Clustering Mixtures of Bounded Covariance Distributions Under Optimal Separation*

Ilias Diakonikolas†
University of Wisconsin-Madison
ilias@cs.wisc.edu

Daniel M. Kane‡
University of California, Davis
dakane@cs.ucsd.edu

Jasper C.H. Lee§
University of Wisconsin-Madison
jasperlee@ucdavis.edu

Thanasis Pittas¶
University of Wisconsin-Madison
pittas@wisc.edu

Abstract

We study the clustering problem for mixtures of bounded covariance distributions, under a fine-grained separation assumption. Specifically, given samples from a k-component mixture distribution $D = \sum_{i=1}^{k} w_i P_i$, where each $w_i \geq \alpha$ for some known parameter α, and each P_i has unknown covariance $\Sigma_i \preceq \sigma_i^2 \cdot I_d$ for some unknown σ_i, the goal is to cluster the samples assuming a pairwise mean separation in the order of $(\sigma_i + \sigma_j)/\sqrt{\alpha}$ between every pair of components P_i and P_j. Our main contributions are as follows:

- For the special case of nearly uniform mixtures, we give the first polynomial-time algorithm for this clustering task. Prior work either required separation scaling with the maximum cluster standard deviation (i.e. $\max_i \sigma_i$) [DKK+22b] or required both additional structural assumptions and mean separation scaling as a large degree polynomial in $1/\alpha$ [BKK22].
- For arbitrary (i.e. general-weight) mixtures, we point out that accurate clustering is information-theoretically impossible under our fine-grained mean separation assumptions. We introduce the notion of a *clustering refinement* — a list of not-too-small subsets satisfying a similar separation, and which can be merged into a clustering approximating the ground truth — and show that it is possible to efficiently compute an accurate clustering refinement of the samples. Furthermore, under a variant of the "no large sub-cluster" condition introduced in prior work [BKK22], we show that our algorithm will output an accurate clustering, not just a refinement, even for general-weight mixtures. As a corollary, we obtain efficient clustering algorithms for mixtures of well-conditioned high-dimensional log-concave distributions.

Moreover, our algorithm is robust to a fraction of adversarial outliers comparable to α.

At the technical level, our algorithm proceeds by first using list-decodable mean estimation to generate a polynomial-size list of possible cluster means, before successively pruning candidates using a carefully constructed convex program. In particular, the convex program takes as input a candidate mean $\hat{\mu}$ and a scale parameter $\hat{s}$, and determines the existence of a subset of points that could plausibly form a cluster with scale $\hat{s}$ centered around $\hat{\mu}$. While the natural way of designing this program makes it non-convex, we construct a convex relaxation which remains satisfiable by (and only by) not-too-small subsets of true clusters..

1 Introduction

Clustering mixture models is one of the most basic and widely-used statistical primitives on data samples from high-dimensional distributions, with applications in a variety of fields, including bioinformatics, astrophysics, and marketing [Lin95, GEGMMI10]; see [TSM85] for an extensive list of applications. Informally, the input is a set of n samples drawn from a mixture distribution $D = \sum_{i=1}^{k} w_i P_i$ over $\mathbb{R}^d$, where w_i is the mixing weight of component

*The arXiv version of the paper can be accessed at https://arxiv.org/abs/2312.11769

†Supported by NSF Medium Award CCF-2107079 and NSF Award CCF-1652862 (CAREER).
‡Supported by NSF Medium Award CCF-2107547 and NSF Award CCF-1553288 (CAREER).
§Jasper C.H. Lee's work was partially done while he was at the University of Wisconsin-Madison, supported by a Croucher Fellowship for Postdoctoral Research and NSF Medium Award CCF-2107079.
¶Supported by NSF Medium Award CCF-2107079 and NSF Award DMS-2023239 (TRIPODS).

P_i. The goal is to cluster (most of) the samples such that the clustering is approximately equal to partitioning the data according to the ground truth; namely, partitioning samples according to which mixture component they were drawn from. For the clustering task to be information-theoretically possible, it is common to make concentration assumptions on each mixture component P_i (e.g. sub-Gaussianity, or a bounded moments assumption), as well as on the pairwise separation between the means of the components.

The prototypical case is that of Gaussian mixtures and has been extensively studied in the literature; see, e.g. [VW02, KSV05, AM05] and references therein. In more detail, [VW02] studied the clustering of data drawn from mixtures of separated spherical Gaussians. Subsequent work [KSV05, AM05] built on the approach of [VW02] to design clustering algorithms for mixtures of separated Gaussians with general covariances. The main algorithmic technique underlying these papers is to apply k-PCA in order to discover the subspace spanned by the means of the mixture components.

The focus of this paper is the more general *heavy-tailed* setting, where each component is only assumed to have *bounded covariance* instead of stronger concentration. Specifically, suppose that each component P_i has unknown covariance matrix Σ_i that satisfies $\Sigma_i \preceq \sigma^2 \cdot I_d$, for some unknown parameter $\sigma > 0$. For notational simplicity, we restrict this discussion to *uniform* mixtures (corresponding to the case that $w_i = 1/k$ for all $i \in [k]$). Then, unless the component means have pairwise ℓ_2-distance $\gg \sigma\sqrt{k}$, accurate clustering is information-theoretically impossible in the worst-case. On the positive side, the recent work of [DKK+22b] gave a computationally efficient algorithm which achieved the best worst-case separation: if all the components P_i have covariances $\Sigma_i \preceq \sigma^2 \cdot I_d$, then [DKK+22b] showed that it is possible to accurately cluster when given a pairwise separation of $C\sigma\sqrt{k}$, where $C > 0$ is a sufficiently large universal constant[1].

The preceding discussion suggests that the algorithmic problem of clustering mixtures of bounded covariance distributions under the information-theoretically optimal mean estimation (within constant factors) is fully resolved. Yet, consider the simple example shown in Figure 1 below.

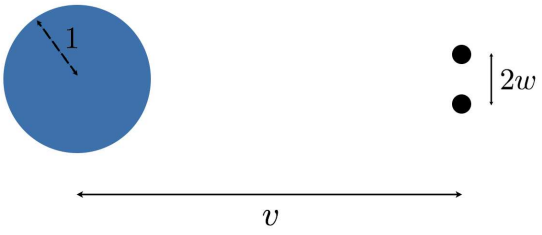

Figure 1: Example "well-separated" mixture distribution that cannot be handled by the algorithm of [DKK+22b].

In this example, we have an identity-covariance distribution on the left, separated by distance $v \gg 1$ from a pair of 0-covariance distributions on the right, which are in turn separated by some small distance $2w \ll 1$. This example is clearly clusterable and "well-separated", since there is essentially no overlap between any of the mixture components. However, the example cannot be handled by the algorithm of [DKK+22b] or earlier algorithms[2] for the following reason: the largest variance is $\sigma^2 = 1$, but the two 0-covariance distributions are separated only by $2w \ll 1$ — instead of the required $\Theta(\sigma\sqrt{k}) = \Theta(1)$ separation.

The above example illustrates an important conceptual weakness of prior work in the heavy-tailed (bounded covariance) setting: it requires that the pairwise mean separation is measured by the *maximum* covariance across all the mixture components — even if the pair of components in question both have small covariances. This distinction can make a large quantitative difference in both theory and practice. Indeed, even for the special case that the components are of approximately the same size (cardinality), their relative radii may dramatically differ.

Motivation: Achieving fine-grained separation A more reasonable separation assumption that we focus on in this paper is as follows. Suppose that the components P_i and P_j have maximum standard deviations σ_i and σ_j respectively. Then we require the corresponding means μ_i and μ_j to be separated in ℓ_2-distance by a quantity scaling with $\sigma_i + \sigma_j$. Note that this is much weaker than the prior assumption scaling with $\max_i \sigma_i$. We also

[1]We note that [DKK+22b] gave an almost-linear time algorithm that succeeds under slightly stronger separation (within a $\log(k)$-factor of the optimal). If one allows polynomial-time algorithms, this extra factor does not appear.

[2]We note, for example, that the algorithm [AS12] produces an accurate clustering under separation $\Delta \gg k\sigma$.

point out that clustering under such fine-grained separation has been achieved for Gaussian components in earlier works [AM05, KSV05, Bru09]. However, to the best of our knowledge, no such result was previously known for the bounded covariance setting. Motivated by this gap in the literature, in this paper we ask:

> *Is it possible to efficiently cluster data from mixtures of bounded covariance distributions under the fine-grained separation assumption? Specifically, can we efficiently achieve accurate clustering under pairwise mean separation in the order of* $\sqrt{k}(\sigma_i + \sigma_j)$*?*

As our main contribution, in this paper we study and *essentially resolve* this question.

We emphasize that the heavy-tailed setting introduces a number of technical challenges that do not appear in the presence of strong concentration. For the sake of intuition, we explain below how k-PCA — a standard spectral technique used in prior work — provably fails in our setting.

Failure of k-PCA One of the main standard techniques for clustering mixtures of separated components is to perform k-PCA: find the top-k dimensional subspace of the sample covariance, and show that with high probability, this subspace captures the span of the mixture component means. However, this technique fails for bounded covariance distributions under our fine-grained separation assumption, even with infinitely many samples. This can be demonstrated through a variant of the example in Figure 1. Consider the uniform (i.e. equal weights) mixture with a component with unit covariance on a subspace V at the origin, and two components with 0-covariance, located at points $v + w$ and $v - w$ with $\|v\|_2 \gg 1$ and $\|w\|_2 \ll 1$. Suppose also that V is $\Omega(d)$-dimensional, and V, v, w are orthogonal to each other. Denoting the identity matrix in the subspace V by I_V, the covariance of the full distribution is equal to $\frac{1}{3}I_V + \frac{2}{9}vv^\top + \frac{2}{3}ww^\top$. Given that $\|v\|_2 \gg 1$ and $\|w\|_2 \ll 1$, the eigenvectors of this covariance are v, any $\Omega(d)$-dimensional basis of V, finally followed by w. Thus, in order to have the direction w in the subspace found by k-PCA, we might need as many as $k = \Omega(d)$ dimensions, which reduces the dimensionality only mildly.

Summary of contributions Our first goal focuses on uniform-weight mixture distributions, with the aim of clustering assuming only a pairwise separation of $C \cdot (\sigma_i + \sigma_j)\sqrt{k}$ between mixture components P_i and P_j satisfying $\Sigma_i \preceq \sigma_i^2 \cdot I_d$ and $\Sigma_j \preceq \sigma_j^2 \cdot I_d$, for some sufficiently large universal constant C. We note that the individual standard deviations σ_i are *unknown* to the algorithm.

For this setting, we give the first efficient algorithm (Algorithm 1) achieving this guarantee in Theorem 1.1. We point out that the recent work of [BKK22] also studies the heavy-tailed setting under a fine-grained separation assumption. However, they require separation which scales like $(\sigma_i+\sigma_j)\,\mathrm{poly}(k, \log n)$, for a large degree polynomial[3]. More importantly, they also require an additional "no large sub-cluster assumption" on the samples beyond bounded covariance — even for the uniform-weight mixture setting.

Our second, more general goal is to study the limits of clustering general-weight mixtures of bounded covariance distributions, under the same fine-grained pairwise separation assumption. Perhaps surprisingly, we point out that it is information-theoretically impossible to achieve accurate clustering due to identifiability issues — there can be multiple valid ground truths for the same mixture and there is no way to tell which one is the "correct" one — if the mixing weights are (highly) non-uniform. Nonetheless, our main algorithm (Algorithm 1) efficiently produces an accurate *refinement* of the ground truth clustering (Theorem 1.2): informally, a clustering refinement is a list of not-too-small and disjoint subsets of samples such that there *exists* a way to combine them into a clustering close to the ground truth, and furthermore, these subset are themselves well-separated like the ground truth distribution. This essentially amounts to the information-theoretically strongest possible guarantee in our setting. We further show that, under a "no large sub-cluster" condition (à la [BKK22]), the same algorithm outputs exactly the correct k clusters (up to some small fraction of misclassified points).

Finally, we remark that our algorithm is robust to a fraction of adversarial outliers that is comparable to the size of the smallest cluster.

1.1 Our results

Even in the special case of uniform-weight mixtures, no prior work can find an accurate clustering under a fine-grained separation assumption scaling with $\sigma_i + \sigma_j$ between components P_i and P_j, even if we allow a sub-optimal $\mathrm{poly}(k)$ scaling. Here we present our first result, solving both issues simultaneously.

[3]Their results do not explicit state the degree, but we believe it is at least degree-4 for k according to their algorithm, as opposed to our optimal $\sqrt{k}$ dependence.

Algorithm 1 finds an accurate clustering in polynomial time, assuming the optimal (up-to-constants) separation in the order of $(\sigma_i + \sigma_j)\sqrt{k}$, which is both fine-grained and has the information-theoretically optimal $\sqrt{k}$ dependence.

THEOREM 1.1. (CLUSTERING UNIFORM-WEIGHT BOUNDED COVARIANCE MIXTURES) *Let C be a sufficiently large constant. Consider a uniform-weight mixture distribution $D = \sum_{i=1}^{k} \frac{1}{k} P_i$ with k components on $\mathbb{R}^d$. Suppose that α is a parameter in $[0.6/k, 1/k]$. Let μ_i and Σ_i be the (unknown) mean and covariance of each P_i, and assume that $\Sigma_i \preceq \sigma_i^2 \cdot I_d$ (with σ_i being unknown) and $\|\mu_i - \mu_j\|_2 > C(\sigma_i + \sigma_j)/\sqrt{\alpha}$ for all $i \neq j$.*

Draw n samples from D, and let S_i be the samples from the i^{th} mixture component. Further fix a failure probability $\delta > 0$. If $n > C(d\log(d) + \log(1/(\alpha\delta)))/\alpha^2$, then Algorithm 1 when given the samples, α, and δ as input, runs in polynomial time and outputs k disjoint sets $\{B_i\}_{i\in k}$ so that with probability at least $1 - \delta$ the following are true, up to a permutation of indices of the output sets:

1. *$|S_i \triangle B_i| \leq 0.045n/k$ for every $i \in [k]$.*
2. *The mean of B_i is close to S_i: $\|\mu_{B_i} - \mu_i\|_2 = O(\sigma_i)$ for every $i \in [k]$.*

Algorithm 1 is given as input a minimum-weight parameter $\alpha \in [0.6/k, 1/k]$, and in polynomial-time it returns a list of exactly k sets, $\{B_i\}$, such that, up to a permutation, each B_i has a 95% overlap with the set S_i of samples drawn from the i^{th} mixture component P_i and that the mean μ_{B_i} of B_i is indeed close to the mean μ_i of P_i, under the minimal assumption that the means of the i^{th} and j^{th} clusters are separated by at least a large constant multiple of $(\sigma_i + \sigma_j)/\sqrt{\alpha}$. We note that *i*) the 95% overlap can be made an arbitrarily close constant to 1 by increasing the hidden constant in the separation assumption and adapting corresponding constants in the algorithm, and *ii*) we do not require any "no large sub-cluster condition" in the uniform mixture setting.

We also stress that Algorithm 1 does not require knowing k precisely, and only needs to know a lower bound α for $1/k$, which can be a (small) constant factor different.

We further remark that Item 1 above lower bounds the size of the union of all the B_is by $0.95n$, namely that at least 95% of all the points are clustered and returned. As noted above, the 95% can be made into any constant arbitrarily close to 100%, by increasing the constant C in the separation assumption. Alternatively, if we drop Item 2 in the theorem statement, namely that the requirement that the mean of B_i is indeed close to the mean μ_i of component P_i, then it is possible to return all the input samples in the output clustering.

Moreover, Theorem 1.1 holds even for almost-uniform mixtures, where each mixing weight $w_i \in [0.9/k, 1.1/k]$, and if $\alpha \in [0.7/k, \min_i w_i]$.

The situation of general (non-uniform) mixing weights is somewhat more complicated. For example, in the situation described below (also shown in Figure 2), even if we know the number k of components and even if we have infinitely many samples, it is information-theoretically impossible to reliably achieve a 90%-accurate clustering.

Example: Non-identifiability of general mixtures Consider a distribution consisting of 4 equal weight 0-covariance components, separated into 2 pairs. Each pair is at unit distance, and the two pairs are separated by a large distance. Suppose we are given that $k = 3$ and $\alpha = 1/4$, then there are two possible clusterings that disagree with each other by at least 25% of the total mass: either group the first pair as a large cluster with weight 1/2 and leave the second pair as two smaller clusters, or by symmetry we can group the second pair instead. It is allegorically impossible to determine which of these is the "true" ground truth clustering even with infinitely many samples from the mixture.

Given the above impossibility example, the question remains, what *is* possible given only the mixing weight lower bound parameter α, and a separation assumption of $C(\sigma_i + \sigma_j)/\sqrt{\alpha}$ between components P_i and P_j? The example highlights the core of the non-identifiability issue: an impossibility to identify which small subsets to group together. Consequently, we can perhaps hope to compute all the information in the ground truth clustering *except for* such subset grouping. That is, we can try to identify only the small subsets themselves. Motivated by this observation, we instead aim to return a *refinement* of the clustering: we will return a list of $\geq k$ subsets (which we will call sub-clusters), each of size at least $\approx \alpha n$, such that there exists some way of grouping the subsets into k larger clusters which then correspond to the ground truth mixture distribution.

For example, in the concrete setting of Figure 2, we could return the 4 small sub-clusters individually, which is a common refinement of the two possible clusterings shown in the figure. Furthermore, (as we will show) we can

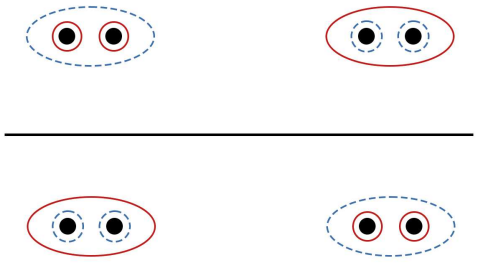

Figure 2: Two different ground truth clusterings for $k = 3$.

even guarantee that the returned subsets satisfy a pairwise separation guarantee similar to what we assume of our underlying mixture distribution.

Our main result (Theorem 1.2) of the paper shows that it is indeed possible to find an accurate refinement of the ground truth clustering, using $\tilde{O}(d/\alpha^2)$ samples and in polynomial time, with Algorithm 1. We define an accurate refinement below, as well as state a simplified version of our main theorem.

DEFINITION 1.1. (ACCURATE REFINEMENT OF GROUND TRUTH CLUSTERING) *Let $c > 0$ be an absolute constant. Suppose we draw n samples from the mixture distribution $D = \sum_{i=1}^{k} w_i P_i$, where each $w_i \geq \alpha$ and each P_i has mean μ_i and standard deviation σ_i. Let S_i be the set of samples drawn from P_i.*

An accurate refinement of the clustering S_i is a list of m disjoint sets of samples $\{B_j\}_{j\in[m]}$ for some $m \in [k, O(1/\alpha)]$, such that:

1. *The sets $B_1, \ldots, B_m$ each have size $|B_j| \geq 0.92\alpha n$ for all $j \in [m]$.*
2. *The indices $[m]$ can be partitioned into k sets $H_1, \ldots, H_k$, such that if $\mathcal{B}_i$ are defined as $\mathcal{B}_i := \cup_{j\in H_i} B_j$, the following hold:*
   (a) *$|S_i \setminus \mathcal{B}_i| \leq 0.045|S_i|$ for every $i \in [k]$.*
   (b) *$|\mathcal{B}_i \setminus S_i| \leq 0.03\alpha n$ for every $i \in [k]$.*
   (c) *For any $i \in [k]$ and any $j \in H_i$ we have that $\|\mu_{B_j} - \mu_i\|_2 \leq c\,\sigma_i\sqrt{|S_i|/|B_j|}$.*
   (d) *For any pair $j \neq j'$ we have that $\|\mu_{B_j} - \mu_{B_{j'}}\|_2 > 100\,c\,(\sigma_{B_j} + \sigma_{B_{j'}})/\sqrt{\alpha}$, where σ_{B_j} is the maximum standard deviation of B_j.*
3. *As a consequence of Item 2a we have that $|\cup_{j\in[m]} B_j| \geq 0.95n$, namely that 95% of the input points are classified into the output sets.*

Item 1 above says that each returned set must have size at least $\approx \alpha n$, given that each mixture component is supposed to have weight at least α. Item 2 captures the core idea of a refinement: there exists some way to combining the returned sets into sets $\mathcal{B}_1, \ldots, \mathcal{B}_k$, each corresponding to a mixture component $P_1, \ldots, P_k$, with the following guarantees. Items 2a and 2b say that the symmetric difference between the samples S_i drawn from component P_i and the set $\mathcal{B}_i$ is small. Item 2c says that each output set B_j must be close to the true mean of its corresponding mixture component P_i, with error scaling with σ_i as well as $\sqrt{|S_i|/|B_j|}$ — the larger B_j is, containing more samples in S_i, the closer μ_{B_j} should be to μ_i. Item 2d says that the returned subsets $\{B_j\}$ must themselves satisfy a mean separation akin to the one satisfied by the mixture components, up to a constant factor loss. Lastly, Item 3 guarantees that at least 95% of the samples are indeed classified and returned in one of the output sets.

REMARK 1.1. *The guarantees of Definition 1.1 imply that for every output set B_j there exists a true cluster S_i such that $|B_j \cap S_i| = |B_j| - |B_j \setminus S_i| \geq |B_j| - |\mathcal{B}_i \setminus S_i| \geq |B_j| - 0.03\alpha n \geq (1 - 0.03/0.92)|B_j| \geq 0.967|B_j|$, i.e. more than 96% of the points in the output set come from the true cluster S_i.*

THEOREM 1.2. (SIMPLIFIED VERSION OF THEOREM 3.1) *Consider a mixture distribution on $\mathbb{R}^d$, $D = \sum_{i=1}^k w_i P_i$ with unknown positive weights $w_i \geq \alpha$ for some known parameter $\alpha \in (0,1)$. Let μ_i and Σ_i be the (unknown) mean and covariance for each P_i, and assume that $\Sigma_i \preceq \sigma_i^2 \cdot I_d$ for all $i \in [k]$ (with σ_i being unknown) and $\|\mu_i - \mu_j\|_2 > C(\sigma_i + \sigma_j)/\sqrt{\alpha}$ for every $i \neq j$, for a sufficiently large constant C.*

There is an algorithm (Algorithm 1) which, when given α and n independent samples from D for n at least a sufficiently large multiple of $(d \log d + \log(1/(\alpha\delta)))/\alpha^2$, runs in polynomial time and with probability at least $1 - \delta$ (over the randomness of both the samples and the algorithm), outputs an accurate refinement clustering of these samples in the sense of Definition 1.1.

As in Theorem 1.1, for Definition 1.1 to be satisfied by the algorithm output, we can make the constant 0.92 in Item 1 arbitrarily close to 1, and the constants in Items 2a and 2b arbitrarily close to 0, if we increased the constant in the mean separation assumption in Theorem 1.2.

We also remark that the same algorithm (Algorithm 1) can even tolerate adversarial corruption in an $\Omega(\alpha)$-fraction of the samples. See the full theorem, Theorem 3.1, for the detailed statement.

Clustering under "no large sub-clusters" We can further guarantee that Algorithm 1 returns only k clusters (thereby corresponding exactly to the k ground truth components), if we also assume a "no large sub-cluster" condition à la [BKK22], stated in Section 1.1. Informally, the condition says that for any large subset S' of samples S_i drawn from the i^{th} mixture component, the standard deviation $\sigma_{S'}$ of S' is comparable to σ_{S_i}. This is intuitively the contrapositive of not having any large sub-clusters: a large sub-cluster can be understood as a large subset that is separated from the rest of the clusters, meaning that it has a substantially smaller covariance. Our condition below is qualitatively the same condition as that of [BKK22], but with a stronger quantitative requirement on the parameters of a sub-cluster. In Section 8.1, we show that such a stronger condition is information-theoretically necessary, due to our much weaker (and optimal) mixture separation assumption. Afterwards, in Section 8.2, we also show (see Corollary 1.1, an informal version of Corollary 8.1), that if the condition is satisfied, then there can only be one possible ground truth (i.e., there are no identifiability issues anymore) and thus Algorithm 1 indeed returns only one output set per real mixture component.

[NLSC condition] We say that the disjoint sets $S_1, \ldots, S_k$ of total size n satisfy the "No Large Sub-Cluster" condition with parameter α if for any cluster S_i and any subset $S' \subset S_i$ with $|S'| \geq 0.8\alpha n$, it holds that $\sigma_{S'} \geq 0.1\sigma_{S_i}$, where $\sigma_{S'}$ is the square root of the largest eigenvalue of the covariance matrix of S'.

COROLLARY 1.1. (INFORMAL VERSION OF COROLLARY 8.1) *If the samples S_i from the i^{th} mixture component jointly satisfy the NLSC assumption with parameter α across all $i \in [k]$, then Algorithm 1 returns exactly one sample set per mixture component (with high probability). As a consequence, if B_i is the output set corresponding to the i^{th} mixture component, then we have $\|\mu_{B_i} - \mu_i\| \leq O(\sigma_i)$, just like in Theorem 1.1.*

Later in Section 8.2, we also show that well-conditioned and high-dimensional log-concave distributions have samples that satisfy the NLSC condition with high probability. We remark that the high-dimensionality assumption is necessary: the thin-shell behavior of log-concave distributions in high dimensions is critical to satisfy our NLSC condition.

PROPOSITION 1.1. (INFORMAL VERSION OF PROPOSITION 8.2) *A sample of size $\tilde{O}(d/\alpha^2)$ drawn from a well-conditioned and high-dimensional log-concave distribution satisfies the NLSC condition (Section 1.1) with high probability.*

Before moving on to an overview of our algorithmic techniques, we emphasize that, even though we presented multiple results in multiple settings (uniform vs general weight mixtures, with and without the NLSC condition), they all apply to the same algorithm without needing any changes even in the hardcoded constants. Algorithm 1 does not need any knowledge of whether any of the conditions hold; it achieves the corresponding results automatically whenever the corresponding assumptions are satisfied.

1.2 Technical overview In this section, we give an overview of the components and techniques used in Algorithm 1, our main algorithm.

Since the mixture component means are assumed to be well-separated, Algorithm 1 works by finding a list of candidate mean vectors, each of which is close to a mixture component, with the entire list "covering" all the

components. Once we have such a list, it suffices to consider the Voronoi partition of the samples; that is, to assign each point to the cluster of the closest candidate mean. The mean separation assumption, along with the concentration of bounded covariance distributions, guarantee that such a Voronoi partition will be close to a refinement of the ground truth clustering.

The high-level idea of finding such a list of candidate mean vectors is to first generate a much larger (but still polynomially-sized) list which potentially contains candidates that are far from all mixture components, and then prune all the invalid candidate means out of the list. The first part is relatively straightforward, since there are standard list-decodable mean estimation algorithms for bounded-covariance distributions (e.g. [DKK+21]). The only minor complication is that, for these algorithms to return means with tight error guarantees, they need good upper bounds on the standard deviation of each mixture component. We thus first generate a list of possible standard deviations $\hat{s}$ (Proposition 2.1), and for each $\hat{s}$, run the list-decodable mean estimation algorithm. After this step, we have a list of candidate means such that, for each mixture component, there is at least one candidate mean close to it.

The next step is at the heart of our algorithm: to prune candidate means that are not sufficiently close to any mixture component (with distance threshold scaling with the standard deviation of the mixture component). A natural way to do this would be to test each candidate mean by trying to find its corresponding cluster and seeing if that exists. In particular, given a candidate mean $\hat{\mu}$ and candidate standard deviation $\hat{s}$, we would like to find a subset of at least an $\approx \alpha$-fraction of the samples whose covariance matrix is bounded by $O(\hat{s}^2) \cdot I_d$ and whose mean is within $O(\hat{s}/\sqrt{\alpha})$ of $\hat{\mu}$. If we can find this, it suggests that the cluster we are looking for actually exists.

Unfortunately, the natural approach of finding such a cluster is computationally hard, so we need to find appropriate relaxations to make it tractable. Immediately, to avoid computational hardness from integrality issues, we begin by allowing a weighted subset rather than an actual subset, which concretely is to find weights $w_i \in [0, 1]$ over each sample x_i, such that $\sum_i w_i$ is at least $\approx \alpha n$. This nearly turns our problem into a convex program. In particular, if we knew the mean of the cluster exactly, the covariance would be a linear function of $\{w_i\}_i$, making it a convex program. However, as we do not know the real mean, the covariance matrix centered at μ_w — the mean of the weighted cluster defined by $\{w_i\}_i$ — is no longer linear in $\{w_i\}_i$, and the constraint bounding its operator norm is no longer a convex constraint. So, instead, we compute the second moment matrix of $\{w_i\}_i$ centered at the candidate mean $\hat{\mu}$ (i.e. proportional to $\sum_i w_i(x_i - \hat{\mu})(x_i - \hat{\mu})^\top$). This gives us a convex program, but unfortunately one that might not be satisfiable even by a correct cluster C whose mean is indeed $O(\hat{s}/\sqrt{\alpha})$ close to the candidate mean $\hat{\mu}$: the second moment matrix of C would actually be $\mathrm{Cov}(C) + (\hat{\mu} - \mu_C)(\hat{\mu} - \mu_C)^\top$, and the latter term might contribute to an eigenvector of size as large as $\Omega(\hat{s}^2/\alpha)$, which is too large when α is small. We must therefore further relax our convex program. Instead of finding $\{w_i\}$ whose second moment matrix centered at $\hat{\mu}$ has operator norm bounded by $O(\hat{s}^2) \cdot I_d$, we constrain its $O(1/\alpha)$-Ky-Fan norm by $O(\hat{s}^2/\alpha)$. This new, final program (Program (4.1) in Section 4) is now both convex and satisfiable by a true cluster.

The next obstacle, however, is that a solution $\{w_i\}_i$ to the above convex program might not actually correspond to a true cluster or mixture component. In particular, if there are other clusters with standard deviation much smaller than $\hat{s}$, we might have found a solution that shares bits and pieces of these smaller clusters. This problem can only occur though if there are other clusters with standard deviation smaller than $\hat{s}$, but which are close to $\hat{\mu}$. Thus, we can avoid it by searching for clusters in *increasing* order of $\hat{s}$ and then throwing away any $\hat{\mu}$ that is within $O(\hat{s}/\sqrt{\alpha})$ of some previously un-pruned candidate mean. Formally, Lemma 4.1 shows that if $\hat{\mu}$ is far from all clusters with standard deviation smaller than $\hat{s}$, and if a solution to Program (4.1) exists for the pair $(\hat{\mu}, \hat{s})$, then the found solution must overlap substantially with a true cluster. An induction applying Lemma 4.1 repeatedly then shows that, after this pruning, all candidate means must be close to some true cluster, and that all clusters have candidate means close to them.

As discussed at the beginning of the section, we can now consider the Voronoi partition of the samples based on the candidate means. A few issues remain, that this partition does not satisfy the guarantees of Theorem 1.2. First, if there are too many candidate means remaining at this stage, a cluster in the Voronoi partition might be too small in size (Section 5). To solve this, we repeatedly remove candidate means whose Voronoi cluster is too small, noting that *i*) this never decreases the cluster size of un-removed candidate means, and *ii*) by the separation of the mixture components, we will never accidentally remove all candidate means close to any true cluster. Second, due to heavy-tailed noise and adversarial corruption, even for the Voronoi clusters that overlap well with true clusters, their means might be very far from the candidate means we started out with. We fix this using the standard filtering technique in robust statistics, removing at most 2% of the samples in each Voronoi cluster. Lastly, we

need to guarantee that the returned clusters also satisfy (up to constant factors) the same separation assumption we have on our underlying mixture distribution (Section 6). We enforce this again by removing candidate means whenever we detect a pair of (filtered) clusters that are too close to each other. Crucially, we carefully choose which corresponding candidate mean from the pair to remove, so that we never remove all the candidate means close to a true cluster.

1.3 Related work Here we survey the most relevant prior work on clustering mixture models and algorithmic robust statistics.

Mixture models A long line of work in theoretical computer science and machine learning has focused on developing efficient clustering methods for various mixture models (with mixtures of Gaussians being the prototypical example) under mean separation conditions; see, e.g. [Das99, AK01, VW02, AM05, KSV05, KK10, AS12, CSV17, HL18, KSS18, DKK+22b, BKK22].

Early work [AM05] gave an efficient spectral algorithm for clustering mixtures of bounded covariance Gaussians that succeeds under mean separation $\Theta((\sigma_i + \sigma_j)/\sqrt{\alpha})$ between components P_i and P_j, when the minimum mixing weight α is much smaller than $1/k$. However, even for the special case of uniform-weight k-mixtures of Gaussians (and log-concave distributions), their result requires a separation of $(\sigma_i + \sigma_j)\Omega(k)$ — instead of scaling with $\sqrt{k}$ — and, in fact, also has additional spurious terms in the separation containing a logarithmic dependence on the sample complexity n. It should be noted that the algorithm of [AM05] built on an earlier algorithm developed in [VW02], which only works for mixtures of spherical Gaussians. They can cluster under the weaker mean separation condition which (roughly) scales as $(1/\alpha)^{1/4}$; their separation condition also has a mild logarithmic dependence on the ambient dimensionality d. The works mentioned in this line all employ k-PCA as a core algorithmic technique; see the beginning of the introduction on why k-PCA fails in our heavy-tailed problem setting, under our fine-grained separation assumption.

[AS12] provided another spectral algorithm, designed to cluster mixtures of bounded covariance data. Their algorithm is able to cluster under a separation of (roughly) $\Omega(k)(\max_i \sigma_i)$. Their specific separation assumption can in fact be smaller than $\Omega(k)(\max_i \sigma_i)$ in certain instances, but the bound is not improvable to $o(k)(\max_i \sigma_i)$ in the worst case, contrasting the $\sqrt{k}$ dependence we achieve. More importantly, their separation condition between μ_i, μ_j scales with the maximum standard deviation $\max_i \sigma_i$, as opposed to the fine-grained pair-dependent sum $\sigma_i + \sigma_j$ achieved by our algorithm.

Recently, [DKK+22b] gave an almost linear-time clustering algorithm for mixtures of bounded covariance distributions. Their techniques inherently also cluster only under a $\max_i \sigma_i$ separation for the following reason: their algorithm runs a list-decodable mean estimation routine once (with the goal to list-decode the mean of a distribution with covariance $\Sigma \preceq (\max_i \sigma_i^2) \cdot I_d$) to generate a list of $O(1/\alpha)$ possible candidate cluster means. It then uses a coarse distance-based method to prune the means down to exactly k of them. As a result, their approach only works under a uniform separation between all pairs of components.

Another recent work [BKK22] also studied efficient clustering of mixtures of bounded covariance distributions, achieving a mean separation (between μ_i, μ_j) scaling with $\sigma_i + \sigma_j$. However, their separation assumption has a highly sub-optimal $\mathrm{poly}(1/\alpha)$ dependence, as well as an unnecessary logarithmic dependence on the sample complexity n. More importantly, their clustering algorithm inherently requires an additional structural condition on the components (which they term "no large sub-cluster" condition) beyond just bounded covariance, even for the special case of uniform-weight mixtures.

A related line of work has obtained clustering algorithms with significantly improved separation using more sophisticated algorithmic tools; see, e.g. [DKS18b, HL18, KSS18, DK20, LL22]. These works apply for families of distributions with controlled higher moments (e.g. sub-Gaussians), and in particular have no implication for the bounded covariance setting studied here.

Beyond clustering, a line of research developed efficient algorithms for learning mixtures of Gaussians, even in the presence of a constant fraction of corruptions; see, e.g. [MV10, BS10, BDH+20, Kan21, LM20, BDJ+20]. The aforementioned algorithms make essential use of the assumption that the mixture components are Gaussian.

Robust statistics and list-decodable learning Our paper is also related to the field of algorithmic robust statistics in high dimensions. Early work in the statistics community [Hub64, Tuk75] solidified the statistical foundations of this field. Unfortunately, the underlying estimators lead to exponential time algorithms. A line of work in computer science, starting with [DKK+16, LRV16], developed polynomial-time algorithms for a wide range of robust high-dimensional estimation tasks. The reader is referred to the recent book [DK23] for an overview.

The list-decodable learning setting that we leverage in this work was defined, in a somewhat different context, in [BBV08]. [CSV17] gave the first polynomial-time algorithm for the task of list-decodable mean estimation under a bounded covariance assumption. Specifically, if the clean data has covariance bounded by the identity, their achieved error guarantee is $\tilde{O}(1/\sqrt{\alpha})$. This error bound was slightly refined in [CMY20] to $O(1/\sqrt{\alpha})$ with an asymptotically faster algorithm; a matching information-theoretic lower bound of $\Omega(1/\sqrt{\alpha})$ was shown in [DKS18a]. We note that [CSV17] also obtains a corollary for clustering mixtures, but their method requires sub-Gaussian components, and it only outputs a clustering refinement with $O(1/\alpha)$ subsets. Finally, [DKK+22b], building on [DKK20a, DKK+20b], developed an almost-linear time algorithm for this task; in fact, they built their clustering result for mixtures of bounded covariance distributions via a reduction to list-decodable mean estimation.

In this work, we also use list-decodable mean estimation as a blackbox (Fact 2.4 in Section 2). An important difference compared to prior work is that our processing of the candidate means is substantially more involved, which is required due to our fine-grained separation assumption.

Finally, we point out other work which developed efficient list-decodable mean estimators with significantly improved error guarantees under much stronger distributional assumptions [DKS18a, KSS18, DKK+22a].

1.4 Organization Section 2 gives basic notations and facts that we use in the rest of the paper. Section 3 states our main algorithm (Algorithm 1) as well as the full version of our main result (Theorem 3.1). Sections 4 to 6 analyzes the three main steps of the algorithm. Section 7 uses the guarantees from the prior three sections to prove our main result. Finally, Section 8 discusses the implications of the no large sub-cluster condition in our problem setting.

2 Preliminaries

In this section, we state useful notations and facts that the rest of the paper depends on.

2.1 Notation For a vector v, we let $\|v\|_2$ denote its ℓ_2-norm. We use I_d to denote the $d \times d$ identity matrix; We will drop the subscript when it is clear from the context. For a matrix A, we use $\|A\|_{\mathrm{F}}$ and $\|A\|_{\mathrm{op}}$ to denote the Frobenius and spectral (or operator) norms, respectively. We use $\|A\|_{(k)}$ to denote the Ky-Fan norm which is defined as $\|A\|_{(k)} = \sum_{j=1}^{k} s_j(A)$, where $s_j(A)$ for $j = 1, \ldots, k$ are the first k singular values of A. If V is a subspace, we denote by Proj_V its the orthogonal projection matrix.

We use $X \sim D$ to denote that a random variable X is distributed according to the distribution D. We use $\mathcal{N}(\mu, \Sigma)$ for the Gaussian distribution with mean μ and covariance matrix Σ. For a set S, we use $X \sim S$ to denote that X is distributed uniformly at random from S. When S is a set of points in $\mathbb{R}^d$, we will use the shorter notation $\mu_S := \mathbb{E}_{X \sim S}[X], \mathrm{Cov}(S) := \mathbb{E}_{X \sim S}[(X - \mu_S)(X - \mu_S)^\top]$, and $\sigma_S := \sqrt{\|\mathrm{Cov}(S)\|_{\mathrm{op}}}$.

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

2.2 Deterministic conditions and useful facts

Stability condition Our algorithm will succeed if the following condition is satisfied for the samples of each true cluster. The condition, referred to as "stability", is standard in algorithmic robust statistics. Intuitively, it requires that any large subset of the dataset has mean and covariance that do not shift significantly. We provide the definition below. In the fact that follows, we state that large sets of points from bounded covariance distributions indeed satisfy the stability condition with high probability.

Definition 2.1. (Stability condition) *For $C > 0$ and $\epsilon \in (0, 1/2)$, a multiset S of m points $x_1, \ldots, x_m$ in $\mathbb{R}^d$ is called (C, ϵ)-stable with respect to $\mu \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}^+$ if, for any weights $w_1, \ldots, w_m \in [0, 1]$ with $\sum_{x_i \in S} w_i \geq (1 - \epsilon)m$ it holds:*

- $\left\| \frac{1}{\sum_{x_i \in S} w_i} \sum_{x_i \in S} w_i x_i - \mu \right\|_2 \leq C\sigma\sqrt{\epsilon}$
- $\Sigma_{w,\mu} \preceq C^2\sigma^2 \cdot I_d$,

where $\Sigma_{w,\mu} := \frac{1}{\sum_{x_i \in S} w_i} \sum_{x_i \in S} w_i (x_i - \mu)(x_i - \mu)^\top$.

FACT 2.1. (SAMPLE COMPLEXITY OF STABILITY [DKP20]) *Let S be a set of m points drawn i.i.d. from a distribution on $\mathbb{R}^d$ with mean μ and covariance $\Sigma \preceq \sigma^2 \cdot I_d$. If $m \gg (d\log(d) + \log(1/\delta))/\min\{\epsilon, \alpha\}$ then, with probability $1-\delta$, there exists a $(1-0.001\alpha)m$-sized subset S' of S that is $(100, \epsilon)$-stable with respect to $\mu \in \mathbb{R}^d$ and σ.*

Facts from robust statistics We record the following facts that will be useful later on. First, we recall in Fact 2.2 a stability-based filtering algorithm that, given any stable set of samples with bounded covariance and with 4% of its points arbitrarily corrupted, removes 4% of the points in a way that the resulting output set is guaranteed to have bounded covariance and mean close to the true one.

DEFINITION 2.2. (STRONG CONTAMINATION MODEL) *Given a parameter $0 < \epsilon < 1/2$, the strong adversary operates as follows: The algorithm specifies a set of n samples, then the adversary inspects the samples, removes up to ϵn of them and replaces them with arbitrary points. The resulting set is given as input to the learning algorithm. We call a set ϵ-corrupted if it has been generated by the above process.*

FACT 2.2. (FILTERING; SEE, E.G. [DK23]) *There exists an algorithm for which the following is true: Let $\delta \in (0,1)$ be a parameter. Let S be a set of points in $\mathbb{R}^d$ that is (C, ϵ)-stable with respect to μ and σ for some $C > 0$ and $\epsilon \leq 0.04$. Let T be an ϵ-corrupted version of S (cf. Definition 2.2) and assume $|T| \gg \log(1/\delta)$. Then the algorithm having as input any set T of the above form and δ terminates in time $\mathrm{poly}(|T|, d)$ and returns a subset $T' \subseteq T$ such that, with probability at least $1-\delta$, the following hold:*

- $|T'| \geq (1-\epsilon)|T|$.
- $\|\mu_{T'} - \mu\|_2 \leq 10C\sigma\sqrt{\epsilon}$.
- $\Sigma_{T',\mu} \preceq 10C^2\sigma^2 \cdot I_d$.

The following fact states that taking subsets of a set S with bounded covariance does not shifts the mean significantly. This (or its contrapositive version) will be used in a lot of the core arguments. In particular, one corollary of this fact is Lemma 2.1, stating that subsets of stable sets are also stable with worse parameters. This will be useful for applying the aforementioned filtering algorithm at the very last step of our main algorithm to ensure that the final clusters have means and covariances that are close to what they should be. For completeness, we provide a proof of Lemma 2.1 in Section A.

FACT 2.3. *Let S be a multiset, and denote by μ_S, Σ_S the mean vector and covariance matrix of the uniform distribution on S. If S satisfies $\Sigma_S \preceq \sigma^2 \cdot I_d$ and $w_x \in [0,1]$ are weights for the points $x \in S$ that satisfy $\sum_{x\in S} w_x \geq \alpha|S|$, then we have that*

$$\left\| \frac{\sum_{x\in S} w_x x}{\sum_{x\in S} w_x} - \mu_S \right\|_2 \leq \frac{\sigma}{\sqrt{\alpha}} \,.$$

LEMMA 2.1. *Let S be a set of points that is (C, ϵ)-stable with respect to μ and σ for some $C \geq 1$ and $\epsilon < 1/2$. Then, any subset $S' \subseteq S$ with $|S'| \geq \alpha|S|$ is $(1.23C/\sqrt{0.04\alpha}, 0.04)$-stable with respect to μ and σ.*

We finally state in Proposition 2.1 and fact 2.4 the subroutines that we will use for creating a list of candidate covariances and means of the true clusters. We defer the proof of Proposition 2.1 to Section A. The algorithm consists of simply returning a list with all the values starting from $\|x-y\|_2$ down to $\|x-y\|_2/(2|S|^2)$ in multiples of $\sqrt{2}$, for all pairs of points x, y. By the definition of the covariance matrix as $\mathrm{Cov}(S) = \frac{1}{2|S|^2}\sum_{x,y\in S}(x-y)(x-y)^\top$, one of these quantities should be within a factor of two from $\|\mathrm{Cov}(S)\|_{\mathrm{op}}$.

PROPOSITION 2.1. *Let T be a set of m points in $\mathbb{R}^d$. There is a $\mathrm{poly}(m,d)$-time algorithm that outputs a list of size $O(m^2\log(m))$ that for any $S \subseteq T$ contains an estimate $\hat{s}$ such that $\|\mathrm{Cov}(S)\|_{\mathrm{op}} \leq \hat{s}^2 \leq 2\|\mathrm{Cov}(S)\|_{\mathrm{op}}$.*

FACT 2.4. (LIST-DECODABLE MEAN ESTIMATION; SEE, E.G. [DKK+21]) *Let S be a multi-set in $\mathbb{R}^d$ that satisfies $\frac{1}{m}\sum_{x\in S}(x-\mu)(x-\mu)^\top \preceq \sigma^2 \cdot I_d$ for some $\mu \in \mathbb{R}^d$ and $\sigma > 0$, and T be another multi-set in $\mathbb{R}^d$ such that $S \subseteq T$ and $|S| \geq \alpha|T|$. There exists an algorithm and absolute constant $C > 1$, that on any input T of the aforementioned form and the standard deviation parameter σ, the algorithm runs in polynomial time and returns a $O(1/\alpha)$-sized list of vectors that contains at least one vector $\hat{\mu}$ such that $\|\hat{\mu} - \mu\|_2 \leq C\sigma/\sqrt{\alpha}$.*

3 Main algorithm and result

We present our main algorithm in the paper, Algorithm 1, which follows the outline described in Section 1.2. Lines 1 and 2 first generates a list of plausible component means and standard deviations. Then, Line 4 is responsible for pruning the list such that every remaining candidate mean is indeed close to a true component. This is useful because the Voronoi partition of the samples based on such a list is an accurate refinement of the ground truth clustering. Lines 5 and 6 further prune the list, to ensure that the returned refinement have subsets that are not too small (at least $\approx \alpha n$ in size) and that they are well-separated. Finally, Line 7 returns filtered versions of the final Voronoi partition, in order to filter out adversarial and heavy-tailed outliers, to make sure that the mean of each returned subset is reasonably close to its corresponding mixture component.

Algorithm 1 Clustering algorithm

Input: Parameter $\alpha \in (0,1)$, and multi-set T of n points in $\mathbb{R}^d$ for which there exists a ground truth clustering $S_1, \ldots, S_k$ according to the assumptions of Theorem 3.2.
Output: Disjoint subsets of T that form an accurate refinement (cf. Definition 1.1) of the ground truth clustering.

1. Generate a list L_{stdev} of candidate standard deviations using the algorithm from Proposition 2.1.
2. Generate a list of candidate means, L_{mean}, by applying the list-decoding algorithm of Fact 2.4 for each candidate s in the list L_{stdev}, and appending the output of each run to L_{mean}.
3. Initialize $L \leftarrow \emptyset$.
4. For every $s \in L_{\text{stdev}}$ in increasing order of s:
   (a) For every $\mu \in L_{\text{mean}}$:
      i. If $\|\mu - \hat{\mu}\|_2 > 99Cs/\sqrt{\alpha}$ for all $\hat{\mu} \in L$, decide the satisfiability of the convex program defined in (4.1) in Section 4.
      ii. If satisfiable, add μ to the list L.
5. $L' \leftarrow \textsc{SizeBasedPruning}(L, T, \alpha)$. ▷ cf. Algorithm 2
6. $L'' \leftarrow \textsc{DistanceBasedPruning}(L', T, \alpha)$. ▷ cf. Algorithm 4
7. Output $\textsc{FilteredVoronoi}(L'', T)$.

We will now state the full version of our main theorem (Theorem 3.1). As discussed in the introduction, our algorithm can also handle a small amount of adversarial corruption in the samples. Recall the "Strong Contamination Model" from Definition 2.2, commonly used in the robust statistics literature, capturing the powerful adversary that our algorithm can handle. In that model, a computationally unbounded adversary can inspect and edit a small fraction of the input points however it wants.

We now give the version of our main result (Theorem 3.1) that works under this adversarial corruption. The statement says that Algorithm 1 outputs an accurate refinement of the ground truth clustering of the samples: a list of sets $\{B_j\}_{j\in[m]}$ for some $m \in [k, O(1/\alpha)]$, each of which has size at least $0.92\alpha n$, such that the sets are 90% close to a refinement of the ground truth clustering. We also ensure that the output clusters also enjoy a mean separation guarantee that is qualitatively similar to the one at the distributional level (Item 2d below). Furthermore, if the output set B_j corresponds a subset of the samples S_i drawn from component i, then the mean μ_{B_j} of B_j is close to μ_i (Item 2c), by a distance bound that depends on the ratio $|S_i|/|B_j|$, namely that the larger the fraction that B_j covers in S_i, the closer their means are.

THEOREM 3.1. (MAIN RESULT, FORMAL STATEMENT) *Consider a mixture distribution on $\mathbb{R}^d$, $D = \sum_{i=1}^k w_i P_i$ with unknown positive weights $w_i \geq \alpha$ for some known parameter $\alpha \in (0,1)$. Let μ_i and Σ_i be the (unknown) mean and covariance for each P_i, and assume that $\Sigma_i \preceq \sigma_i^2 \cdot I_d$ for all $i \in [k]$ (with σ_i being unknown) and $\|\mu_i - \mu_j\|_2 > 591\, c^2(\sigma_i + \sigma_j)/\sqrt{\alpha}$ for every $i \neq j$, for a sufficiently large constant c.*

Let a set T_0 of n samples drawn from D independently, and let S_i be the samples from the i^{th} mixture

component. Let T be any 0.01α-corruption of T_0 according to the model defined in Definition 2.2. Further fix a failure probability $\delta \in (0,1)$.

If $n \gg (d\log(d) + \log(1/(\alpha\delta)))/\alpha^2$, then on input the set T and the parameter α, with probability at least $1-\delta$ (over the randomness of both the samples and the algorithm), Algorithm 1 runs in time $\mathrm{poly}(nd/\alpha)$ and outputs $m \le 1/(0.92\alpha)$ disjoint sets $\{B_j\}_{j\in[m]}$ such that:

1. *The output sets $B_1, \dots, B_m$ each have size $|B_j| \ge 0.92\alpha n$ for all $j \in [m]$.*
2. *The set of indices $[m]$ can be partitioned into k subsets $H_1, \dots, H_k$, such that if $\mathcal{B}_i$ are defined as $\mathcal{B}_i := \cup_{j\in H_i} B_j$, the following hold:*
   (a) *$|S_i \setminus \mathcal{B}_i| \le 0.045|S_i|$ for every $i \in [k]$.*
   (b) *$|\mathcal{B}_i \setminus S_i| \le 0.03\alpha n$ for every $i \in [k]$.*
   (c) *For any $i \in [k]$ and any $j \in H_i$, we have that $\|\mu_{B_j} - \mu_i\|_2 \le c\,\sigma_i\sqrt{|S_i|/|B_j|}$.*
   (d) *For any pair $j \ne j'$, we have that $\|\mu_{B_j} - \mu_{B_{j'}}\|_2 > 366\,c\,(\sigma_{B_j} + \sigma_{B_{j'}})/\sqrt{\alpha}$.*
3. *As a consequence of Item 2a, we have that $|\cup_{j\in[m]} B_j| \ge 0.95n$, namely that 95% of the input points are classified into the output sets.*

Before we prove Theorem 3.1, we first show Theorem 1.1 concerning the special case of uniform-weight mixture distributions. As we show below, Theorem 1.1 is a direct consequence of Theorem 3.1.

Proof. [Proof of Theorem 1.1] Theorem 1.1 is a special case of Theorem 3.1. It can be readily checked that all the assumptions of Theorem 3.1 are satisfied for $\alpha \in [0.6/k, 1/k]$. Moreover, the sizes $|S_i|$ have expected value n/k, and thus by the Chernoff-Hoefding bound it must be the case that $0.999n/k \le |S_i| \le 1.001n/k$ with high probability. Since the sets B_j ($j \in [m]$) mentioned in Theorem 3.1 are disjoint with sizes $|B_j| \ge 0.92\alpha n > 0.552n/k$ (Item 1 of the theorem statement) and their unions $\mathcal{B}_i$ corresponding to i^{th} cluster satisfy $|\mathcal{B}_i \setminus S_i| \le 0.03n/k$ (Item 2b), this means that each $\mathcal{B}_i$ has size $|\mathcal{B}_i| \le 1.031n/k$ and thus every $\mathcal{B}_i$ must consist of exactly one of the B_j's. Thus, the algorithm outputs exactly k sets $B_1, \dots, B_k$, where (up to a permutation of the labels) B_i corresponds to the i^{th} mixture component. Then, Items 2a and 2b of Theorem 3.1 imply that $|S_i \triangle B_i| \le 0.044\max(|S_i|, \alpha n) \le 0.045n/k$ since $\alpha \le 1/k$ and $|S_i| \le 1.001n/k$. Item 2 of Theorem 1.1 follows from Item 2c of Theorem 3.1 after noting that

$$|B_j| \ge |B_j \cap S_j| \ge |S_j| - |S_j \triangle B_j| \ge 0.999n/k - 0.044n/k = 0.955n/k \ge (0.955/1.001)|S_j|\,.$$

This completes the proof of Theorem 1.1.
∎

It remains to analyze Algorithm 1, which we do in Sections 4 to 7. Section 4 states and analyzes the convex program used in Line 4 of the algorithm, as well as the guarantees-by-induction right after Line 4 finishes. Section 5 gives Algorithm 2 used in Line 5, which ensures that every set in the Voronoi partition computed from the remaining candidate means is of size at least $\approx \alpha n$. Section 6 gives Algorithm 4 used in Line 6, which in turn ensures that the Voronoi partition from the remaining means corresponds to a refinement with well-separated subsets. Finally, in Section 7, we prove Theorem 3.2 stated below, which is a version of Theorem 3.1 conditioned on samples satisfying deterministic stability conditions (cf. Section 2.2).

THEOREM 3.2. (STABLE SET VERSION OF THEOREM 3.1) *Let $d \in \mathbb{Z}_+$, $\delta, \alpha \in (0,1)$ be parameters, and let $C > 1$ be a sufficiently large absolute constant. Consider a (multi-)set T of $n \gg \log(1/(\alpha\delta))/\alpha$ points in $\mathbb{R}^d$ with k disjoint subsets $S_1, \dots, S_k \subseteq T$, where $|\cup_i S_i| \ge (1-0.02\alpha)|T|$, satisfying the following for each $i \in [k]$: (i) $|S_i| \ge 0.97\alpha n$, (ii) S_i is $(C, 0.04)$-stable (cf. Definition 2.1) with respect to mean μ_i and maximum standard deviation parameter σ_i (where μ_i, σ_i are unknown), (iii) for every pair $i \ne j$ we have $\|\mu_i - \mu_j\|_2 > 10^5 C^2(\sigma_i + \sigma_j)/\sqrt{\alpha}$. Then Algorithm 1 on input T, α, runs in $\mathrm{poly}(nd/\alpha)$-time and with probability at least $1-\delta$ (over the internal randomness of the algorithm) outputs $m \le 1.07/\alpha$ disjoint sets $\{B_j\}_{j\in[m]}$ that satisfy the following:*

1. *The output sets $B_1, \dots, B_m$ are disjoint and have size $|B_j| \ge 0.92\alpha n$ for all $j \in [m]$.*
2. *The set $[m]$ can be partitioned into k sets $H_1, \dots, H_k$, such that if $\mathcal{B}_i$ are defined as $\mathcal{B}_i := \cup_{j\in H_i} B_j$, the following hold:*

(a) $\mathcal{B}_i \neq \emptyset$ *for* $i \in [k]$.

(b) $|S_i \setminus \mathcal{B}_i| \leq 0.033|S_i|$ *for every* $i \in [k]$.

(c) $|\mathcal{B}_i \setminus S_i| \leq 0.03\alpha n$ *for every* $i \in [k]$.

(d) For any $i \in [k]$ *and any* $j \in H_i$ *we have that* $\|\mu_{B_j} - \mu_i\|_2 \leq 13C\sigma_i\sqrt{|S_i|/|B_j|}$.

(e) For any pair $j \neq j'$ *we have that* $\|\mu_{B_j} - \mu_{B_{j'}}\|_2 > 4761C(\sigma_{B_j} + \sigma_{B_{j'}})/\sqrt{\alpha}$.

To end this section, we prove that Theorem 3.1 does indeed follow from Theorem 3.2.

Proof. [Proof of Theorem 3.1] Before we begin the proof, we note that, despite the notation S_i appearing in both Theorem 3.1, Theorem 3.2, they mean slightly different sets in the context. In Theorem 3.1, the S_i sets refer to all the samples generated from the i^{th} mixture component, prior to any corruptions. On the other hand, when applying Theorem 3.2, we will instead consider large subsets of the samples that are stable. For this proof, we will use the notation $\tilde{S}_1, \ldots, \tilde{S}_k$ to denote the samples from the i^{th} component, and we will later choose S_i in the context of Theorem 3.2 to be large subsets of $\tilde{S}_i$ that are stable, essentially guaranteed by Fact 2.1.

We now check explicitly that with high probability (i.e. at least $1 - \delta/2$), the set T in Theorem 3.1 has subsets $S_1, \ldots, S_k$ satisfying the assumptions of Theorem 3.2. We choose the constant c that appears in the statement of Theorem 3.1 to be the same as $13C$ in Theorem 3.2.

We can think of the mixture model as first deciding the number of samples drawn from each component, and then generating each set of samples by drawing i.i.d. samples from the component. Since each component has weight at least α and the number of samples is $n \gg (d\log(d) + \log(1/(\alpha\delta)))/\alpha^2$, by Chernoff-Hoeffding bounds and a union bound, with probability at least $1 - \delta/100$, $|\tilde{S}_i| \geq 0.999\alpha n \gg (d\log(d) + \log(1/(\alpha\delta)))/\alpha$ for all $i \in [k]$. Then, by Fact 2.1 applied to the samples $\tilde{S}_i$ from each component, and a union bound over all components, we have that with probability at least $1 - \delta/100$, there exist subsets $S_i' \subseteq \tilde{S}_i$ for $i \in [k]$ with $|S_i'| \geq (1 - 0.001\alpha)|\tilde{S}_i|$ that are $(C/2, 0.05)$-stable with respect to μ_i and σ_i. This, combined with the fact that the adversary can corrupt only $0.01\alpha n$ points, means that if we let S_i for $i \in [k]$ be the sets $S_i' \cap T$ (i.e. parts of S_i' that are not corrupted by the adversary), the assumptions of Theorem 3.2 that $|\cup_i S_i| \geq (1 - 0.02\alpha)|T|$, $|S_i| \geq 0.97\alpha n$ and S_i being $(C, 0.04)$-stable are all satisfied with probability at least $1 - \delta/2$.

Continuing our check of the assumptions of Theorem 3.2, the separation assumption $\|\mu_i - \mu_j\|_2 > 10^5C^2(\sigma_i + \sigma_j)/\sqrt{\alpha}$ trivially follows from the corresponding assumption in Theorem 3.1 (and the fact that we have chosen $c = 13C$).

The conclusion of Theorem 3.2 is guaranteed to hold with probability $1 - \delta/2$ over the randomness of the algorithm. By a union bound over the failure event of the Theorem 3.2 and the failure event of Fact 2.1 (which are both at most $\delta/2$), we get that the conclusion holds with probability at least $1 - \delta$ over both the randomness of the samples and the randomness of the algorithm.

We finally check that the conclusion of Theorem 3.2 implies the conclusion in Theorem 3.1. Item 1 of Theorem 3.1, stating that $|B_j| \geq 0.92\alpha n$, is the same as in Theorem 3.2. Item 2a of Theorem 3.1, stating that $|\tilde{S}_i \setminus \mathcal{B}_i| \leq 0.034|\tilde{S}_i|$ is derived from Item 2b of Theorem 3.2 as follows: $|\tilde{S}_i \setminus \mathcal{B}_i| \leq |S_i \setminus \mathcal{B}_i| + |\tilde{S}_i \setminus S_i| \leq 0.033|S_i| + |\tilde{S}_i \setminus S_i| \leq 0.033|\tilde{S}_i| + 0.001|\tilde{S}_i| + 0.01\alpha n = 0.045|\tilde{S}_i|$, where the second step used Item 2b of Theorem 3.2, the third step used that $S_i \subseteq S_i' \subseteq \tilde{S}_i$, $|S_i'| \geq (1 - 0.001\alpha)|\tilde{S}_i|$ and that the adversary can edit at most $0.01\alpha n$ points. The last step used that $|\tilde{S}_i| \geq 0.999\alpha n$. Item 2b of Theorem 3.1, stating that $|\mathcal{B}_i \setminus \tilde{S}_i| \leq 0.03\alpha n$ can be derived from Item 2c of Theorem 3.2 as follows: $|\mathcal{B}_i \setminus \tilde{S}_i| \leq |\mathcal{B}_i \setminus S_i| \leq 0.03\alpha n$, the first step is because $S_i \subseteq \tilde{S}_i$ and the second step uses the guarantee from Theorem 3.2. The last two parts of the conclusion of Theorem 3.1 follow similarly.

□

4 Candidate mean pruning via convex programming

This section states and analyzes the convex program (in (4.1) below) used in Line 4 of Algorithm 1. Line 4 assumes that for all mixture components P_i and its stable subset of samples S_i, the list L_{stdev} contains an $\hat{s} \in [\sigma_{S_i}, \sqrt{2}\sigma_{S_i}]$ by Proposition 2.1, and the list L_{mean} contains a $\hat{\mu}$ with $\|\hat{\mu} - \mu_{S_i}\| \leq O(\sigma_{S_i}/\sqrt{\alpha})$ by Fact 2.4—recall that we denote by $\sigma_{S_i} = \sqrt{\|\operatorname{Cov}(S_i)\|_{\text{op}}}$ the maximum standard deviation of the points in S_i. At the end of the section, we will then guarantee that, after the double-loop of Line 4 finishes, the list $L \subset L_{\text{mean}}$ also contains mean estimates close to every S_i, and moreover, every $\hat{\mu} \in L$ is close to some S_i.

We will use the notation of Theorem 3.2 in the following. Recall that we denote by T the input set of samples. For every vector $\mu \in \mathbb{R}^d$ and $s > 0$, we define the convex program below, where the constant C is the same constant appearing in Fact 2.4.

$$\begin{aligned} &\text{Find:} && w_x \in [0,1] \text{ for all } x \in T \\ &\text{s.t.:} && \left\| \sum_{x \in T} w_x (x-\mu)(x-\mu)^\top \right\|_{(1/\alpha)} \le \frac{2C^2 s^2}{\alpha} \sum_{x \in T} w_x, \\ & && 0.97\alpha n \le \sum_{x \in T} w_x \end{aligned} \tag{4.1}$$

The following lemma (Lemma 4.1) analyzes the convex program (4.1). If for some standard deviation candidate s and candidate mean μ, we are guaranteed that μ is far from all S_j with $\sigma_{S_j} \ll s$, and furthermore, there is a solution for the program (4.1), then every S_j whose mean is far away from μ has negligible overlap with the solution $\{w_x\}_x$. The first assumption corresponds to the check in Line 4(a)i—Lemma 4.1 will be used in the context of an induction over the outer loop, where we assume that all clusters S_j with $\sigma_{S_j} \le s$ have some "representative" candidate mean in L that is close to μ_{S_j}. The conclusion of Lemma 4.1 certifies that μ must be close to some true cluster S_i if Line 4(a)i passes, thus allowing us to safely add this μ to the list L.

LEMMA 4.1. *Consider the setting of Theorem 3.2 and consider an arbitrary pair of parameters $\mu \in \mathbb{R}^d$ and $s > 0$. Suppose that: (i) for every cluster S_j with $\sigma_{S_j} < s/100$ it holds that $\|\mu - \mu_{S_j}\|_2 \ge 46Cs/\sqrt{\alpha}$, and (ii) a solution w_x for $x \in T$ to the program defined in (4.1) exists. Then there exists a unique true cluster S_i with $\sigma_{S_i} \ge s/100$ such that $\|\mu_{S_i} - \mu\|_2 \le 4600C\sigma_{S_i}/\sqrt{\alpha}$.*

Proof. By the constraint $0.97\alpha n \le \sum_{x \in T} w_x$ of the program, it suffices to show that all clusters S_j with $\|\mu_{S_i} - \mu\|_2 > 4600C\sigma_{S_i}/\sqrt{\alpha}$ have (in the aggregate) small overlap with the solution of the program $\{w_x\}_x$, namely, that $\sum_{j : \|\mu_{S_i} - \mu\|_2 > 4600C\sigma_{S_i}/\sqrt{\alpha}} \sum_{x \in S_j} w_x \le 0.01 \sum_{x \in T} w_x$. In order to show this, we consider a number of cases. We first consider clusters that have standard deviation at most $s/100$ (which satisfy assumption (i) in the lemma statement), and then clusters with bigger standard deviation. At the end, we combine the two analyses to conclude the proof of the lemma.

For clusters S_j with $\sigma_{S_j} < s/100$: We first show that across cluster indices j with $\sigma_{S_j} < s/100$, we must have that $\sum_{j\,:\,\sigma_{S_j} < s/100} \sum_{x \in S_j} w_x \le 0.003 \sum_{x \in T} w_x$. For every cluster index $j \in [k]$, we denote by v_j the unit vector in the direction of $\mu_{S_j} - \mu$ and consider the partition $S_j = S_j^{\le} \cup S_j^{>}$, where $S_j^{>} = \{x \in S_j : v_j^\top (x - \mu) > 45Cs/\sqrt{\alpha}\}$ and $S_j^{\le} = S \setminus S_j^{>}$. That is, $S_j^{>}$ is the part of the cluster S_j that is far away from μ in the direction $\mu_{S_j} - \mu$ and $S_j^{\le}$ the points that are close. We bound the overlap of the solution $\{w_x\}_{x \in T}$ with each kind of points individually in Claims 1 and 2 that follow. The argument for the points that are far away is that a large number of them would cause a violation of the Ky-Fan norm constraint of the program defined in (4.1). For the points that are close to μ, the argument is that a large number of these points would move the mean of the cluster close to μ and violate our assumption that $\|\mu - \mu_{S_j}\|_2 \ge 46Cs/\sqrt{\alpha}$ for every cluster S_j with $\sigma_{S_j} < s/100$.

CLAIM 1. $\sum_{j=1}^{k} \sum_{x \in S_j^{>}} w_x \le 0.001 \sum_{x \in T} w_x$.

Proof. This follows by the Ky-Fan norm constraint of the the program defined in (4.1). Let V be an arbitrary $(1/\alpha)$-dimensional subspace containing the span of $v_1, \ldots, v_k$ (where the v_i's are defined as the unit vectors in the directions $\mu_{S_i} - \mu$ for $i \in [k]$). Then we have that:

$$\begin{aligned} \text{(by the Ky-Fan norm constraint)} \qquad \frac{2C^2 s^2}{\alpha} \sum_{x \in T} w_x &\ge \left\| \sum_{x \in T} w_x (x-\mu)(x-\mu)^\top \right\|_{(1/\alpha}) \\ \text{(by def. of the Ky-Fan norm)} \qquad &\ge \operatorname{tr}\left(\sum_{x \in T} w_x \mathrm{Proj}_V (x-\mu)(x-\mu)^\top \mathrm{Proj}_V^\top \right) \end{aligned}$$

$$= \sum_{x\in T} w_x \left\|\mathrm{Proj}_V(x-\mu)\right\|_2^2$$

$$\geq \sum_{j=1}^{k}\sum_{x\in S_j^{>}} w_x \left\|\mathrm{Proj}_V(x-\mu)\right\|_2^2$$

(since $v_j \in V$) $$\geq \sum_{j=1}^{k}\sum_{x\in S_j^{>}} w_x (v_j(x-\mu))^2$$

(by definition of set $S_j^{>}$) $$\geq 2000C^2 \sum_{j=1}^{k}\sum_{x\in S_j^{>}} w_x \frac{1}{\alpha}s^2 \,.$$

The above implies that $\sum_{j=1}^{k}\sum_{x\in S_j^{>}} w_x \leq 0.001\sum_{x\in T} w_x$. ∎

Claim 2. *We have that* $\sum_{j\,:\,\sigma_{S_j}<s/100}\sum_{x\in S_j^{\leq}} w_x \leq 0.002\sum_{x\in T} w_x$.

Proof. Let $\alpha_j^{\leq} := \left(\sum_{x\in S_j^{\leq}} w_x\right)/|S_j^{\leq}|$ denote the intersection of the solution with $S_j^{\leq}$; the part of the j-th cluster that is close to μ. We will show that $\sum_{j\in[k]:\sigma_{S_j}<s/100}\alpha_j^{\leq} \leq 0.001$.

Since $S_j^{\leq}$ contains by definition the points $x \in S_j$ that $v_j^{\top}(x-\mu) \leq 45Cs/\sqrt{\alpha}$, then their mean satisfies $v_j^{\top}(\mu_{S_j^{\leq}}-\mu) \leq 45Cs/\sqrt{\alpha}$. Then we can write

$$v_j^{\top}(\mu_{S_j}-\mu_{S_j^{\leq}}) = v_j^{\top}(\mu_{S_j}-\mu) - v_j^{\top}(\mu_{S_j^{\leq}}-\mu) \geq 46Cs/\sqrt{\alpha} - 45Cs/\sqrt{\alpha} \geq Cs/\sqrt{\alpha} > 100C\sigma_{S_j}/\sqrt{\alpha}\,,$$

where the first inequality used the assumption that $\|\mu_{S_j}-\mu\|_2 \geq 46Cs/\sqrt{\alpha}$ for $\sigma_{S_j} < s/100$ (and that v_j is the unit vector in the direction of $\mu_{S_j}-\mu$), and the last inequality used that we consider only clusters with $\sigma_{S_j} < s/100$.

The above implies that $\|\mu_{S_j}-\mu_{S_j^{\leq}}\|_2 > 100C\sigma_{S_j}/\sqrt{\alpha}$. If, for the sake of contradiction, we had $\alpha_j^{\leq} \geq 0.001\alpha$, then Fact 2.3 (and the fact that $C > 1$) implies $\|\mu_{S_j^{\leq}}-\mu_{S_j}\|_2 \leq 100\sigma_{S_j}/\sqrt{\alpha} \leq 100C\sigma_{S_j}/\sqrt{\alpha}$, which is a contradiction. Thus, it must be the case that $\alpha_j^{\leq} < 0.001\alpha$.

The above implies that $\sum_{j\,:\,\sigma_{S_j}<s/100}\sum_{x\in S_j^{\leq}} w_x \leq 0.001\alpha\sum_{j\,:\,\sigma_{S_j}<s/100}|S_j^{\leq}| \leq 0.001\alpha n \leq (0.001/0.97)\sum_{x\in T} w_x < 0.002\sum_{x\in T} w_x$, where the last inequality used that $\sum_{x\in T} w_x \geq 0.97\alpha n$ is a constraint in the program (4.1). ∎

For clusters S_j with $\sigma_{S_j} \geq s/100$: For every cluster S_j, we define a similar notation as in the previous case $\alpha_j := \left(\sum_{x\in S_j} w_x\right)/|S_j|$, which quantifies the overlap of the cluster with the solution of the program. As explained in the beginning, the goal is to show that all clusters S_j with mean far away from μ have (in the aggregate) small overlap with the solution $\{w_x\}_x$ of the program. In the previous paragraph (the one analyzing clusters with $\sigma_{S_j} \geq s/100$), we did not have to use that the means are far from μ because we could argue separately for the points that are close to μ; but here considering only clusters with mean far away from μ will become crucial. We will furthermore only consider clusters for which $\alpha_j > 0.001\alpha$—since our goal is to show small overlap in the aggregate, it suffices to do so for the clusters that individually have non-trivial overlap. In summary, the clusters that we consider in this paragraph are ones from the set $\mathrm{Bad} := \{j \in [k] : \sigma_{S_j} \geq s/100, \alpha_j > 0.001\alpha, \|\mu_{S_j}-\mu\|_2 > 4600C\sigma_{S_j}/\sqrt{\alpha}\}$, and the goal is to show that $\sum_{j\in\mathrm{Bad}}\sum_{x\in S_j} w_x \leq 0.001\sum_{x\in T} w_x$. To do this, we will show that the part of the solution coming from clusters in the set Bad causes large variance in the subspace connecting the μ_{S_j}'s with μ; thus, by the Ky-Fan norm constraint, such contributions should be limited.

Recall that for any cluster S_j, the notation v_j denotes the unit vector in the direction of $\mu_{S_j}-\mu$, and V denotes a subspace of dimension $1/\alpha$ that includes the span of $v_1,\ldots,v_k$ (recall $k \leq 1/\alpha$). Using calculations similar to Claim 1, we have that

$$\sum_{j=1}^{k}\sum_{x\in S_j} w_x \left(v_j(x-\mu)\right)^2 \leq \sum_{j=1}^{k}\sum_{x\in S_j} w_x \left\|\mathrm{Proj}_V(x-\mu)\right\|_2^2$$

$$
\begin{aligned}
&= \sum_{x\in T} w_x \left\|\mathrm{Proj}_V(x-\mu)\right\|_2^2 \\
&\le \left\|\sum_{x\in T} w_x (x-\mu)(x-\mu)^\top\right\|_{(1/\alpha)} \\
&\le \frac{2C^2 s^2}{\alpha}\sum_{x\in T} w_x \ , \qquad (4.2)
\end{aligned}
$$

where the last inequality is, again, by definition of the the program (4.1).

Now consider a cluster S_j with $j \in$ Bad, i.e. a cluster for which $\sigma_{S_j} > s/100$, $\alpha_j > 0.001\alpha$ and $\|\mu_{S_j} - \mu\|_2 > 4600C\sigma_{S_j}/\sqrt{\alpha}$. Let $\mu'_j := \left(\sum_{x\in S_j} w_x x\right) / \left(\sum_{x\in S_j} w_x\right)$. We have the following by Fact 2.3:

$$
\|\mu'_j - \mu_{S_j}\|_2 \le \sigma_{S_j}/\sqrt{\alpha_j} \le 100\sigma_{S_j}/\sqrt{\alpha} \le 100C\sigma_{S_j}/\sqrt{\alpha} \ . \qquad (4.3)
$$

The above implies that $v_j^\top(\mu'_j - \mu) \ge 4500C\sigma_{S_j}/\sqrt{\alpha}$, because otherwise we would have

$$
\begin{aligned}
\|\mu_{S_j} - \mu\|_2 &= v_j^\top(\mu_{S_j} - \mu) \\
&= v_j^\top(\mu_{S_j} - \mu'_j) + v_j^\top(\mu'_j - \mu) \\
&\le \|\mu_{S_j} - \mu'_j\|_2 + 4500C\sigma_{S_j}/\sqrt{\alpha} \\
&\le 100C\sigma_{S_j}/\sqrt{\alpha} + 4500C\sigma_{S_j}/\sqrt{\alpha} \qquad \text{(by (4.3))} \\
&\le 4600C\sigma_{S_j}/\sqrt{\alpha} \ ,
\end{aligned}
$$

which is a contradiction to $j \in$ Bad. Thus,

$$
\begin{aligned}
\sum_{x\in S_j} w_x (v_j^\top(x-\mu))^2 &= |S_j| \frac{\alpha_j}{\sum_{x\in S_j} w_x} \sum_{x\in S_j} w_x (v_j^\top(x-\mu))^2 \\
&\ge |S_j|\alpha_j \left(\frac{1}{\sum_{x\in S_j} w_x}\sum_{x\in S_j} w_x v_j^\top(x-\mu)\right)^2 \qquad \text{(by Jensen's inequality)} \\
&= |S_j|\alpha_j \left(v_j^\top(\mu'_j - \mu)\right)^2 \\
&\ge |S_j|\alpha_j \cdot 2\cdot 10^7 \cdot C^2\alpha^{-1}\sigma_{S_j}^2 \\
&\ge |S_j|\alpha_j \cdot 2000C^2\cdot\alpha^{-1}s^2 \qquad \text{(since } \sigma_{S_j} \ge s/100\text{)} \\
&\ge \left(\sum_{x\in S_j} w_x\right) 2000C^2\alpha^{-1}s^2 \ .
\end{aligned}
$$

Combining with (4.2) the above shows that $\sum_{j\in\text{Bad}}\sum_{x\in S_j} w_x \le 0.001\sum_{x\in T} w_x$.

Putting everything together: We now show how the two previous analyses for the clusters j with $\sigma_{S_j} < s/100$ and $\sigma_{S_j} \ge s/100$ for $j \in$ Bad can be combined to conclude the proof of Lemma 4.1.

We first argue that there exists exactly one cluster i with $\|\mu_{S_j} - \mu\|_2 \le 4600C\sigma_{S_i}/\sqrt{\alpha}$: Indeed, there cannot be more than one such clusters because if there were two clusters $i \ne j$ then by the triangle inequality and stability condition we would have

$$
\begin{aligned}
&\text{(by the triangle inequality)} && \|\mu_i - \mu_j\|_2 \le \|\mu_{S_i} - \mu\|_2 + \|\mu_{S_j} - \mu\|_2 + \|\mu_i - \mu_{S_i}\|_2 + \|\mu_j - \mu_{S_j}\|_2 \\
&\text{(by stability condition for means)} && \le \|\mu_{S_i} - \mu\|_2 + \|\mu_{S_j} - \mu\|_2 + C(\sigma_i + \sigma_j) \\
& && \le 4600C(\sigma_{S_i} + \sigma_{S_j})/\sqrt{\alpha} + C(\sigma_i + \sigma_j) \\
&\text{(by stability condition for covariances)} && \le 4600C^2(\sigma_i + \sigma_j)/\sqrt{\alpha} + C(\sigma_i + \sigma_j) \\
&\text{(using } C > 1\text{)} && \le 4601C^2(\sigma_i + \sigma_j)/\sqrt{\alpha} \ ,
\end{aligned}
$$

which would violate our separation assumption in Theorem 3.2. It also cannot be the case that none of the clusters satisfy the condition that $\|\mu_{S_i} - \mu\|_2 \le 4600C\sigma_{S_i}/\sqrt{\alpha}$, because in that case we will show that we could also obtain a contradiction. Recall that in our notation T is the entire dataset and S_j's are the stable sets (which we often call "clusters"). The contradiction can be derived as follows (step by step explanations are provided in the next paragraph):

$$\sum_{x\in T} w_x = \sum_{j:\sigma_{S_j}<s/100} \sum_{x\in S_j} w_x + \sum_{j:\sigma_{S_j}\ge s/100} \sum_{x\in S_j} w_x + \sum_{x\in T\setminus\cup_j S_j} w_x \tag{4.4}$$

$$= \sum_{j:\sigma_{S_j}<s/100} \sum_{x\in S_j} w_x + \sum_{j\in \text{Bad}} \sum_{x\in S_j} w_x + \sum_{j:\sigma_{S_j}\ge s/100, j\notin \text{Bad}} \sum_{x\in S_j} w_x + \sum_{x\in T\setminus\cup_j S_j} w_x \tag{4.5}$$

$$\le 0.003 \sum_{x\in T} w_x + 0.002 \sum_{x\in T} w_x + 0.001 \sum_{x\in T} w_x + 0.02\alpha n \tag{4.6}$$

$$\le 0.003 \sum_{x\in T} w_x + 0.002 \sum_{x\in T} w_x + 0.001 \sum_{x\in T} w_x + 0.021 \sum_{x\in T} w_x \le 0.05 \sum_{x\in T} w_x \ . \tag{4.7}$$

We explain the steps here: (4.4) splits the summation into a part for the large covariance clusters and one for the small covariance ones, and the part of the dataset that does not belong to any of the clusters. (4.5) further splits the sum due to large variance clusters into two parts: the clusters that belong in the set Bad and the rest of them. (4.6) bounds each one of the resulting terms as follows: The bound of the first term uses the analysis of small covariance clusters. The bound of the second term uses the analysis of large covariance clusters. The bound of the third term uses that, since we have assumed that $\|\mu_{S_j} - \mu\|_2 > 4600C\sigma_{S_j}/\sqrt{\alpha}$ for all clusters, the only way that $j \notin$ Bad can happen is because of $\alpha_j < 0.001\alpha$. The bound of the last term comes from the assumption in Theorem 3.2 that $\cup_i S_i$ contains most of the points in T (this is one of the assumptions in Theorem 3.2). Finally, (4.7) uses the fact that $\sum_{x\in T} w_x \ge 0.97\alpha n$ by construction of the the program constraints in (4.1).

Equation (4.7) yields the desired contradiction, thus there must be exactly one cluster S_i with $\|\mu_{S_i} - \mu\|_2 \le 4600C\sigma_{S_i}/\sqrt{\alpha}$. This completes the proof of Lemma 4.1.

□

Having shown Lemma 4.1 which gives guarantees about solutions of the convex program (4.1), we can now state and prove the induction (Lemma 4.2) which guarantees that throughout the execution of the double loop in Line 4, every candidate mean added to the list L must be close to some true cluster S_i, and every true cluster S_i with standard deviation at most s must have a corresponding candidate mean in L.

LEMMA 4.2. (INDUCTION) *Consider the setting of Theorem 3.2 and Algorithm 1. The first statement below holds throughout the execution and the second statement holds at the start of every iteration of the loop of line 4:*

1. *(Every element from the list is being mapped to a true cluster): For every element $\hat{\mu}_i$ in the list L there exists a true cluster S_j such that $\|\hat{\mu}_i - \mu_{S_j}\| \le 4600C\sigma_{S_j}/\sqrt{\alpha}$.*
2. *(Every cluster of smaller covariance has already been found): For every true cluster S_i with $\sigma_{S_i} \le s$, there exists $\hat{\mu}_j$ in the list L such that $\|\hat{\mu}_j - \mu_{S_i}\|_2 \le 4600C\sigma_{S_i}/\sqrt{\alpha}$.*

Before we prove the lemma, we note that the guarantee of the lemma involves the empirical quantities μ_{S_i} and σ_{S_i} as opposed to the "true" means and standard deviations μ_i, σ_i of the mixture components, which are the parameters that each S_i is stable with respect to. Later on in the paper, we will use the following straightforward corollary of Lemma 4.2, which can be derived directly by the two stability conditions $\sigma_{S_i} \le C\sigma_i$ and $\|\mu_{S_i} - \mu_i\|_2 \le C\sigma_i$.

COROLLARY 4.1. *In the setting of Lemma 4.2, the first statement holds throughout the execution of the algorithm and the second holds at the start of every iteration of the loop of line 4:*

1. *For every element $\hat{\mu}_i$ in the list L, there exists a true cluster S_j such that $\|\hat{\mu}_i - \mu_j\| \le 4601C^2\sigma_j/\sqrt{\alpha}$.*
2. *For every true cluster S_i with $\sigma_{S_i} \le s$, there exists $\hat{\mu}_j$ in the list L such that $\|\hat{\mu}_j - \mu_i\|_2 \le 4601C^2\sigma_i/\sqrt{\alpha}$.*

We now prove Lemma 4.2.

Proof. [Proof of Lemma 4.2]

In everything that follows, we will informally use the phrase that "cluster S_i has been found" as a shorthand to the statement that there exists $\hat{\mu}_j$ in the list L such that $\|\hat{\mu}_j - \mu_{S_i}\|_2 \leq 4600C\sigma_i/\sqrt{\alpha}$.

We prove the lemma by induction. Suppose the algorithm enters a new iteration of the outer loop (line 4), and suppose that Items 1 and 2 (our inductive hypothesis) hold for all prior steps of the algorithm. We will show that Item 1 remains true each time a new element is inserted into the list L in iterations of the inner loop and that Item 2 remains true in the next iteration of the outer loop. Since showing Item 2 is more involved, we will start with that.

Proof of Item 2: For Item 2 we want to show that every cluster S_j with $\sigma_{S_j} \leq s$ will be found. We consider two cases: The first case is $\sigma_{S_j} < s/100$. In that case, by the guarantee of list-decoding for the covariances (Proposition 2.1), there must exist a candidate standard deviation $\hat{s}$ in the list L_{stdev} such that $\sigma_{S_j} \leq \hat{s} \leq \sqrt{2}\sigma_{S_j}$. Note that combining with $\sigma_{S_j} < s/100$ this implies that $\hat{s} < s$. This means that, as the algorithm has gone through the list L_{stdev}, it must have examined that candidate covariance $\hat{s}$ in an earlier step. For that step, the inductive hypothesis along with the fact that $\sigma_{S_j} \leq \hat{s}$ implies that the cluster S_j must have already been found at that earlier step.

Now let us consider the case $s/100 \leq \sigma_{S_j} \leq s$. We will show that, if the cluster has not been already found, then it will be found at the current iteration of the loop of line 4. We will do this by showing that there exists a candidate mean $\mu \in L_{\mathrm{mean}}$ such that:

(a) $\|\mu - \mu_{S_j}\|_2 \leq C\sigma_{S_j}/\sqrt{\alpha}$.

(b) $\|\mu - \hat{\mu}_i\|_2 > 99Cs/\sqrt{\alpha}$ for every $\hat{\mu}_i$ in the list L.

(c) The program defined by (4.1) is satisfiable.

Before establishing the individual claims, we point out that they indeed imply that the cluster j will be found at the current iteration. To see this, first note that claim (b) above implies that the algorithmic check in line 4(a)i will go through when the algorithm uses the candidate mean μ. Then, because of claim (c), the program will be satisfiable, and an application of Lemma 4.1 combined with claim (a) will yield that $\|\mu - \mu_{S_j}\|_2 \leq 4600C\sigma_{S_j}/\sqrt{\alpha}$, i.e. the cluster S_j is indeed found. We explain the application of Lemma 4.1 in detail in the next two paragraphs.

First, we check that the preconditions of Lemma 4.1 are established, i.e. we will check that for every cluster ℓ with $\sigma_{S_\ell} < s/100$ it holds that $\|\mu - \mu_{S_\ell}\|_2 \geq 46Cs/\sqrt{\alpha}$ and that a solution to the program exists. The satisfiability of the program is due to claim (c). In the reminder of the paragraph, we show the part that $\|\mu - \mu_{S_\ell}\|_2 \geq 46Cs/\sqrt{\alpha}$ for all clusters ℓ with $\sigma_{S_\ell} < s/100$: By the inductive hypothesis, all clusters with standard deviation at most $s/100$ have already been found, meaning that if S_ℓ is a cluster with $\sigma_{S_\ell} < s/100$, then there is a $\hat{\mu}_t$ in the list with $\|\hat{\mu}_t - \mu_{S_\ell}\|_2 \leq 4600C\sigma_{S_\ell}/\sqrt{\alpha}$. Putting everything together, if S_ℓ is a cluster with $\sigma_{S_\ell} < s/100$, then $\|\mu - \mu_{S_\ell}\|_2 \geq \|\mu - \hat{\mu}_t\|_2 - \|\hat{\mu}_t - \mu_{S_\ell}\|_2 \geq 99Cs/\sqrt{\alpha} - 4600C\sigma_{S_\ell}/\sqrt{\alpha} \geq 99Cs/\sqrt{\alpha} - 46Cs/\sqrt{\alpha} \geq 46Cs/\sqrt{\alpha}$ (where the first step uses the reverse triangle inequality, the second step uses claim (b) for the first term and $\|\hat{\mu}_t - \mu_\ell\|_2 \leq 4600C\sigma_{S_\ell}/\sqrt{\alpha}$ for the second term and the next step uses that $\sigma_{S_\ell} < s/100$).

We have thus checked that Lemma 4.1 is applicable. We now check that the conclusion of the lemma indeed implies that cluster S_j will be found. The conclusion of the lemma (after a renaming of the index) is that there exists a unique true cluster S_t with $\sigma_{S_t} \geq s/100$ such that $\|\mu - \mu_{S_t}\|_2 \leq 4600C\sigma_{S_t}/\sqrt{\alpha}$. Note the "unique" part: there cannot be any other cluster $S_{t'}$ for which $\|\mu - \mu_{S_{t'}}\|_2 \leq 4600C\sigma_{S_{t'}}/\sqrt{\alpha}$ (otherwise the separation assumption between clusters is violated). This combined with claim (a) means that the cluster S_t from the conclusion of Lemma 4.1 must be the same cluster that we originally denoted by S_j. Thus, we showed that cluster S_j is found, as desired.

We now show that the claims (a), (c), and (b) hold for μ being the mean candidate for which it holds $\|\mu - \mu_{S_j}\|_2 \leq C\sigma_{S_j}/\sqrt{\alpha}$ by the list-decoding guarantee (Fact 2.4). Thus, (a) is satisfied by that fact. We now show that this μ also satisfies (c): Using (a) and that the standard deviation of S_j in every direction is at most σ_{S_j} (by definition), we can show the following for the Ky-Fan norm of the centered around μ second moment of

that true cluster:

$$\begin{aligned}
\left\| \frac{1}{|S_j|} \sum_{x \in S_j} (x-\mu)(x-\mu)^\top \right\|_{(1/\alpha)} &\le \left\| \frac{1}{|S_j|} \sum_{x \in S_j} (x-\mu_{S_j})(x-\mu_{S_j})^\top \right\|_{(1/\alpha)} + \|\mu - \mu_{S_j}\|_2^2 \\
&\le \frac{1}{\alpha} \left\| \frac{1}{|S_j|} \sum_{x \in S_j} (x-\mu_{S_j})(x-\mu_{S_j})^\top \right\|_{(\mathrm{op})} + C^2 \frac{1}{\alpha} \sigma_{S_j}^2 \\
&\le \frac{1}{\alpha} \left(\sigma_{S_j}^2 + C^2 \sigma_{S_j}^2 \right) \\
&\le 2C^2 \frac{s^2}{\alpha} \,,
\end{aligned}$$

where the first step uses the inverse triangle inequality and the last step uses that we only consider true clusters with $\sigma_{S_j} \le s$. Thus, the program is satisfiable by the binary weights $w_x = \mathbb{1}(x \in S_j)$.

We now move to establishing the claim (b), i.e. that $\|\mu - \hat{\mu}_i\|_2 > 99Cs/\sqrt{\alpha}$ for every $\hat{\mu}_i$ in the list L. Consider an arbitrary $\hat{\mu}_i$ from the list L corresponding to a previously found cluster. By the inductive hypothesis, for every $\hat{\mu}_i \in L$, there exists a true cluster S_ℓ for which $\|\hat{\mu}_i - \mu_{S_\ell}\|_2 \le 4600C\sigma_{S_\ell}/\sqrt{\alpha}$. By assumption in the context of the claim, cluster j has not been found, and thus $\ell \neq j$. Then, by the reverse triangle inequality, we obtain:

$$\begin{aligned}
\|\mu - \hat{\mu}_i\|_2 &\ge \|\mu_j - \mu_\ell\|_2 - \|\mu_j - \mu_{S_j}\|_2 - \|\mu_\ell - \mu_{S_\ell}\|_2 - \|\mu_{S_\ell} - \hat{\mu}_i\|_2 - \|\mu - \mu_{S_j}\|_2 \\
&> 10^4 C^2(\sigma_\ell + \sigma_j)/\sqrt{\alpha} - C\sigma_j - C\sigma_\ell - 4600C\sigma_{S_\ell}/\sqrt{\alpha} - C\sigma_{S_j}/\sqrt{\alpha} \\
&\ge (10^4 - 1)C^2(\sigma_j + \sigma_\ell)/\sqrt{\alpha} - 4600C\sigma_{S_\ell}/\sqrt{\alpha} - C\sigma_{S_j}/\sqrt{\alpha} \\
&\quad (\sigma_{S_j} \le C\sigma_j \text{ by stability condition for covariances}) \\
&\ge (10^4 - 1)C(\sigma_{S_j} + \sigma_{S_\ell})/\sqrt{\alpha} - 4600C\sigma_{S_\ell}/\sqrt{\alpha} - C\sigma_{S_j}/\sqrt{\alpha} \\
&\ge (10^4 - 2)C\sigma_{S_j} \\
&\quad (\text{using } s/100 < \sigma_{S_j}) \qquad \ge 99Cs/\sqrt{\alpha} \,,
\end{aligned}$$

where the second line uses the separation assumption between clusters ℓ, j to bound below the first term, the stability condition to bound the next two terms, and the facts that $\|\mu - \mu_j\|_2 \le C\sigma_{S_j}/\sqrt{\alpha}$ and $\|\hat{\mu}_i - \mu_{S_\ell}\|_2 \le 4600C\sigma_\ell/\sqrt{\alpha}$ that we had already established in the previous paragraph. The last line uses that we are analyzing only the case $s/100 < \sigma_{S_j}$.

Proof of Item 1: Consider an iteration of the (inner) loop of the algorithm. We assume that the inductive hypothesis holds for the past iterations and we will show that Item 1 continues to be true after the current one is finished. It suffices to only consider an iteration where a new element $\hat{\mu}$ gets inserted to the list L in line 4(a)ii (otherwise the claim is trivial). The fact that $\hat{\mu}$ corresponds to a true cluster will be a direct consequence of Lemma 4.1.

It remains to check that Lemma 4.1 is applicable, i.e. we will check that for every cluster ℓ with $\sigma_{S_\ell} < s/100$ it holds that $\|\mu - \mu_{S_\ell}\|_2 \ge 46Cs/\sqrt{\alpha}$ and that a solution to the program exists. The satisfiablitity of the program is due to the fact that the algorithm has reached line 4(a)ii. In the reminder of the paragraph, we show the part that $\|\mu - \mu_{S_\ell}\|_2 \ge 46Cs/\sqrt{\alpha}$ for all clusters ℓ with $\sigma_{S_\ell} < s/100$: By the inductive hypothesis, all clusters with standard deviation at most $s/100$ have already been found, meaning that if S_ℓ is a cluster with $\sigma_{S_\ell} < s/100$, then there is a $\hat{\mu}_t$ in the list with $\|\hat{\mu}_t - \mu_{S_\ell}\|_2 \le 4600C\sigma_{S_\ell}/\sqrt{\alpha}$. Putting everything together, if S_ℓ is a cluster with $\sigma_{S_\ell} < s/100$, then $\|\mu - \mu_{S_\ell}\|_2 \ge \|\mu - \hat{\mu}_t\|_2 - \|\hat{\mu}_t - \mu_{S_\ell}\|_2 \ge 99Cs/\sqrt{\alpha} - 4600C\sigma_{S_\ell}/\sqrt{\alpha} \ge 99Cs/\sqrt{\alpha} - 46Cs/\sqrt{\alpha} \ge 46Cs/\sqrt{\alpha}$, where the inequalities used are the following: The first step uses the reverse triangle inequality, the second step uses the condition in line 4(a)i of the pseudocode line 4(a)ii in order to bound the first term and $\|\hat{\mu}_t - \mu_{S_\ell}\|_2 \le 4600C\sigma_{S_\ell}/\sqrt{\alpha}$ for the second term, and the next inequality uses that $\sigma_{S_\ell} < s/100$.

□

5 Cardinality-based pruning of candidate means

This section concerns Line 5 of Algorithm 1. Right before Line 5 is executed, we are guaranteed that the list L of candidate means consists only of candidates close to one of the S_i sets. Concretely, every $\hat{\mu} \in L$ is close

to some S_i with distance at most $O(\sigma_{S_i}/\sqrt{\alpha})$, and that every S_i has some $\hat{\mu} \in L$ close to it. At this point, the Voronoi partition of the samples is already an accurate refinement of the ground truth clustering (Lemma 5.1 below). However, we want to further ensure that the returned clustering "looks like" what we assume of our underlying mixture distribution; namely, that each subset has at least $\approx \alpha$ mass, and that the subsets are pairwise well-separated. Line 5 prunes candidate means, via Algorithm 2 stated below, to ensure that the corresponding Voronoi cell has sufficient mass.

We first show Lemma 5.1, which states that the Voronoi partition based on the candidate means in L does form an accurate refinement to the ground truth clustering.

LEMMA 5.1. (VORONOI CLUSTERING PROPERTIES) *Consider the notation and assumptions of Theorem 3.2. Let L be an m-sized list of vectors $\hat{\mu}_1, \dots, \hat{\mu}_m$ with $m \geq k$. Suppose the list L can be partitioned into sets $H_1, \dots, H_k$ such that for every $i \in [k]$, H_i consists of the vectors $\hat{\mu}_j$ with $\|\hat{\mu}_j - \mu_i\|_2 \leq 4601C^2\sigma_i/\sqrt{\alpha}$, and further assume that $H_i \neq \emptyset$ for all $i \in [k]$. Let $A_j = \{x \in T : \arg\min_{j' \in [m]} \|x - \hat{\mu}_{j'}\|_2 = j\}$ for $j \in [m]$ be the Voronoi partition (recall that T denotes the entire dataset). For each $i \in [k]$ define $\mathcal{A}_i := \cup_{j:\hat{\mu}_j \in H_i} A_j$. Then, the following hold:*

1. *(Points from S_i assigned to sub-clusters associated with the wrong true cluster are few) $|S_i \setminus \mathcal{A}_i| \leq 0.011|S_i|$ for every $i \in [k]$, and*
2. *(Points from the sub-clusters associated with a true cluster mostly include points from that true cluster) $|\mathcal{A}_i \setminus S_i| \leq 0.03\alpha n$ for every $i \in [k]$.*
3. *$|\mathcal{A}_i| \geq 0.959\alpha n$ for $i \in [k]$.*

Proof. First, observe that Item 3 in the lemma follows directly from Item 1 and the assumption $|S_i| \geq 0.97\alpha n$. Namely,

$$|\mathcal{A}_i| \geq |\mathcal{A}_i \cap S_i| \geq |S_i| - |S_i \setminus \mathcal{A}_i| \geq 0.989|S_i| \geq 0.959\alpha n \,. \tag{5.8}$$

Thus it suffices to prove Items 1 and 2.

For $i \in [k]$ and for every $i' \neq i$ define the intersection of the true cluster i with the union of the sub-clusters associated with cluster i' as $S'_{i,i'} := S_i \cap \mathcal{A}_{i'}$. We claim that it suffices to show that $|S'_{i,i'}| < (0.01\alpha)|S_i|$ for every $i' \neq i$, that Items 1 and 2 follow.

For the first part of the lemma statement (Item 1), we have that

$$|S_i \setminus \mathcal{A}_i| = \sum_{i' \neq i} |S_i \cap \mathcal{A}_{i'}| = \sum_{i' \neq i} |S'_{i,i'}| \leq 0.01|S_i|\alpha k \leq 0.011|S_i| \,,$$

where we used that the sets $A_1, \dots, A_m$ form a partition of T, and the number of true clusters is $k \leq 1/(0.97\alpha)$ (since we assumed $|S_i| \geq 0.97\alpha n$).

Similarly, for the second part of the lemma statement (Item 2),

$$|\mathcal{A}_i \setminus S_i| \leq \sum_{i' \neq i} |\mathcal{A}_i \cap S_{i'}| + 0.02\alpha n \leq 0.01\alpha \sum_{i' \in [k]} |S_{i'}| + 0.02\alpha n \leq 0.01\alpha n + 0.02\alpha n \leq 0.03\alpha n \,,$$

where the first inequality uses the assumption from Theorem 3.2, that there are at most $0.02\alpha n$ points that do not belong to any of the sets $S_1, \dots, S_n$.

We now show the claim that $|S'_{i,i'}| < (0.01\alpha)|S_i|$ for every $i, i' \in [k]$ with $i' \neq i$. Recall our notation μ_i (for $i \in [k]$) representing the vectors that each true cluster S_i is stable for (see setup of Theorem 3.2). These vectors should not be confused with the $\hat{\mu}_j$ ones (for $j \in [m]$), which are the approximate centers used to produce the Voronoi partition. Since we have assumed that the μ_i's are separated from each other and H_i contains (by definition) the candidate means that are close to μ_i, every pair of vectors $\hat{\mu} \in H_i$ and $\hat{\mu}' \in H_{i'}$ for $i \neq i'$ must also be separated:

$$\begin{aligned}
\text{(by reverse triangle inequality)} \qquad \|\hat{\mu} - \hat{\mu}'\|_2 &\geq \|\mu_i - \mu_{i'}\|_2 - \|\hat{\mu} - \mu_i\|_2 - \|\hat{\mu}' - \mu_{i'}\|_2 \\
&\geq 10^5 C^2(\sigma_i + \sigma_{i'})/\sqrt{\alpha} - 4601C^2\sigma_i/\sqrt{\alpha} - 4601C^2\sigma_{i'}/\sqrt{\alpha} \\
&\geq 95399C^2(\sigma_i + \sigma_{i'})/\sqrt{\alpha} \,. \qquad (5.9)
\end{aligned}$$

Given that every point in $S'_{i,i'}$ is closer to some $\hat{\mu}' \in H_{i'}$ than every $\hat{\mu} \in H_i$, and furthermore given that $\hat{\mu}$ and $\hat{\mu}'$ are far from each other according to (5.9), we now show that $\|\mu_{S'_{i,i'}} - \mu_i\|_2 > 10C^2\sigma_i/\sqrt{\alpha}$. Combining this with Fact 2.3, we can extract that $|S'_{i,i'}| < (0.01\alpha)|S_i|$. To see that by contradiction, assume that $|S'_{i,i'}| \geq (0.01\alpha)|S_i|$. Then, Fact 2.3 ensures that $\|\mu_{S'_{i,i'}} - \mu_i\|_2 \leq 10\sigma_{S_i}/\sqrt{\alpha} \leq 10C\sigma_{S_i}/\sqrt{\alpha} \leq 10C^2\sigma_i/\sqrt{\alpha}$, where we used $C > 1$ as well as the stability condition for the covariance (the fact that $\sigma_{S_i} \leq C\sigma_i$).

To see that $\|\mu_{S'_{i,i'}} - \mu_i\|_2 > 10C^2\sigma_i/\sqrt{\alpha}$, consider an arbitrary point

$x \in S'_{i,i'}$ and let $\hat{\mu}' \in H_{i'}$ be the center from L that is the closest one to x (by definition of $S'_{i,i'}$ that closest center belongs in $H_{i'}$). Letting $\hat{\mu}$ again be an arbitrary center from H_i, since x is closer to $\hat{\mu}'$ than $\hat{\mu}$, we have $\|x - \hat{\mu}\|_2 \geq \frac{1}{2}\|\hat{\mu} - \hat{\mu}'\|_2$. Finally,

$$
\begin{aligned}
\text{(by reverse triangle inequality)} \qquad \|x - \mu_i\|_2 &\geq \|x - \hat{\mu}\|_2 - \|\hat{\mu} - \mu_i\|_2 \\
&\geq \frac{1}{2}\|\hat{\mu} - \hat{\mu}'\|_2 - \|\hat{\mu} - \mu_i\|_2 \\
\text{(by (5.9) and } \hat{\mu} \in H_i\text{)} \qquad &\geq \frac{1}{2} \cdot 95399C^2(\sigma_i + \sigma_{i'})/\sqrt{\alpha} - 4601C^2\sigma_i/\sqrt{\alpha} \\
&> 10C^2\sigma_i/\sqrt{\alpha}\,.
\end{aligned}
$$

Since the above holds for every $x \in S'_{i,i'}$, it also holds for the mean of that set, i.e. $\|\mu_{S'_{i,i'}} - \mu_i\|_2 > 10C^2\sigma_i/\sqrt{\alpha}$. As we mentioned above, combining this with Fact 2.3 shows that $|S'_{i,i'}| < (0.01\alpha)|S_i|$, as desired.

□

We now state Algorithm 2, which is used in Line 5 of Algorithm 1.

Algorithm 2 Pruning of sub-clusters based on cardinality.

Input: Dataset T of n points, centers $\hat{\mu}_1, \ldots, \hat{\mu}_m$ and parameter $\alpha \in (0, 1)$.
Output: A subset $\hat{\mu}_1, \ldots, \hat{\mu}_{m'}$ of the input centers.

1. $J_{\text{deleted}} \leftarrow \emptyset$.
2. Construct the Voronoi partition $A_j = \{x : \arg\min_{j' \in [m]} \|x - \hat{\mu}_{j'}\|_2 = j\}$ for $j \in [m]$.
3. While there exists $j \in [m] \setminus J_{\text{deleted}}$ with $|A_j| < 0.96\alpha n$ do:
   (a) Update $J_{\text{deleted}} \leftarrow J_{\text{deleted}} \cup \{j\}$.
   (b) For all $j \notin J_{\text{deleted}}$, update $A_j = \{x : \arg\min_{j' \in [m] \setminus J_{\text{deleted}}} \|x - \hat{\mu}_{j'}\|_2 = j\}$.
4. Return $\{\hat{\mu}_j\}_{j \in [m] \setminus J_{\text{deleted}}}$.

Lemma 5.2 below analyzes Algorithm 2.

LEMMA 5.2. (PRUNING OF SUB-CLUSTERS BASED ON CARDINALITY) *Consider the notation and assumptions of Theorem 3.2. Let L be an m-sized list of vectors $\hat{\mu}_1, \ldots, \hat{\mu}_m$ with $m \geq k$. Suppose the list L can be partitioned into sets $H_1, \ldots, H_k$ such that for every $i \in [k]$, H_i consists of the vectors $\hat{\mu}_j$ with $\|\hat{\mu}_j - \mu_i\|_2 \leq 4061C^2\sigma_i/\sqrt{\alpha}$, and further assume that $H_i \neq \emptyset$ for all $i \in [k]$.*

Suppose that we run Algorithm 2 on L as the input and denote by $\hat{\mu}_1, \ldots, \hat{\mu}_{m'}$ the sublist of centers output by the algorithm. Then, if we define the sets $H'_i := \{\hat{\mu}_j \text{ for } j \in [m'] : \|\hat{\mu}_j - \mu_i\|_2 \leq 4061C^2\sigma_i/\sqrt{\alpha}\}$ for $i \in [k]$, then $H'_1, \ldots, H'_k$ is a partition of $\{\hat{\mu}_1, \ldots, \hat{\mu}_{m'}\}$ and it also holds that $H'_i \neq \emptyset$ for all $i \in [k]$. Moreover, in the final Voronoi clustering that corresponds to these output centers, $A_j := \{x : \arg\min_{j' \in [m']} \|x - \hat{\mu}_{j'}\|_2 = j\}$ for $j \in [m']$, it holds true that $|A_j| \geq 0.96\alpha n$.

Proof. Consider the notation A_j for the Voronoi clusters as in the pseudocode of Algorithm 2. The claim that $|A_j| \geq 0.96\alpha n$ for all $j \in [m']$ follows by construction of the algorithm (line 3). We thus focus on the remaining part of the lemma conclusion (the one about the sets H'_i).

To show the remaining parts of the lemma conclusion, it suffices to show that at any point during the algorithm's execution, if we define the sets $H'_i := \{\hat{\mu}_j \text{ for } j \in [m] \setminus J_{\text{deleted}} : \|\hat{\mu}_j - \mu_i\|_2 \le 4061C^2\sigma_j/\sqrt{\alpha}\}$, then $H'_i \neq \emptyset$ for all $i \in [k]$ (the fact that $H'_1, \dots, H'_k$ is a partition of L holds trivially by our assumption on the input).

In order to show that $H'_i \neq \emptyset$ for all $i \in [k]$, suppose that at some point during the algorithm's execution there exists $i \in [k]$ for which we are left with only a single center $\hat{\mu}_j$ satisfying $\|\hat{\mu}_j - \mu_i\|_2 \le 4061C^2\sigma_i/\sqrt{\alpha}$. Then, we will show that this $\hat{\mu}_j$ will never get deleted. To do so, we claim that at least $0.99|S_i|$ points of S_i have $\hat{\mu}_j$ as their closest center among the non-deleted centers $\{\hat{\mu}_t\}_{t\in[m]\setminus J_{\text{deleted}}}$. From this claim, it follows that the set A_j in the Voronoi partition corresponding to that center will have size $|A_j| \ge 0.99|S_i| \ge 0.99 \cdot 0.97\alpha n \ge 0.96\alpha n$ (using our assumption $|S_i| > 0.97\alpha n$) and therefore $\hat{\mu}_j$ will never be deleted because of the deletion condition in line 3.

We now prove the above claim that at least $0.99|S_i|$ points of S_i have $\hat{\mu}_j$ as their closest non-deleted center. Denote by $S'_{i,i'} := \{x \in S_i : \arg\max_{t\in[m]\setminus J_{\text{deleted}}} \|x - \hat{\mu}_t\|_2 \in H_{i'}\}$, i.e. the part of S_i consisting of the points that are closer to centers belonging in $H_{i'}$ than H_i. First we argue that it suffices to show that $|S'_{i,i'}| < 0.01\alpha|S_i|$. This implies $\sum_{i'\neq i} |S'_{i,i'}| \le 0.01k\alpha|S_i| \le 0.01|S_i|$, which means that, at least $0.99|S_i|$ of the points from S_i must have $\arg\max_{j'\in[m]\setminus J_{\text{deleted}}} \|x - \hat{\mu}_{j'}\|_2 \in H_i$. Finally, since we are under the assumption that $\hat{\mu}_j$ is the only center in H_i from the non-deleted ones ($j \in [m] \setminus J_{\text{deleted}}$), the previous implies that at least $0.99|S_i|$ points of S_i have $\hat{\mu}_j$ as their closest center.

In order to show $|S'_{i,i'}| < 0.01\alpha|S_i|$ for any $i' \neq i$, we will show that $\|\mu_{S'_{i,i'}} - \mu_i\|_2 > 10C^2\sigma_i/\sqrt{\alpha}$; this is enough because of Fact 2.3 and the fact that $S_{i,i'} \subseteq S_i$.

It thus remains to show that $\|\mu_{S'_{i,i'}} - \mu_i\|_2 > 10C^2\sigma_i/\sqrt{\alpha}$. To do so, consider any center $\hat{\mu}_\ell$ that satisfies $\|\hat{\mu}_\ell - \mu_{i'}\|_2 \le 4061C^2\sigma_{i'}/\sqrt{\alpha}$ and observe the following (recall that in our notation $\hat{\mu}_j$ is the only center from $\{\hat{\mu}_t\}_{t\in[m]\setminus J_{\text{deleted}}}$ that satisfies $\|\hat{\mu}_j - \mu_i\|_2 \le 4061C^2\sigma_i/\sqrt{\alpha}$):

$$\begin{aligned}
\text{(by reverse triangle inequality)} \qquad \|\hat{\mu}_j - \hat{\mu}_\ell\|_2 &\ge \|\mu_i - \mu_{i'}\|_2 - \|\hat{\mu}_\ell - \mu_{i'}\|_2 - \|\hat{\mu}_j - \mu_i\|_2 \\
&\ge 10^5C^2(\sigma_i + \sigma_{i'})/\sqrt{\alpha} - 4061C^2\sigma_{i'}/\sqrt{\alpha} - 4061C^2\sigma_i/\sqrt{\alpha} \\
&\ge 95399C^2(\sigma_i + \sigma_{i'})/\sqrt{\alpha}\,. \qquad (5.10)
\end{aligned}$$

Now, consider $S'_{i,i'} := \{x \in S_i : \arg\max_{t\in[m]\setminus J_{\text{deleted}}} \|x - \hat{\mu}_t\|_2 \in H_{i'}\}$ and fix an $x \in S'_{i,i'}$. If ℓ denotes the $\arg\max_{t\in[m]\setminus J_{\text{deleted}}} \|x - \hat{\mu}_t\|_2$, then it holds $\|x - \hat{\mu}_j\|_2 \ge \frac{1}{2}\|\hat{\mu}_j - \hat{\mu}_\ell\|_2$. Then,

$$\begin{aligned}
\|x - \mu_i\|_2 &\ge \|x - \hat{\mu}_j\|_2 - \|\hat{\mu}_j - \mu_i\|_2 \\
&\ge \frac{1}{2}\|\hat{\mu}_j - \hat{\mu}_\ell\|_2 - \|\hat{\mu}_j - \mu_i\|_2 \\
\text{(by (5.10))} \qquad &\ge \frac{1}{2} \cdot 95399C^2(\sigma_i + \sigma_{i'})/\sqrt{\alpha} - 4061C^2\sigma_i/\sqrt{\alpha} \\
&> 10C^2\sigma_i/\sqrt{\alpha}\,.
\end{aligned}$$

Since, the above holds for every $x \in S'_{i,i'}$, then it must also hold for their mean of the set, i.e. $\|\mu_{S'_{i,i'}} - \mu_i\|_2 > 10C^2\sigma_i/\sqrt{\alpha}$.

□

6 Distance-based pruning of candidate means

In the previous section, we gave Algorithm 2 used in Line 5 of Algorithm 1, which ensures that the list L of candidate means corresponds to a Voronoi partition that is an accurate refinement of the true clustering $\{S_i\}_i$, and furthermore, that each subset in the partition has size at least $\approx \alpha n$.

This section concerns Line 6 of Algorithm 1, which additionally prunes the list L so that the Voronoi cells are in fact far apart from each other, satisfying a pairwise separation that is qualitatively identical to the separation assumption we impose on the underlying mixture distribution.

Due to the existence of adversarial corruptions and heavy-tailed noise in the data set, we first need to use filtering on each Voronoi cell (Algorithm 3), in order to make sure that the mean of the filtered Voronoi cell is actually close to the mean of the S_i that the cell corresponds to. Corollary 6.1 states the guarantees after such filtering.

Algorithm 3 Filtered Voronoi partitioning

Input: Dataset T of n points and centers $\hat{\mu}_1, \dots, \hat{\mu}_m$.
Output: Disjoint subsets $B_1, \dots, B_m$ of T.

1. Construct the Voronoi partition $A_j = \{x \in T : \arg\min_{j' \in [m]} \|x - \hat{\mu}_{j'}\|_2 = j\}$.
2. $B_j \leftarrow \textsc{Filter}(A_j)$ for $j \in [m]$, where $\textsc{Filter}$ denotes the filtering algorithm from Fact 2.2.
3. Output $B_1, \dots, B_m$.

Corollary 6.1. (Filtered Voronoi clustering properties) *Consider the setting of Lemma 5.1 and furthermore assume that the Voronoi sets have size $|A_j| \geq 0.96\alpha n$ for every $j \in [m]$. Then the algorithm $\textsc{FilteredVoronoi}(T, \{\hat{\mu}_i\}_{i \in [m]})$ outputs disjoint sets $B_1, \dots, B_m$ such that with probability $1 - \alpha\delta/10$, the following are true (denote $\mathcal{B}_i = \cup_{j : \hat{\mu}_j \in H_i} B_j$, where H_i's are defined as in Lemma 5.1):*

1. *$|S_i \setminus \mathcal{B}_i| \leq 0.033|S_i|$ for every $i \in [k]$.*
2. *$|\mathcal{B}_i \setminus S_i| \leq 0.03\alpha n$ for every $i \in [k]$ and $|A_j \setminus B_j| \leq 0.04|A_j|$ for every $j \in [m]$.*
3. *For any $j \in [m]$ such that $\hat{\mu}_j \in H_i$, it holds $\|\mu_{B_j} - \mu_i\|_2 \leq 13C\sigma_i\sqrt{|S_i|/|B_j|}$ and $\sigma_{B_j} \leq 20C\sigma_i\sqrt{|S_i|/|B_j|}$.*
4. *$|\mathcal{B}_i| \geq 0.93\alpha n$ for $i \in [k]$.*

Proof. As in the previous lemma, we first note that Item 4 follows directly from Item 1.

$$|\mathcal{B}_i| \geq |\mathcal{B}_i \cap S_i| \geq 0.967|S_i| \geq 0.93\alpha n \,,$$

where the second inquality uses Item 1 and the last inequality uses $|S_i| \geq 0.97\alpha n$ by the setup in Theorem 3.2.

If $A_1, \dots, A_m$ is the Voronoi clustering before filtering and $\mathcal{A}_1, \dots, \mathcal{A}_k$ as in Lemma 5.1, then by that lemma: $|S_i \setminus \mathcal{A}_i| \leq 0.011|S_i|$, $|\mathcal{A}_i \setminus S_i| \leq 0.03\alpha n$ and $|\mathcal{A}_i| \geq 0.959\alpha n$ for all $i \in [k]$. In everything that follows we assume $|A_j| \geq 0.96\alpha n$. Let B_j denote the filtered sets output by the algorithm of Fact 2.2 on input A_j.

Proof of Item 3: Recall that the outputs B_j of Algorithm 3 are filtered versions of the sets A_j from the Voronoi partition. Item 3 states that the filtered version $B_j \subseteq \mathcal{B}_i$ must have mean close to μ_i and covariance not too large. We check this by showing the preconditions of Fact 2.2 (applied with $\epsilon = 0.04$), and then Item 3 follows from applying the fact with A_j as the set T from the fact statement and $A_j \cap S_i$ as the set S in that statement, where i here is the index for which $A_j \subseteq \mathcal{A}_i$.

We will apply Fact 2.2 with $\epsilon = 0.04$. For this to be applicable, we need to ensure that $|T \setminus S| \leq 0.04|T|$, which using A_j in place of T and $A_j \cap S_i$ in place of S becomes $|A_j \setminus S_i| \leq 0.04|A_j|$. Applying Fact 2.2 also requires that $A_j \cap S_i$ is stable (Definition 2.1). We start by establishing the first requirement, that $|A_j \setminus S_i| \leq 0.04|A_j|$:

$$\begin{aligned}
|A_j \cap S_i| &= |A_j| - |A_j \setminus S_i| \\
&\geq |A_j| - |\mathcal{A}_i \setminus S_i| && \text{(since } A_j \subseteq \mathcal{A}_i\text{)} \\
&\geq |A_j| - 0.03\alpha n && (|\mathcal{A}_i \setminus S_i| \leq 0.03\alpha n \text{ by Lemma 5.1}) \\
&\geq 0.96|A_j| \,, && (6.11)
\end{aligned}$$

where the last line uses that we have assumed $|A_j| \geq 0.96\alpha n$. Using the above $|A_j \setminus S_i| = |A_j| - |A_j \cap S_i| \leq 0.04|A_j|$, as desired.

We now establish the second requirement, that $A_j \cap S_i$ is stable (Definition 2.1). To this end, since S_i was assumed to be $(C, 0.04)$-stable with respect to μ_i and σ_i, then using Lemma 2.1 we have that $A_j \cap S_i$ is $(1.23C\sqrt{|S_i|}/\sqrt{0.04|A_j \cap S_i|}, 0.04)$-stable with respect to μ_i, σ_i.

The first part of the conclusion of Fact 2.2 is that if B_j denotes the output of the filtering algorithm run on A_j, it holds $|B_j| \geq 0.96|A_j|$, the second part states that

$$\|\mu_{B_j} - \mu_i\|_2 \leq 12.3\, C\sigma_i \sqrt{\frac{|S_i|}{0.04|A_j \cap S_i|}}\sqrt{0.04} \leq 13C\sigma_i\sqrt{\frac{|S_i|}{|B_j|}}$$

where the last inequality above is because $|B_j| \le |A_j| \le |A_j \cap S_i|/0.96$, where the last step here is because of (6.11).

Similarly, the third part of the conclusion of Fact 2.2 is that $\sigma_{B_j} \le 20C\sigma_i\sqrt{|S_i|/|B_j|}$. Lastly, we check that the condition on the size of the sets from Fact 2.2 is indeed satisfied because $|A_j| \ge 0.96\alpha n \gg \log(1/(\alpha\delta))$, where we used the assumption on the size of n from Theorem 3.2.

Proof of Item 1: We have already shown that Fact 2.2 is applicable for analyzing the effect of the filtering algorithm on input A_j and thus $|A_j \setminus B_j| \le 0.04|A_j|$ (first part of the conclusion of Fact 2.2). Then,

$$
\begin{aligned}
|S_i \setminus \mathcal{B}_i| &= |S_i \setminus \mathcal{A}_i| + \sum_{j:A_j \subseteq \mathcal{A}_i} |A_j \setminus B_j| \\
\text{(by Lemma 5.1 and Fact 2.2)} \quad &\le 0.011|S_i| + 0.04 \sum_{j:A_j \subseteq \mathcal{A}_i} |A_j| \\
(A_j\text{'s are disjoint}) \quad &= 0.011|S_i| + 0.04|\mathcal{A}_i| \\
&= 0.011|S_i| + 0.04(|\mathcal{A}_i \cap S_i| + |\mathcal{A}_i \setminus S_i|) \\
(\ |\mathcal{A}_i \setminus S_i| \le 0.012\alpha n \text{ by Lemma 5.1}) \quad &\le 0.011|S_i| + 0.04(|S_i| + 0.03\alpha n) \\
\text{(by assumption that } |S_i| \ge 0.97\alpha n) \quad &\le 0.011|S_i| + 0.04\left(|S_i| + \frac{0.03}{0.97}|S_i|\right) \le 0.033|S_i|
\end{aligned}
$$

Proof of Item 2: We have that $|\mathcal{A}_i \setminus S_i| \le 0.03\alpha n$ before the filtering takes place. Since filtering only removes points, $\mathcal{B}_i \subseteq \mathcal{A}_i$ and thus $|\mathcal{B}_i \setminus S_i| \le 0.03\alpha n$ continues to hold after the filtering.
□

Having shown guarantees on the filtered Voronoi cells, we now give Algorithm 4, used in Line 6 of Algorithm 1, which is responsible for further pruning the candidate means in L such that the resulting filtered Voronoi cells are well-separated. Lemma 6.1 gives the guarantees of Algorithm 4.

Algorithm 4 Distance-based pruning of sub-clusters

Input: Dataset T of n points, centers $\hat{\mu}_1, \dots, \hat{\mu}_m$, and parameter $\alpha \in (0, 1)$.
Output: A subset $\hat{\mu}_1, \dots, \hat{\mu}_{m'}$ of the input centers.

1. $\{B_1, \dots, B_m\} \leftarrow$ FilteredVoronoi$(\{\hat{\mu}_1, \dots, \hat{\mu}_m\}, T)$.
2. $J_{\text{deleted}} \leftarrow \emptyset$.
3. While there exist j, j' with $\|\mu_{B_j} - \mu_{B_{j'}}\|_2 \le 4761C(\sigma_{B_j} + \sigma_{B_{j'}})/\sqrt{\alpha}$:
   (a) Calculate $d = \min_{t\in[m]} \|\mu_{B_j} - \mu_{B_t}\|_2/\sigma_{B_j}$ and $d' = \min_{t\in[m]} \|\mu_{B_{j'}} - \mu_{B_t}\|_2/\sigma_{B_{j'}}$.
   (b) If $d < d'$:
      i. $j_{\text{deleted}} \leftarrow j$.
   (c) Else:
      i. $j_{\text{deleted}} \leftarrow j'$.
   (d) Update $J_{\text{deleted}} \leftarrow J_{\text{deleted}} \cap \{j_{\text{deleted}}\}$
   (e) Update $\{B_j\}_{j\in[m]\setminus J_{\text{deleted}}} \leftarrow$ FilteredVoronoi$(\{\hat{\mu}_j\}_{j\in[m]\setminus J_{\text{deleted}}}, T)$.
4. Output $\hat{\mu}_j$ for $j \in [m] \setminus J_{\text{deleted}}$ after relabeling the indices so that they are from 1 to $m - |J_{\text{deleted}}|$.

LEMMA 6.1. (DISTANCE-BASED PRUNING OF SUB-CLUSTERS) *Consider the setting and notation of Theorem 3.2. Let $L = \{\hat{\mu}_1, \ldots, \hat{\mu}_m\}$ be a list of vectors for some $m \geq k$. Suppose the list L can be partitioned into sets $H_1, \ldots, H_k$ such that for every $i \in [k]$, H_i consists of the vectors $\hat{\mu}_j$ with $\|\hat{\mu}_j - \mu_i\|_2 \leq 4061C^2\sigma_i/\sqrt{\alpha}$, and that $H_i \neq \emptyset$ for all $i \in [k]$. Also assume that every set in the Voronoi partition $A_j = \{x : \arg\min_{j'} \|x - \hat{\mu}_{j'}\|_2 = j\}$ for $j \in [m]$ has size $|A_j| \geq 0.96\alpha n$. Consider an execution of DISTANCEBASEDPRUNING(L, T, α) algorithm (Algorithm 4) with the list L, the entire dataset of points T and the parameter α as input.*

After the algorithm terminates, let $\hat{\mu}'_1, \ldots, \hat{\mu}'_{m'}$ be the output list (where we denote by m' its size). Then the following three statements hold with probability at least $1 - \delta/2$:

1. *The output list $\{\hat{\mu}'_j\}_{j\in[m']}$ can be partitioned into sets $H'_1, \ldots, H'_k$ such that for every $i \in [k]$, H'_i consists of the vectors of $\hat{\mu}'_j$ with $\|\hat{\mu}'_j - \mu_i\|_2 \leq 4061C^2\sigma_i/\sqrt{\alpha}$ and it holds $H'_i \neq \emptyset$ for all $i \in [k]$.*
2. *Every set in the Voronoi partition corresponding to the output centers $A'_j = \{x : \arg\min_{j'\in[m']} \|x - \hat{\mu}_{j'}\|_2 = j\}$ for $j \in [m']$ has size $|A'_j| \geq 0.96\alpha n$.*
3. *If $B'_1, \ldots, B'_{m'}$ denote the output of FILTEREDVORONOI$(\{\hat{\mu}'_1, \ldots, \hat{\mu}'_{m'}\}, T)$ for the non-deleted centers, then it holds that $\|\mu_{B'_j} - \mu_{B'_{j'}}\|_2 \geq 4761C(\sigma_{B'_j} + \sigma_{B'_{j'}})/\sqrt{\alpha}$ for every $j, j' \in [m']$ with $j \neq j'$.*

Proof. The final part of the lemma conclusion, Item 3, holds by design of the stopping condition of our algorithm (line 3).

We show the remaining parts (Items 1 and 2) by induction. That is, we will fix an iteration of the algorithm, assume that Items 1 and 2 hold just before the iteration starts, and prove that they continue to hold after the iteration ends. More specifically, since in each iteration we use FILTEREDVORONOI, which is randomized, we may allow a probability of failure for each step in our inductive hypothesis, in particular, we will use probability of failure ($\delta/2$) divided by the maximum number of iterations (so that by Fact 6.1, the conclusion holds after all iterations end with probability at least $1 - \delta/2$).

FACT 6.1. *If event A happens with probability $1 - \tau_1$ and event B happens with probability $1 - \tau_2$ conditioned on event A, then the probability of both A and B happening is at least $1 - \tau_1 - \tau_2$.*

The upper bound on the number of iterations can be trivially seen to be $1/(0.96\alpha)$. This is because we assumed that every Voronoi set in the beginning has size $|A_j| \geq 0.96\alpha n$ and the algorithm only deletes one of the candidate means at a time, thus the algorithm will trivially terminate after $1/(0.96\alpha)$ steps. It therefore suffices to show that the inductive step of our proof holds with probability at least $1 - 0.1\alpha\delta$.

Since the iteration under consideration alters the list of vectors and some associated quantities, we must ensure that our notation reflects the specific moment within the algorithm. To achieve this, we will use unprimed letters to represent quantities at the moment just before the iteration begins ($J_{\text{deleted}}, H_i, A_i, B_i$) for the set of deleted indices appearing in the pseudocode, the partition, the Voronoi clustering, and the filtered Voronoi clustering), and primes to denote the quantities ($J'_{\text{deleted}}, H'_i, A'_i, B'_i$) after the iteration ends. That is, our inductive hypothesis is that

(a) The list $\{\hat{\mu}_j\}_{j\in[m]\setminus J_{\text{deleted}}}$ can be partitioned into sets $H_1, \ldots, H_k$ such that for every $i \in [k]$, H_i consists of the vectors of $\hat{\mu}_j$ with $\|\hat{\mu}_j - \mu_i\|_2 \leq 4061C^2\sigma_i/\sqrt{\alpha}$ and it holds $H_i \neq \emptyset$ for all $i \in [k]$.

(b) Every set in the Voronoi partition corresponding to the centers $A_j = \{x : \arg\min_{j\in[m]\setminus J_{\text{deleted}}} \|x - \hat{\mu}_{j'}\|_2 = j\}$ for $j \in [m] \setminus J_{\text{deleted}}$ has size $|A_j| \geq 0.96\alpha n$.

And we will show that after the iteration ends, if J'_{deleted} denotes the updated set of deleted indices (i.e. the set that also includes the index that was deleted during the current iteration), and A'_j, B'_j denote the Voronoi sets and filtered Voronoi sets corresponding to the centers $\hat{\mu}_j$ for $j \in [m] \setminus J'_{\text{deleted}}$, the following hold with probability at least $1 - 0.1\alpha\delta$:

1. The updated list $\{\hat{\mu}_j\}_{j\in[m]\setminus J'_{\text{deleted}}}$ can be partitioned into sets $H'_1, \ldots, H'_k$ such that for every $i \in [k]$, H'_i consists of the vectors of $\hat{\mu}_j$ with $\|\hat{\mu}_j - \mu_i\|_2 \leq 4061C^2\sigma_i/\sqrt{\alpha}$ and it holds $H'_i \neq \emptyset$ for all $i \in [k]$.
2. Every set in the Voronoi partition corresponding to the updated centers A'_j, where $A'_j = \{x : \arg\min_{j\in[m]\setminus J'_{\text{deleted}}} \|x - \hat{\mu}_{j'}\|_2 = j\}$ for $j \in [m] \setminus J'_{\text{deleted}}$ has size $|A'_j| \geq 0.96\alpha n$.

Now observe that, by construction, every iteration only deletes a vector from the list, and therefore the list $\{\hat{\mu}_j\}_{j\in[m]\setminus J'_{\text{deleted}}}$ can be partitioned into the sets $H'_1,\dots,H'_k$ satisfying the first part of Item 1 (that H'_i consists of the vectors of $\hat{\mu}_j$ with $\|\hat{\mu}_j-\mu_i\|_2\le 4061C^2\sigma_i/\sqrt{\alpha}$). Regarding Item 2, this trivially holds because deleting a point, can only make the Voronoi clusters bigger in size. The only nontrivial condition to check is that H'_i remains non-empty for all $i\in[k]$. Equivalently, we need to show that, if at the beginning of an iteration, H_i consists of only a single vector, then it will never be removed in the iteration.

By our inductive hypothesis that the partition $H_1,\dots,H_k$ with the aforementioned properties exists (Item (a)) and our assumption that $|A_j|\ge 0.96\alpha n$ (Item (b) of inductive hypothesis), Corollary 6.1 is applicable. The application of that implies that the following holds with probability at least $1-0.1\alpha\delta$: Denote by $\mathcal{B}_i=\cup_{j\in[m]\setminus J_{\text{deleted}}:\hat{\mu}_j\in H_i}B_j$ for $i\in[k]$, i.e. $\mathcal{B}_i$ is the union of all Voronoi clusters corresponding to (non-deleted) centers in H_i. Then,

(i) $\mathcal{B}_i\neq\emptyset$ for $i\in[k]$.

(ii) For any $j\in[m]\setminus J_{\text{deleted}}$ such that $\hat{\mu}_j\in H_i$ it holds $\|\mu_{B_j}-\mu_i\|_2\le 14C\sigma_i/\sqrt{\alpha}$ and $\sigma_{B_j}\le 21C\sigma_i/\sqrt{\alpha}$.

(iii) For every $i\in[k]$ with $|H_i|=1$, if $j\in[m]\setminus J_{\text{deleted}}$ denotes the unique index for which $B_j=\mathcal{B}_i$, then it holds $\sigma_{B_j}\le 21C\sigma_i$.

The second statement above can be extracted from Item 3 of Corollary 6.1 after noting that $|B_j|\ge|A_j|-|A_j\setminus B_j|\ge 0.96|A_j|\ge 0.92\alpha n\ge 0.94\alpha|S_i|$, where we used $|A_j\setminus B_j|\le 0.04|A_j|$ (Item 3 of Corollary 6.1) and the assumption that $|A_j|\ge 0.96\alpha n$. The third statement ((iii) above) can be extracted from Item 3 of Corollary 6.1 after noting that $|B_j|\ge|B_j\cap S_i|\ge|S_i|-|S_i\setminus B_j|\ge 0.967|S_i|$, where we used that $|S_i\setminus B_j|\le 0.033|S_i|$ by Item 1 of Corollary 6.1.

We will also use the notation $\text{par}(B_j)$ to denote the index $i\in[k]$ for which it holds $\|\mu_{B_j}-\mu_i\|_2\le 35C\sigma_i/\sqrt{\alpha}$ (by the fact that $\mathcal{B}_i\neq\emptyset$ mentioned above and the separation assumption for the μ_i's, such an index indeed exists and it is unique). We will call $\text{par}(B_j)$ the "parent" of B_j. By slightly overloading this notation, we will also use $\text{par}(\hat{\mu}_j)$ to denote the index i for which it holds $\hat{\mu}_j\in H_i$, i.e. $\|\hat{\mu}_j-\mu_i\|_2\le 4061C^2\sigma_i/\sqrt{\alpha}$. We will also informally call the B_j's "sub-clusters" (as opposed to the sets S_i that we call "true" or "parent" clusters).

Using this notation, and further denoting by j_{deleted} the index of the vector deleted in the current iteration, what remains to check is equivalent to the statement that $|H_{\text{par}(j_{\text{deleted}})}|>1$.

To show this, we need Claim 3 below, which states the straightforward fact that sub-clusters with the same parent cluster will have means close to each other, and sub-clusters with different parents necessarily have means much farther. This in particular implies that, given a sub-cluster B_j, the closest sub-cluster must share the same parent if $|B_j|>1$. We will now use Claim 3 to show the statement that the deleted vector $\hat{\mu}_{j_{\text{deleted}}}$ must have $|H'_{\text{par}(j_{\text{deleted}})}|>1$, and provide the simple proof of Claim 3 at the end, which follows from straightforward applications of the reverse triangle inequality.

CLAIM 3. *The following holds for every for $j,j'\in[m]\setminus J_{\text{deleted}}$ with $j\neq j'$: Denote by $\ell:=\text{par}(B_j)$, $\ell':=\text{par}(B_{j'})$. If $\ell=\ell'$, then $\|\mu_{B_j}-\mu_{B_{j'}}\|_2\le 28C\sigma_\ell/\sqrt{\alpha}$, otherwise, $\|\mu_{B_j}-\mu_{B_{j'}}\|_2>10^4C^2(\sigma_\ell+\sigma_{\ell'})/\sqrt{\alpha}$.*

We will now show our end goal using a case analysis (and Claim 3). Denote by $B_j,B_{j'}$ the sub-clusters that are identified in line 3 of Algorithm 4 (i.e. one of j or j' will eventually be what we called j_{deleted} before). We need to show that, if the index j is the one that gets deleted, then $|H'_{\text{par}(j)}|>1$, and similarly for j'. We check each of the following cases:

1. (Case where $|H_{\text{par}(B_j)}|=1,|H_{\text{par}(B_{j'})}|>1$) Let ℓ,ℓ' be the parents of B_j and $B_{j'}$ respectively. We first note that $\sigma_{B_j}\le 21C\sigma_\ell$ by the third property of B_j (Item (iii)). Now we argue that, since j and j' are flagged by line 3 of Algorithm 4, it must be the case that $\sigma_{B_{j'}}>21C\sigma_{\ell'}$, for otherwise:

$$\begin{aligned}
&\text{(reverse triangle inequality)} && \|\mu_{B_j}-\mu_{B_{j'}}\|\ge\|\mu_\ell-\mu_{\ell'}\|-\|\mu_{B_{j'}}-\mu_{\ell'}\|-\|\mu_{B_j}-\mu_\ell\|\\
&\text{(by separation assumption and Item (ii))} && \ge 10^5C^2(\sigma_\ell+\sigma_{\ell'})/\sqrt{\alpha}-14C\sigma_{\ell'}/\sqrt{\alpha}-14C\sigma_\ell/\sqrt{\alpha}\\
&\text{(6.12)} && \ge(10^5-14)C^2(\sigma_\ell+\sigma_{\ell'})/\sqrt{\alpha}\\
&\text{(using } \sigma_{B_j}\le 21C\sigma_\ell,\ \sigma_{B_{j'}}\le 21C\sigma_{\ell'}) && \ge 4761C(\sigma_{B_j}+\sigma_{B_{j'}})/\sqrt{\alpha}
\end{aligned}$$

Having shown that $\sigma_{B_j} \leq 21C\sigma_\ell$ and $\sigma_{B_{j'}} > 21C\sigma_{\ell'}$, we will now show that the center $\hat{\mu}_j$ corresponding to B_j will not be the one deleted in this loop iteration, and instead the center $\hat{\mu}_{j'}$ corresponding to $B_{j'}$ will be the one that will get deleted. To see that, denote by d and d' the same quantities as in line 3a of the pseudocode, i.e. the normalized distances of the sub-clusters from their closest other sub-clusters.

On the one hand, we have that

$$d := \frac{\min_{t\in[m]\setminus J_{\text{deleted}}} \|\mu_{B_j} - \mu_{B_t}\|_2}{\sigma_{B_j}} \geq \frac{(10^5-14)C^2\sigma_\ell}{\sqrt{\alpha}\sigma_{B_j}} \geq \frac{4761C}{\sqrt{\alpha}} ,$$

where the first step follows by the fact that the closest sub-cluster to B_j must have as parent a different true cluster (because $|\text{par}(B_j)| = 1$), and since true clusters are sufficiently separated, the closest sub-cluster to B_j must be at least $(10^5-14)C^2\sigma_\ell/\sqrt{\alpha}$-away (see the derivation of (6.12) for an identical proof). The last step uses that $\sigma_{B_j} \leq 21C\sigma_\ell$.

On the other hand, for the (normalized) distance of $B_{j'}$ to its closest sub-cluster (denote that sub-cluster by B_{t^*}) we have the following: First note that B_{t^*} must have the same parent as $B_{j'}$ due to Claim 3 and $|\text{par}(B_{j'})| > 1$. Then, since both have ℓ' as their parent,

$$\begin{aligned}
d' &:= \frac{\min_{t\in[m]\setminus J_{\text{deleted}}} \|\mu_{B_{j'}} - \mu_{B_t}\|_2}{\sigma_{B_{j'}}} \\
&= \frac{\|\mu_{B_{j'}} - \mu_{B_{t^*}}\|_2}{\sigma_{B_{j'}}} \\
&\leq \frac{\|\mu_{B_{j'}} - \mu_{\ell'}\|_2}{\sigma_{B_{j'}}} + \frac{\|\mu_{\ell'} - \mu_{B_{t^*}}\|_2}{\sigma_{B_{j'}}} \\
\text{(by Item (ii) and } C>1\text{)} \quad &\leq 2\cdot\frac{14C^2\sigma_{\ell'}}{\sqrt{\alpha}\sigma_{B_{j'}}} \\
\text{(using } \sigma_{B_{j'}} \geq 21C\sigma_{\ell'}\text{)} \quad &\leq 2C/\sqrt{\alpha} .
\end{aligned}$$

This means that $d' < d$ and line 3(c)i of the algorithm will delete $\hat{\mu}_{j'}$, i.e. the eliminated center is not the only center of its parent.

2. (Case $|H_{\text{par}(B_j)}| > 1, |H_{\text{par}(B_{j'})}| = 1$) Symmetric to the previous case.

3. (Case $|H_{\text{par}(B_j)}| > 1, |H_{\text{par}(B_{j'})}| > 1$) This case is straightforward. In this case, both parents have more than one centers, thus no matter which center the algorithm deletes, the eliminated center is not the only center of its parent.

4. ($|H_{\text{par}(B_j)}| = 1, |H_{\text{par}(B_{j'})}| = 1$) In this case we argue that $B_j, B_{j'}$ could not have been identified in line 3 of Algorithm 4, meaning that this is not a valid case to consider. To show this, let ℓ, ℓ' be the parents of B_j and $B_{j'}$ respectively. By the second and third properties of B_j, $\sigma_{B_j} \leq 21C\sigma_\ell$ and $\|\mu_{B_j} - \mu_\ell\| \leq 14C\sigma_\ell/\sqrt{\alpha}$. Similarly, $\sigma_{B_{j'}} \leq 21C\sigma_{\ell'}$ and $\|\mu_{B_{j'}} - \mu_{\ell'}\| \leq 14C\sigma_{\ell'}/\sqrt{\alpha}$.

$$\begin{aligned}
\|\mu_{B_j} - \mu_{B_{j'}}\| &\geq \|\mu_\ell - \mu_{\ell'}\| - \|\mu_{B_j} - \mu_\ell\| - \|\mu_{B_{j'}} - \mu_{\ell'}\| \\
&\geq 10^5C^2(\sigma_\ell + \sigma_{\ell'})/\sqrt{\alpha} - 14C\sigma_\ell/\sqrt{\alpha} - 14C\sigma_{\ell'}/\sqrt{\alpha} \\
&\geq (10^5-14)C^2(\sigma_\ell + \sigma_{\ell'})/\sqrt{\alpha} \\
&> 4761C(\sigma_{B_j} + \sigma_{B_{j'}})/\sqrt{\alpha} .
\end{aligned}$$

The above means that the check of line 3 in Algorithm 4 could not be satisfied for $B_j, B_{j'}$.

It only remains to prove Claim 3.

Proof. [Proof of Claim 3] Let $B_j, B_{j'}$ be sub-clusters with the same parent ℓ. Then by Item (ii) and a triangle inequality, $\|\mu_{B_j} - \mu_{B_{j'}}\| \leq \|\mu_{B_j} - \mu_\ell\| + \|\mu_{B_{j'}} - \mu_\ell\| \leq 28C\sigma_\ell/\sqrt{\alpha}$.

Now, if B_j has parent ℓ and $B_{j'}$ has parent ℓ', then by Item (ii) and reverse triangle inequality:

$$\begin{aligned}\|\mu_{B_j} - \mu_{B_{j'}}\| &\geq \|\mu_\ell - \mu_{\ell'}\| - \|\mu_\ell - \mu_{B_j}\| - \|\mu_{\ell'} - \mu_{B_{j'}}\| \\ &\geq 10^5 C^2(\sigma_\ell + \sigma_{\ell'})/\sqrt{\alpha} - 14C\sigma_\ell/\sqrt{\alpha} - 14C\sigma_{\ell'}/\sqrt{\alpha} \\ (C > 1) \qquad &> 10^4 C^2(\sigma_\ell + \sigma_{\ell'})/\sqrt{\alpha}\,.\end{aligned}$$

□

□

7 Overall analysis of Algorithm 1

In this brief section, we combine the results and analyses in Sections 4 to 6 to prove Theorem 3.2.

Proof. [Proof of Theorem 3.2]

Let $s_{\max}$ denote the maximum element of the list L_{stdev} created in line 1. By Corollary 4.1, after the loop of line 4 ends, the list L of candidate mean vectors that the algorithm has created is such that (i) for every element $\hat{\mu}_j \in L$ there exists a true cluster S_i such that $\|\hat{\mu}_j - \mu_i\|_2 \leq 4061C^2\sigma_j/\sqrt{\alpha}$, and (ii) for every true cluster S_i with $\sigma_{S_i} \leq s_{\max}$, there exists a $\hat{\mu}_j \in L$ such that $\|\hat{\mu}_j - \mu_i\|_2 \leq 4061C^2\sigma_j/\sqrt{\alpha}$. Furthermore, we also know by Proposition 2.1 that, for every true cluster S_i, there exists an $\hat{s}$ in the list such that $\sigma_{S_i} \leq \hat{s}$. This implies that $s_{\max} \geq \max_i \sigma_{S_i}$, and guarantee (ii) above applies to every true cluster S_i.

Following the structure of the algorithm, we use Lemma 5.2 to reason about line 5 of Algorithm 1. To check that the lemma is indeed applicable, we need to show that L can be partitioned into disjoint sets $H_1, \ldots, H_k$ such that for every $i \in [k]$, H_i consists of the vectors $\hat{\mu}_j$ satisfying $\|\hat{\mu}_j - \mu_i\|_2 \leq 4061C^2\sigma_i/\sqrt{\alpha}$, and that $H_i \neq \emptyset$ for all $i \in [k]$. This is indeed true for the sets $H_i := \{\hat{\mu} \in L : \|\hat{\mu} - \mu_i\|_2 \leq 4061C^2\sigma_i/\sqrt{\alpha}\}$. The sets are disjoint because of our assumption that $\|\mu_i - \mu_{i'}\|_2 > 10^5C^2(\sigma_i + \sigma_{i'})/\sqrt{\alpha}$ for every $i \neq i'$, and their union is equal to the entire L because of the guarantee (i) from the previous paragraph. Finally, the fact that $H_i \neq \emptyset$ for all $i \in [k]$ holds because of the guarantee (ii) of the previous paragraph.

The conclusion of Lemma 5.2 is that, after we apply the SizeBasedPruning algorithm in line 5 of Algorithm 1, the resulting list L' will admit a partition $H'_1, \ldots, H'_k$ with the same properties as before, but also with the added property that every Voronoi cluster $A'_j := \{x \in T : \arg\min_{\hat{\mu}_{j'} \in L'} \|x - \hat{\mu}_{j'}\|_2 = j\}$ for $j \in [|L'|]$ that corresponds to the centers of the output list L', satisfies $|A'_j| \geq 0.96\alpha n$.

Next we use Lemma 6.1 to analyze the application of DistanceBasedPruning to the list L' in line 6 of Algorithm 1. Let us use L'' to denote the output of DistanceBasedPruning(L', T, α). The lemma is applicable because of the conclusion of the previous paragraph. In turn, the conclusion of Lemma 6.1 is that with probability at least $1 - \delta/2$ (over the randomness of the algorithm, in particular, the uses of filtering from Fact 2.2),

(a) The list L'' of centers admits a partition $H''_1, \ldots, H''_k$ with the same properties as before.

(b) Every set in the Voronoi partition corresponding to these centers $A''_j = \{x : \arg\min_{\hat{\mu}_{j'} \in L''} \|x - \hat{\mu}_{j'}\|_2 = j\}$ have sizes $|A''_j| \geq 0.96\alpha n$.

(c) If $B''_1, \ldots, B''_{|L''|}$ denote the output of FilteredVoronoi(L'', T) then it holds that $\|\mu_{B''_j} - \mu_{B''_{j'}}\|_2 \geq 4761C(\sigma_{B''_j} + \sigma_{B''_{j'}})/\sqrt{\alpha}$ for every $j \neq j'$.

Note that FilteredVoronoi(L'', T) is the last step of Algorithm 1. We will show that all the guarantees of the output Theorem 3.2 follow by Items (a) to (c) and a final application of Corollary 6.1 (which is applicable because of Items (a) and (b) above):

Item 1 in the conclusion of Theorem 3.2 is true by the fact that $|A''_j| \geq 0.96\alpha n$ from Item (b) and the fact that the filtering in FilteredVoronoi(L, T) only removes 4% of the points in A''_j (see Item 2 in Corollary 6.1) with probability $1 - \delta/2$.

For Item 2 in the conclusion of Theorem 3.2, we have the following: Item 2a holds by Item 4 in Corollary 6.1. Item 2b holds by Item 1 in Corollary 6.1. Item 2c holds by Item 2 in Corollary 6.1. Item 2d follows from Item 3 in Corollary 6.1. Item 2e holds by Item (c) above.

Moreover, the number m of the output sets $B_1, \ldots, B_m$ is at most $1/(0.92\alpha)$ since each set has at least $0.94\alpha n$ points and the sets are disjoint.

Finally, the algorithm runs in time poly(nd/α)-time because the size of the lists $L_{\text{mean}}, L_{\text{stdev}}$ is polynomial in n and $1/\alpha$, which means that the size of L is also polynomial, and finally since the two pruning algorithms in lines 5 and 6 delete one element of L at each step until termination, the overall number of steps is polynomial. It can also be checked that each step involves calculations that can be implemented in poly(nd/α)-time.

□

8 Clustering under the no large sub-cluster condition

The previous section analyzes Algorithm 1 in the general case, where the underlying mixture satisfies information-theoretically optimal separation, and the algorithm only knows a lower bound α to the mixing weight. As we have shown in the introduction, we cannot aim to return an accurate clustering close to the ground truth, but instead, we return an accurate *refinement* of the ground truth clustering.

In this section, we study the no large sub-cluster (NLSC) condition (Section 1.1), which is a deterministic condition on the sample set that guarantees that Algorithm 1 in fact returns an accurate clustering instead of just a refinement. We first compare our NLSC condition with that proposed by [BKK22]. Even though the conditions are qualitatively similar, our choice of parameters makes our NLSC condition a stronger assumption. We explain in Section 8.1 why the stronger NLSC assumption is necessary due to our weaker separation assumption.

We then show in Section 8.2 that, under the NLSC condition (Section 1.1), Algorithm 1 will return exactly k sets, one per mixture component, despite not knowing k. This is stated as Corollary 8.1, the formal version Corollary 1.1. Afterwards, we also show that the general class of well-conditioned log-concave distributions yield samples that satisfy this condition with high probability, as long as the dimensionality is large and the sample complexity is polynomially large.

8.1 Comparison with the NLSC condition from [BKK22]

In this subsection, we compare our NLSC condition (Section 1.1) with the NLSC condition proposed by [BKK22]. For the reader's convenience, we restate Section 1.1 below.

[NLSC condition] We say that the disjoint sets $S_1, \ldots, S_k$ of total size n satisfy the "No Large Sub-Cluster" condition with parameter α if for any cluster S_i and any subset $S' \subset S_i$ with $|S'| \geq 0.8\alpha n$, it holds that $\sigma_{S'} \geq 0.1\sigma_{S_i}$, where $\sigma_{S'}$ is the square root of the largest eigenvalue of the covariance matrix of S'.

For contrast, the NLSC condition of [BKK22] is weaker. Instead of $\sigma_{S'}$ being within a constant factor of σ_{S_i}, their requirement can be as small as an α factor of σ_{S_i}.

Definition 8.1. (NLSC condition of [BKK22]) *We say that the disjoint sets $S_1, \ldots, S_k$ of total size n satisfy the "No Large Sub-Cluster" condition of [BKK22] with parameter α if for any cluster S_i and any subset $S' \subset S_i$ with $|S'| \geq 0.01\sqrt{n}\log n$, it holds that $\sigma_{S'} \geq \frac{1}{5\sqrt{5}}\frac{|S'|}{|S_i|}\sigma_{S_i}$, where $\sigma_{S'}$ is the square root of the largest eigenvalue of the covariance matrix of S'.*

The two differences between the definitions are *(i)* the minimum size of the subset S' being considered, which is an insignificant difference, and more importantly *(ii)* the lower bound of $\sigma_{S'}$. In our definition, the lower bound is a small constant factor of σ_{S_i}, but their definition uses a factor that scales with the ratio of the set sizes, potentially interpolating between $\Theta(\alpha)$ and $\Theta(1)$.

We will now show that, under the separation assumption of $C \cdot (\sigma_i + \sigma_j)/\sqrt{\alpha}$ between pairs of clusters, for any large constant C, there is an explicit construction of a sample set where the NLSC condition of [BKK22] allows for two substantially different clusterings satisfying the separation assumption, whereas our NLSC condition (by the result of Corollary 8.1 below) only allow clusterings that are essentially the same as each other.

First consider a 1-dimensional set of points U of αn points, distributed as a uniform grid over the interval $[-\frac{1}{2}, \frac{1}{2}]$. Its mean is 0, and its variance is $\frac{1}{12} - o(1)$, where the $o(1)$ term goes to 0 as $\alpha n \to \infty$. It is straightforward to check via a "swapping" argument that, for any subset U' of size at least $0.8\alpha n$ (which is thus at least a 0.8-fraction of U), we have $\sigma_{U'} \geq 0.8\sigma_U \geq \frac{1}{5\sqrt{5}}\frac{|U'|}{|U|}\sigma_U$.

We then use U to construct a high-dimensional sample set. Set the ambient dimensionality to be $d = 1/(2\alpha)$. We will embed a set U along each Euclidean axis, symmetrically in the positive and negative coordinates. For $i \in [d]$, construct the set $S_i^+ = \{(x + (C/\sqrt{\alpha}))e_i : x \in U\}$, which is a set of points that are non-zero only in the i^{th} coordinate, embedded on the positive side of the i^{th} axis, and similarly construct $S_i^- = \{(x - (C/\sqrt{\alpha}))e_i : x \in U\}$. Let the set S be the union of all these S_i^+ and S_i^- across $i \in [d]$, giving a total of n points.

We claim that, according to the NLSC condition of [BKK22], there are two very different but both valid clusterings of S: i) treating every S_i^+ and S_i^- as a separate cluster, and ii) treating the entire S as a single cluster. We now verify both clusterings.

Recall that, to verify the validity of a clustering, we need to check that a) each cluster has size at least αn, b) the clusters are well-separated, and c) the NLSC condition of [BKK22] is satisfied.

For the clustering treating each S_i^+ and S_i^- as separate clusters, point (a) is trivial, and (c) is true by construction of U. It remains to check the cluster separation assumption (point (b) above). The minimum distance between (the means of) a pair of clusters is $\sqrt{2}C/\sqrt{\alpha}$, and each cluster has variance upper bounded by $1/12$. On the other hand, the required separation is $C \cdot (1/\sqrt{12} + 1/\sqrt{12})/\sqrt{\alpha} < \sqrt{2}C/\sqrt{\alpha}$. Thus the separation assumption is indeed satisfied.

Now consider the clustering treating the entire set S as a single cluster. Point (a) is again trivial, and so is point (b). It remains to check point (c), which is the NLSC condition of [BKK22].

By construction of the set S, its mean is 0 and its covariance matrix is a multiple of the identity. We bound above its variance along an axis direction, in order to establish the NLSC condition of [BKK22]. By the law of total variance, we can write

$$\mathrm{Cov}(S)_{ii} = 2\alpha \,\mathrm{Var}(U) + 2\alpha(C/\sqrt{\alpha})^2 \le \alpha/6 + 2C^2,$$

since $\mathrm{Var}(U) \le 1/12$. As long as $\alpha \ll 1$ and $C > 1$, we have that $\|\mathrm{Cov}(S)\|_{\mathrm{op}} = \mathrm{Cov}(S)_{ii} \le 2.1C^2$.

Now consider an arbitrary subset $S' \subseteq S$ (in fact, we will not need to lower bound its size for the analysis). By an averaging argument, there must exist some dimension i such that at least $2\alpha|S'|$ points of S' lie in $S_i^+ \cup S_i^-$. We will lower bound the variance of S' in direction e_i.

Either at least 50% of the points in $|S'|$ lie in $S_i^+ \cup S_i^-$ or at least 50% of the points lie at the origin in direction e_i.

In the former case, since $S_i^+ \cup S_i^-$ has size $2\alpha n$, we know that $|S'| \le 2\alpha n/0.5 = 4\alpha n$. Moreover, by an averaging argument, there are at least $|S'|/4$ points in one of $S' \cap S_i^+$ or $S' \cap S_i^-$. Without loss of generality, we assume it is the $+$ side. By construction of U, the variance of $S' \cap S_i^+$ in the e_i direction is at least $\frac{1}{13}\frac{|S' \cap S_i^+|}{\alpha n} \ge 0.01\frac{|S'|}{\alpha n}$, where the first lower bound follows from having a sufficiently large αn. This in turn lower bounds the variance of S' in the e_i direction by $0.01\frac{|S'|}{\alpha n} \cdot |S' \cap S_i^+|/|S'| \ge 0.002\frac{|S'|}{\alpha n}$. The variance lower bound for S' required by the NLSC condition of [BKK22] is at most $\frac{1}{125}(\frac{|S'|}{n})^2 \cdot 2.1C^2 \le 0.07\alpha\frac{|S'|}{n} \cdot C^2$. Thus, as long as α is upper bounded by some constant much smaller than $1/C$, we will have $0.07\alpha\frac{|S'|}{n} \cdot C^2 \le 0.002\frac{|S'|}{\alpha n}$, and the NLSC condition of [BKK22] is satisfied in this case.

In the latter case, we know that there are at least $0.5|S'|$ points that project to the origin in dimension i, and we also showed previously that there are at least $2\alpha|S'|$ points in $S' \cap (S_i^+ \cup S_i^-)$. Further observe that points in $S' \cap (S_i^+ \cup S_i^-)$ have distance at least $C/\sqrt{\alpha} - \frac{1}{2}$ from the origin in the direction e_i. Using the formula that the variance of S' in direction e_i is equal to

$$\frac{1}{2|S'|^2}\sum_{x \in S'}\sum_{y \in S'}(x_i - y_i)^2,$$

we can thus lower bound this directional variance by

$$\frac{1}{2|S'|^2}2(0.5|S'|)(2\alpha|S'|) \cdot \left(C/\sqrt{\alpha} - \frac{1}{2}\right)^2 \ge 0.25C^2$$

whenever $C > 1$ and $\alpha < 1$. Finally, we note that $0.25C^2 \ge \frac{1}{125}\left(\frac{|S'|}{|S|}\right)^2 \cdot 2.1C^2$, where the right hand side is the NLSC condition (of [BKK22]) variance lower bound, meaning that the NLSC condition is also satisfied in this case.

To summarize, we have exhibited a set S such that, under the separation assumption of $C \cdot (\sigma_i + \sigma_j)/\sqrt{\alpha}$, the NLSC condition of [BKK22] still allows for two very different clusterings of the set S as long as we choose α sufficiently small as a function of the assumed constant C.

On the other hand, our stronger NLSC condition lets us prove Corollary 8.1 below, which shows that the algorithm will always output a clustering close to the ground truth. As such, there cannot be two substantially-different ground truth clusterings under our stronger assumption.

8.2 NLSC implies accurate clustering We prove Corollary 8.1, which states that, if we assume the NLSC condition (Section 1.1), then Algorithm 1 returns a clustering instead of just a refinement. That is, it returns exactly k sets. After that, we show that well-conditioned high-dimensional log-concave distributions give samples that satisfy the NLSC condition with high probability.

COROLLARY 8.1. *If in the setting of Theorem 3.2 we additionally assume that the sets S_i jointly satisfy the NLSC assumption with parameter α across all $i \in [k]$, then the algorithm returns exactly one sample set per mixture component. More precisely, for all $i \in [k]$, the set H_i mentioned in the statement of Theorem 3.2 is a singleton. As a consequence, if j is the unique index in H_i in the context of Theorem 3.2, then we have $\|\mu_{B_j} - \mu_i\| \le O(\sigma_i)$.*

Proof. [Proof of Corollary 8.1] To show this by contradiction, suppose that there are two output sets B, B' that correspond to the same cluster S_i, i.e. $B \subseteq \mathcal{B}_i$ and $B' \subseteq \mathcal{B}_i$ according to Item 2 of Theorem 3.1.

For the set B, observe that we have $|B \cap S_i| \ge 0.96|B| \ge 0.88\alpha|S_i|$. The first inequality is a consequence of Items 1 and 2b (see Remark 1.1), and the second inequality uses $|B| \ge 0.92\alpha n$ and $|S_i| \le n$.

By an application of Fact 2.3, and a subsequent usage of the NLSC assumption, we have that

$$\|\mu_{B\cap S_i} - \mu_{S_i}\|_2 \le \frac{\sigma_{S_i}}{\sqrt{0.88\alpha}} \le \frac{10\sigma_{B\cap S_i}}{\sqrt{0.88\alpha}} \le \frac{10\sigma_B}{0.96\sqrt{0.88\alpha}} \le \frac{12\sigma_B}{\sqrt{\alpha}}\,, \tag{8.13}$$

where the first inequality is an application of Fact 2.3 using $|B \cap S_i| \ge 0.88\alpha|S_i|$, the second inequality uses the NLSC assumption, and the third inequality uses the fact that $|B \cap S_i| \ge 0.96|B|$ implies $\sigma_{B\cap S_i} \le \sigma_B/0.96$. Moreover, since $|B \cap S_i| \ge 0.96|B|$, by another application of Fact 2.3,

$$\|\mu_{B\cap S_i} - \mu_B\|_2 \le \sigma_B/\sqrt{0.96}\,. \tag{8.14}$$

The above two inequalities together imply $\|\mu_B - \mu_{S_i}\| \le 13\sigma_B/\sqrt{\alpha}$. By symmetric arguments for B', we also have $\|\mu_{B'} - \mu_{S_i}\| \le 13\sigma_{B'}/\sqrt{\alpha}$. This then implies $\|\mu_B - \mu_{B'}\|_2 \le 13(\sigma_B + \sigma_{B'})/\sqrt{\alpha}$. We have thus contradicted Item 2e of the theorem (because the constant C there is $C > 1$). ∎

We now show that well-conditioned log-concave distributions yield samples that satisfy the no large sub-cluster condition with high probability. We first start with isotropic distributions.

PROPOSITION 8.1. *Consider an arbitrary d-dimensional isotropic log-concave distribution D. If $d \ge c \cdot \log^8 \frac{1}{\alpha}$ for some sufficiently large constant c, then it suffices to take a set S of $\tilde{O}((d + \log \frac{1}{\delta})/\alpha^2)$ samples from D so that, with probability at least $1 - \delta$, for any subset $S' \subseteq S$ with $|S'| \ge 0.8\alpha$, we have $\|\operatorname{Cov}(S')\|_{\mathrm{op}} \ge 0.7$.*

Proof. First observe that isotropic log-concave distributions D concentrate around a thin spherical shell. Specifically, a result of [Fle10] shows that

$$\mathop{\mathbb{P}}_{X\sim D}\left[\left(1 - \frac{t}{d^{1/8}}\right)\sqrt{d} \le \|X\|_2 \le \left(1 + \frac{t}{d^{1/8}}\right)\sqrt{d}\right] \ge 1 - O\left(e^{-\Omega(t)}\right)$$

for all $t \in [0, d^{1/8}]$. Taking $t = \Theta(\log \frac{1}{\alpha})$ and using the assumption that $d \gg \log^8 \frac{1}{\alpha}$, this implies that with probability at least $1 - \alpha/1000$, we have $\|X\|_2 \ge \sqrt{0.99d}$. Thus, by standard Chernoff bounds, if we take at least $O((1/\alpha)\log \frac{1}{\delta})$ many samples for some sufficiently large hidden constant, then with probability at least $1 - \delta/2$, at most an $\alpha/100$ fraction of the samples have $\|X\|_2 < \sqrt{0.99d}$.

We will further show that the following claim that with high probability over the entire sample set, any α-fraction of the samples must have mean not too far from the origin.

CLAIM 4. *Suppose S is a set of samples drawn from distribution D, of size at least a large constant multiple of $(d + \log \frac{1}{\delta})/\alpha^2$. Then, with probability at least $1 - \delta/2$ over the randomness of S, for any arbitrary subset $S' \subset S$ of size at least $0.9\alpha|S|$, we have $\|\mu_{S'}\|_2 \le O(\log \frac{1}{\alpha})$.*

Proof. We will use the standard fact that isotropic log-concave distributions are sub-exponential, whose samples are in turn stable with high probability, as long as the sample size is sufficiently large (see Exercise 3.1 in [DK23] for example). In particular, with probability at least $1 - \delta/2$ over a set S of $\tilde{O}((d + \log \frac{1}{\delta})/\alpha^2)$ samples from a

log-concave distribution D with unit covariance D and mean 0, it holds that for every subset $\tilde{S} \subseteq S$ of size at least $(1-\alpha)|S|$, we have $\|\mu_{\tilde{S}}\|_2 \le O(\alpha \log \frac{1}{\alpha})$.

Now consider any subset $S'' \subseteq S$ of size between $(\alpha/2)|S|$ and $\alpha|S|$. Its complement $\tilde{S} = S \setminus S''$ satisfies $\|\mu_{\tilde{S}}\|_2 \le O(\alpha \log \frac{1}{\alpha})$. Furthermore, by alternatively taking $\tilde{S} = S$, we have $\|\mu_S\|_2 = O(\alpha \log \frac{1}{\alpha})$. Thus, $\|\mu_{S''}\|_2 \le \frac{2}{\alpha}\|\mu_S - (1-\alpha)\mu_{\tilde{S}}\|_2 \le \frac{1}{\alpha} O(\alpha \log \frac{1}{\alpha}) = O(\log \frac{1}{\alpha})$ by the triangle inequality.

Finally, consider any subset S' of size at least $0.8\alpha|S|$. Observe that this set S' can always be partitioned into sets S'' of sizes between $(\alpha/2)|S|$ and $\alpha|S|$, each of which satisfies $\|\mu_{S''}\|_2 \le O(\log \frac{1}{\alpha})$. Moreover, the mean $\mu_{S'}$ of S' is just the convex combination of the means of these smaller disjoint subsets. This implies that $\|\mu_{S'}\|_2 \le O(\log \frac{1}{\alpha})$. ∎

To summarize, we have shown that, with probability at least $1-\delta$ over the randomness of the samples S, we have (a) at most an $\alpha/100$ fraction of the samples x have $\|x\|_2 < \sqrt{0.99d}$, and (b) for any subset $S' \subseteq S$ of size at least $0.9\alpha n \ge 0.9\alpha|S|$, $\|\mu_{S'}\|_2 \le O(\log \frac{1}{\alpha})$. We are now ready to show the NLSC condition for S conditioned on these two facts.

First, take any subset S' of size at least $0.8\alpha n$. By condition (a) above, there are at least $0.75|S'|$ many points $x \in S'$ with $\|x\|_2 \ge \sqrt{0.99d}$. Thus, we have

$$\operatorname{tr}\left(\frac{1}{|S'|}\sum_{x\in S'} xx^\top\right) \ge 0.75d\,.$$

Second, observe that the covariance of S' is

$$\operatorname{Cov}(S') = \frac{1}{|S'|}\sum_{x\in S'}(x-\mu_{S'})(x-\mu_{S'})^\top = \frac{1}{|S'|}\sum_{x\in S'} xx^\top - \mu_{S'}\mu_{S'}^\top\,.$$

Thus, we have

$$\operatorname{tr}(\operatorname{Cov}(S')) = \operatorname{tr}\left(\frac{1}{|S'|}\sum_{x\in S'} xx^\top\right) - \operatorname{tr}(\mu_{S'}\mu_{S'}^\top) \ge 0.75d - O\left(\log^2\frac{1}{\alpha}\right) \ge 0.7d\,, \tag{8.15}$$

where the last inequality uses $d \gg \log^8 \frac{1}{\alpha} \gg \log^2 \frac{1}{\alpha}$. Since the trace is equal to the sum of all eigenvalues, (8.15) states that the average eigenvalue is at least 0.7, thus the largest one should be $\|\operatorname{Cov}(S')\|_{\mathrm{op}} \ge 0.7$.

∎

Now we use the above proposition to show that samples from well-conditioned log-concave distributions satisfy Section 1.1 with high probability. In fact, the guarantees apply even to log-concave distributions for which there is a high-dimensional subspace V that both *i*) contains the largest variance direction and *ii*) is well-conditioned in the projection onto V.

PROPOSITION 8.2. *Consider an arbitrary d-dimensional log-concave distribution D with covariance matrix Σ such that there exists a subspace V of dimension $\dim(V) \gg c \cdot \log^8 \frac{1}{\alpha}$ which: (i) contains the top eigenvector of Σ and (ii) the covariance matrix of the projected distribution $\operatorname{Proj}_V \Sigma \operatorname{Proj}_V^\top$ has condition number at most 2. Then $\tilde{O}((d+\log\frac{1}{\delta})/\alpha^2)$ samples from D suffice for the sample set to satisfy the NLSC condition (Section 1.1) with probability at least $1-\delta$.*

Proof. We first show the special case when V is the entire $\mathbb{R}^d$, and by assumption, the condition number of Σ is $\kappa \le 2$. Observe that, by the property of the operator norm, for every matrix A it holds $\|A\|_{\mathrm{op}} = \|\Sigma^{-1/2}\Sigma^{1/2}A\Sigma^{1/2}\Sigma^{-1/2}\|_{\mathrm{op}} \le \|\Sigma^{-1/2}\|_{\mathrm{op}}^2 \|\Sigma^{1/2}A\Sigma^{1/2}\|_{\mathrm{op}}$, which rearranging gives

$$\|\Sigma^{1/2}A\Sigma^{1/2}\|_{\mathrm{op}} \ge \|A\|_{\mathrm{op}}/\|\Sigma^{-1/2}\|_{\mathrm{op}}^2\,. \tag{8.16}$$

Let S be the samples from D, and consider an arbitrary subset $S' \subset S$ with $|S'| \ge 0.8\alpha n$. Moreover, let the normalized versions of these sets be $\tilde{S} = \{\Sigma^{-1/2}x : x \in S\}$ and $\tilde{S}' = \{\Sigma^{-1/2}x : x \in S'\}$. The normalization means

that the samples in $\tilde{S}, \tilde{S}'$ come from an isotropic log-concave distribution. Thus, by Proposition 8.1 we know that with probability at least $1 - \delta/2$, it holds

$$\|\operatorname{Cov}(\tilde{S}')\|_{\mathrm{op}} \geq 0.7\,. \tag{8.17}$$

Putting everything together, we have

$$\begin{aligned}
\|\operatorname{Cov}(S')\|_{\mathrm{op}} &\geq \frac{1}{\|\Sigma^{-1/2}\|_{\mathrm{op}}^2}\|\operatorname{Cov}(\tilde{S}')\|_{\mathrm{op}} && \text{(using (8.16) with } A = \operatorname{Cov}(\tilde{S}'))\\
&\geq \frac{0.7}{\|\Sigma^{-1/2}\|_{\mathrm{op}}^2} && \text{(by (8.17))}\\
&\geq 0.35\|\Sigma\|_{\mathrm{op}} && \text{(since condition number of } \Sigma \text{ is at most } \kappa \leq 2)
\end{aligned}$$

Finally, we again note that isotropic log-concave distributions are sub-exponential and thus by standard arguments (see, e.g. Exercise 3.1 in [DK23]), $\tilde{O}((d + \log \frac{1}{\delta}))$ samples suffice to have that $\|\operatorname{Cov}(\tilde{S})\|_{\mathrm{op}} \leq 1.001$ with probability at least $1 - \delta/2$. This means that $\|\operatorname{Cov}(S)\|_{\mathrm{op}} = \|\Sigma^{1/2}\operatorname{Cov}(\tilde{S})\Sigma^{1/2}\|_{\mathrm{op}} \leq \|\Sigma\|_{\mathrm{op}}\|\operatorname{Cov}(\tilde{S})\|_{\mathrm{op}} \leq 1.001\|\Sigma\|_{\mathrm{op}}$. Combining this with the fact that $\|\operatorname{Cov}(S')\|_{\mathrm{op}} \geq 0.35\|\Sigma\|_{\mathrm{op}}$ (that we showed earlier), we obtain that $\|\operatorname{Cov}(S')\|_{\mathrm{op}} \geq 0.1\|\operatorname{Cov}(S)\|_{\mathrm{op}}$, i.e. that the NLSC condition holds.

It is easy to extend the argument for a general subspace V in the corollary statement. To see this, note that orthogonal projections preserve log-concavity, thus if we restrict everything to the subspace V, we could first show that NLSC holds in that subspace. That is, for any subset $S' \subseteq S$ with $|S'| \geq 0.8\alpha|S|$ of the data points, if $S_V := \{\mathrm{Proj}_V x : x \in S\}$ and $S'_V := \{\mathrm{Proj}_V x : x \in S'\}$ denote the projected versions of the sets onto V then $\sigma_{S'_V} \geq 0.1\sigma_{S_V}$. Then, the two inequalities $\sigma_{S'} \geq \sigma_{S'_V}$ and $\sigma_{S_V} \geq 0.99\sigma_S$ would imply that NLSC holds in the full-dimensional space. The first inequality is due to the fact that orthogonal projections can only decrease the variance, and the second inequality is because both σ_{S_V} and σ_S are with high probability close to $\sqrt{\|\Sigma\|_{\mathrm{op}}}$, by concentration of the empirical covariance matrix in every direction (Exercise 3.1 in [DK23]). ∎

Acknowledgements

I.D. is grateful to Ravi Kannan for numerous technical conversations during the Simons Institute program on the Computational Complexity of Statistical Inference. His insights served as an inspiration for this work.

References

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

[AM05] D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *Proceedings of the Eighteenth Annual Conference on Learning Theory (COLT)*, pages 458–469, 2005.

[AS12] P. Awasthi and O. Sheffet. Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 37–49. Springer, 2012.

[BBV08] M.-F. Balcan, A. Blum, and S. Vempala. A discriminative framework for clustering via similarity functions. In *STOC*, pages 671–680, 2008.

[BDH+20] A. Bakshi, I. Diakonikolas, S. B. Hopkins, D. Kane, S. Karmalkar, and P. K. Kothari. Outlier-robust clustering of gaussians and other non-spherical mixtures. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020*, pages 149–159. IEEE, 2020.

[BDJ+20] A. Bakshi, I. Diakonikolas, H. Jia, D. M. Kane, P. K. Kothari, and S. S. Vempala. Robustly learning mixtures of k arbitrary gaussians. *arXiv preprint arXiv:2012.02119*, 2020. Conference version in STOC'22.

[BKK22] C. Bhattacharyya, R. Kannan, and A. Kumar. How many clusters? - an algorithmic answer. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2607–2640, 2022.

[Bru09] S. C. Brubaker. *Extensions of Principle Components Analysis*. PhD thesis, Georgia Institute of Technology, 2009.

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

[CMY20] Y. Cherapanamjeri, S. Mohanty, and M. Yau. List decodable mean estimation in nearly linear time. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020*, 2020.

[CSV17] M. Charikar, J. Steinhardt, and G. Valiant. Learning from untrusted data. In *Proceedings of STOC 2017*, pages 47–60, 2017.

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

[DK20] I. Diakonikolas and D. M. Kane. Small covers for near-zero sets of polynomials and learning latent variable models. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020*, pages 184–195. IEEE, 2020.

[DK23] I. Diakonikolas and D. M. Kane. *Algorithmic high-dimensional robust statistics*. Cambridge University Press, 2023.

[DKK+16] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *Proceedings of FOCS'16*, pages 655–664, 2016.

[DKK20a] I. Diakonikolas, D. Kane, and D. Kongsgaard. List-decodable mean estimation via iterative multi-filtering. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[DKK+20b] I. Diakonikolas, D. M. Kane, D. Kongsgaard, J. Li, and K. Tian. List-decodable mean estimation in nearly-pca time. *CoRR*, abs/2011.09973, 2020. Conference version in NeurIPS'21.

[DKK+21] I. Diakonikolas, D. Kane, D. Kongsgaard, J. Li, and K. Tian. List-decodable mean estimation in nearly-pca time. *Advances in Neural Information Processing Systems*, 34:10195–10208, 2021.

[DKK+22a] I. Diakonikolas, D. Kane, S. Karmalkar, A. Pensia, and T. Pittas. List-decodable sparse mean estimation via difference-of-pairs filtering. In *NeurIPS*, 2022.

[DKK+22b] I. Diakonikolas, D. M. Kane, D. Kongsgaard, J. Li, and K. Tian. Clustering mixture models in almost-linear time via list-decodable mean estimation. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1262–1275, 2022.

[DKP20] I. Diakonikolas, D. M. Kane, and A. Pensia. Outlier Robust Mean Estimation with Subgaussian Rates via Stability. In *Advances in Neural Information Processing Systems 33, NeurIPS 2020*, 2020.

[DKS18a] I. Diakonikolas, D. M. Kane, and A. Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pages 1047–1060, 2018. Full version available at https://arxiv.org/abs/1711.07211.

[DKS18b] I. Diakonikolas, D. M. Kane, and A. Stewart. Sharp bounds for generalized uniformity testing. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, pages 6204–6213, 2018.

[Fle10] B. Fleury. Concentration in a thin euclidean shell for log-concave measures. *Journal of Functional Analysis*, 259(4):832–841, 2010.

[GEGMMI10] L. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. A review of robust clustering methods. *Advances in Data Analysis and Classification*, 4(2):89–109, 2010.

[HL18] S. B. Hopkins and J. Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pages 1021–1034, 2018.

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

[Kan21] D. M. Kane. Robust learning of mixtures of gaussians. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1246–1258. SIAM, 2021.

[KK10] A. Kumar and R. Kannan. Clustering with spectral norm and the k-means algorithm. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 299–308. IEEE, 2010.

[KSS18] P. K. Kothari, J. Steinhardt, and D. Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pages 1035–1046, 2018.

[KSV05] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *Proceedings of the Eighteenth Annual Conference on Learning Theory (COLT)*, pages 444–457, 2005.

[Lin95] B. Lindsay. *Mixture models: theory, geometry and applications*. Institute for Mathematical Statistics, 1995.

[LL22] A. Liu and J. Li. Clustering mixtures with almost optimal separation in polynomial time. In S. Leonardi and A. Gupta, editors, *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, 2022*, pages 1248–1261. ACM, 2022.

[LM20] A. Liu and A. Moitra. Settling the robust learnability of mixtures of gaussians. *arXiv preprint arXiv:2011.03622*, 2020. Conference version in STOC'21.

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

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

[TSM85] D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York, 1985.

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

[VW02] S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. In *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science*, pages 113–122, 2002.

Appendix

A Omitted Proofs from Section 2

We restate and prove the following statements.

LEMMA 2.1. *Let S be a set of points that is (C,ϵ)-stable with respect to μ and σ for some $C \geq 1$ and $\epsilon < 1/2$. Then, any subset $S' \subseteq S$ with $|S'| \geq \alpha|S|$ is $(1.23C/\sqrt{0.04\alpha}, 0.04)$-stable with respect to μ and σ.*

Proof. [Proof] Let S' be a subset of S with $|S'| \geq \alpha|S|$. According to Definition 2.1, in order to show that S' is $(1.23C/\sqrt{0.04\alpha}, 0.04)$-stable, we have to show that for any weight function $w : S' \to [0,1]$ with $\sum_{x\in S'} w_x \geq (1-0.04)|S'|$, the weighted mean and second moment centered around μ are at most $1.23C/\sqrt{\alpha}$ and $38C^2\sigma^2/\alpha$ respectively.

For the mean, by an application of Fact 2.3, we have that

$$\begin{aligned}\left\|\frac{\sum_{x\in S'} w_x x}{\sum_{x\in S'} w_x} - \mu\right\|_2 &\leq \left\|\frac{\sum_{x\in S'} w_x x}{\sum_{x\in S'} w_x} - \mu_S\right\|_2 + \|\mu_S - \mu\|_2 \\ &\leq \frac{\sigma_S}{\sqrt{(1-0.04)\alpha}} + C\sigma\sqrt{0.04} \\ &\leq \frac{C\sigma}{\sqrt{(1-0.04)\alpha}} + C\sigma\sqrt{0.04} \\ &\leq \frac{1.23C\sigma}{\sqrt{\alpha}}\,,\end{aligned}$$

where the first step is triangle inequality, the second step uses Fact 2.3 for the first term and stability of S for the second term, and the next step uses stability condition for the covariance.

For the second moment, we have that

$$\begin{aligned}\frac{1}{\sum_{x\in S'} w_x}\sum_{x\in S'} w_x(x-\mu)(x-\mu)^\top &\preceq \frac{1}{(1-0.04)\alpha}\frac{1}{|S|}\sum_{x\in S} w_x(x-\mu)(x-\mu)^\top \\ &\preceq \frac{1}{(1-0.04)\alpha}\frac{1}{|S|}\sum_{x\in S}(x-\mu)(x-\mu)^\top \\ &\preceq \frac{1}{(1-0.04)\alpha}C^2\sigma^2 I \preceq 38\frac{C^2\sigma^2}{\alpha}I\,,\end{aligned}$$

where the first step uses that $\sum_{x\in S'} w_x \geq (1-0.04)\alpha|S|$, and the last line uses stability for S. ∎

PROPOSITION 2.1. *Let T be a set of m points in $\mathbb{R}^d$. There is a $\mathrm{poly}(m,d)$-time algorithm that outputs a list of size $O(m^2\log(m))$ that for any $S \subseteq T$ contains an estimate $\hat{s}$ such that $\|\operatorname{Cov}(S)\|_{\mathrm{op}} \leq \hat{s}^2 \leq 2\|\operatorname{Cov}(S)\|_{\mathrm{op}}$.*

Proof. The algorithm is the following:

1. $L \leftarrow \emptyset$.
2. For every pair $x, y \in T$:
   (a) Add $\sqrt{2^{-j}\|x-y\|_2^2}$ to the list L for every $j = 0, 1, \ldots, \log(2m^2)$.
3. Let $L' \leftarrow \{\sqrt{2}s : s \in L\}$.
4. Return L'.

Using the definition of the covariance matrix $\operatorname{Cov}(S) = \frac{1}{2|S|^2}\sum_{x,y\in S}(x-y)(x-y)^\top$, we have that $\max_{x,y\in S}\|x-y\|_2^2/(2|S|^2) \leq \|\operatorname{Cov}(S)\|_{\mathrm{op}} \leq \max_{x,y\in S}\|x-y\|_2^2$. The algorithm adds to the output list every number starting from $\max_{x,y\in S}\|x-y\|_2$ down to $\max_{x,y\in S}\|x-y\|_2/\sqrt{2|S|^2}$ in factors of $\sqrt{2}$. This means that the list L contains an s such that $s^2 \leq \|\operatorname{Cov}(S)\|_{\mathrm{op}} \leq 2s^2$. By multiplying each element in the list by $\sqrt{2}$, the final list L' contains an $\hat{s}$ such that $\|\operatorname{Cov}(S)\|_{\mathrm{op}} \leq \hat{s}^2 \leq 2\|\operatorname{Cov}(S)\|_{\mathrm{op}}$. ∎