. Valiant, "Efficiently learning mixtures of two Gaussians," in *STOC*, 2010, pp. 553–562.

[50] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, "Robust estimators in high dimensions without the computational intractability," in *Proc. 57th IEEE Symposium on Foundations of Computer Science (FOCS)*, 2016, pp. 655–664.

[51] S. B. Hopkins and J. Li, "Mixture models, robustness, and sum of squares proofs," in *Proc. 50th Annual ACM Symposium on Theory of Computing (STOC)*, 2018, pp. 1021–1034.

[52] P. K. Kothari, J. Steinhardt, and D. Steurer, "Robust moment estimation and improved clustering via sum of squares," in *Proc. 50th Annual ACM Symposium on Theory of Computing (STOC)*, 2018, pp. 1035–1046.

[53] I. Diakonikolas, D. M. Kane, and A. Stewart, "List-decodable robust mean estimation and learning mixtures of spherical Gaussians," in *Proc. 50th Annual ACM Symposium on Theory of Computing (STOC)*, 2018, pp. 1047–1060.

[54] ——, "Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures," in *Proc. 58th IEEE Symposium on Foundations of Computer Science (FOCS)*, 2017, pp. 73–84.

[55] L. Devroye, A. Mehrabian, and T. Reddad, "The total variation distance between high-dimensional gaussians," 2018.

[56] N. Fleming, P. Kothari, and T. Pitassi, "Semialgebraic proofs and efficient algorithm design," *Foundations and Trends® in Theoretical Computer Science*, vol. 14, no. 1-2, pp. 1–221, 2019. [Online]. Available: http://dx.doi.org/10.1561/0400000086