Clustering Semi-Random Mixtures of Gaussians

Pranjal Awasthi*
pranjal.awasthi@rutgers.edu

Aravindan Vijayaraghavan†
aravindv@northwestern.edu

Abstract

Gaussian mixture models (GMM) are the most widely used statistical model for the k-means clustering problem and form a popular framework for clustering in machine learning and data analysis. In this paper, we propose a natural semi-random model for k-means clustering that generalizes the Gaussian mixture model, and that we believe will be useful in identifying robust algorithms. In our model, a semi-random adversary is allowed to make arbitrary "monotone" or helpful changes to the data generated from the Gaussian mixture model.

Our first contribution is a polynomial time algorithm that provably recovers the ground-truth up to small classification error w.h.p., assuming certain separation between the components. Perhaps surprisingly, the algorithm we analyze is the popular Lloyd's algorithm for k-means clustering that is the method-of-choice in practice. Our second result complements the upper bound by giving a nearly matching information-theoretic lower bound on the number of misclassified points incurred by any k-means clustering algorithm on the semi-random model.

1 Introduction

Clustering is a ubiquitous task in machine learning and data mining for partitioning a data set into groups of similar points. The k-means clustering problem is arguably the most well-studied problem in machine learning. However, designing provably optimal k-means clustering algorithms is a challenging task as the k-means clustering objective is NP-hard to optimize [WS11] (in fact, it is also NP-hard to find near-optimal solutions [ACKS15, LSW17]). A popular approach to cope with this intractability is to study average-case models for the k-means problem. The most widely used such statistical model for clustering is the *Gaussian Mixture Model (GMM)*, that has a long and rich history [Tei61, Pea94, Das99, AK01, VW04, DS07, BV08, MV10, BS10, KK10].

In this model there are k clusters, and the points from cluster i are generated from a Gaussian in d dimensions with mean $\mu_i \in \mathbb{R}^d$, and covariance matrix $\Sigma_i \in \mathbb{R}^{d\times d}$ with spectral norm $\|\Sigma_i\| \leq \sigma^2$. Each of the N points in the instance is now generated independently at random, and is drawn from the ith component with probability $w_i \in [0,1]$ ($w_1, w_2, \ldots, w_k$ are also called mixing weights). If the means of the underlying Gaussians are separated enough, the ground truth clustering is well defined[1]. The algorithmic task is to recover the ground truth clustering for any data set generated from such a model (note that the parameters of the Gaussians, mixing weights and the cluster memberships of the points are unknown).

Starting from the seminal work of Dasgupta [Das99], there have been a variety of algorithms to provably cluster data from a GMM model. Algorithms based on PCA and

*Department of Computer Science, Rutgers University.

†Department of Electrical Engineering and Computer Science, Northwestern University. Supported by the National Science Foundation (NSF) under Grant No. CCF-1652491 and CCF-1637585.

[1] A separation of $\|\mu_i - \mu_j\|_2 \geq \Omega(\sigma\sqrt{\log(Nk)})$ for $i \neq j \in [k]$ suffices w.h.p.

distance-based clustering [AK01, VW04, AM05, KSV08] provably recover the clustering when there is adequate separation between every pair of components (parameters). Other algorithmic approaches include the method-of-moments [KMV10, MV10, BS10], and algebraic methods based on tensor decompositions [HK12, GVX14, BCMV14, ABG$^+$14, GHK15]. (Please see Section 1.1 for a more detailed comparison of the guarantees).

On the other hand, the method-of-choice in practice are iterative algorithms like the Lloyd's algorithm (also called k-means algorithm) [Llo82] and the k-means++ algorithm of [AV07] (Lloyd's algorithm initialized with centers from distance-based sampling). In the absence of good worst-case guarantees, a compelling direction is to use *beyond-worst-case* paradigms like average-case analysis to provide provable guarantees. Polynomial time guarantees for recovering k-means optimal clustering by the Lloyd's algorithm and k-means++ are known when the points are drawn from a GMM model under sufficient separation conditions [DS07, KK10, AS12].

Although the study of Gaussian mixture models has been very fruitful in designing a variety of efficient algorithms, real world data rarely satisfies such strong distributional assumptions. Hence, our choice of algorithm should be informed not only by its computational efficiency but also by its robustness to errors and model misspecification. As a first step, we need theoretical frameworks that can distinguish between algorithms that are tailored towards a specific probabilistic model and algorithms robust to modeling assumptions. In this paper we initiate such a study in the context of clustering, by studying a natural *semi-random model* that generalizes the GMM model and also captures robustness to certain adversarial dependencies in the data.

Semi-random models involve a set of adversarial choices in addition to the random choices of the probabilistic model, while generating the instance. These models have been successfully applied to study the design of robust algorithms for various optimization problems [BS95, FK98, MS10, KMM11, MMV12, MMV14] (see Section 1.1) In a typical semi-random model, there is a "planted" or "ground-truth" solution, and an instance is first generated according to a simple probabilistic model. An adversary is then allowed to make "monotone" or helpful changes to the instance that only make the planted solution more pronounced. For instance, in the semi-random model of Feige and Kilian [FK98] for graph partitioning, the adversary is allowed to arbitrarily add extra edges within each cluster or delete edges between different clusters of the planted partitioning. These adversarial choices only make the planted partition more prominent; however, the choices can be dependent and thwart algorithms that rely on the excessive independence or strong but unrealistic structural properties of these instances.

Hence, the study of semi-random models helps us understand and identify robust algorithms. Our motivation for studying semi-random models for clustering is two-fold: a) design algorithms that are robust to strong distributional data assumptions, and b) explain the empirical success of simple heuristics such as the Lloyd's algorithm.

Semi-random mixtures of Gaussians In an ideal clustering instance, each point x in the ith cluster is significantly closer to the mean μ_i than to any other mean μ_j for $j \neq i$ (for a general instance, in the optimal solution, $\|x - \mu_i\|_2 - \|x - \mu_j\|_2 \leq 0 \ \forall j \neq i$). Moving each point in C_i toward its own mean μ_i only increases this gap between the distance to its mean and to any other mean. Hence, this perturbation corresponds to a monotone perturbation that only make this planted clustering even better. In our semi-random model, the points are first drawn from a mixture of Gaussians (this is the planted clustering). The adversary is then allowed to move each point in the ith cluster closer to its mean μ_i. This allows the points to be even better clustered around their respective means, however these perturbations are allowed to have arbitrary dependencies. We now formally define the semi-random model.

Definition 1.1 (Semi-random GMM model)**.** Given a set of parameters $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}^+$, a clustering instance $\mathcal{X}$ on N points is generated as follows.

1. Adversary chooses an arbitrary partition $\mathcal{C} = (C_1, C_2, \ldots, C_k)$ of $\{1, \ldots, N\}$ and let $N_i = |C_i|$ for all $i \in [k]$.

2. For each $i \in [k]$ and each $t \in C_i$, $y^{(t)} \in \mathbb{R}^d$ is generated independently at random according to a Gaussian with mean μ_i and covariance Σ_i with $\|\Sigma\| \leq \sigma$ i.e., variance at most σ^2 in each direction.

3. The adversary then moves each point $y^{(t)}$ towards the mean of its component by an arbitrary amount i.e., for each $i \in [k], t \in C_i$, the adversary picks $x^{(t)}$ arbitrarily in $\{\mu_i + \lambda(y^{(t)} - \mu_i) : \lambda \in [0, 1]\}$. (Note that these choices can be correlated arbitrarily.)

The instance is $\mathcal{X} = \{x^{(t)} : t \in [N]\}$ and is parameterized by $(\mu_1, \ldots, \mu_k, \sigma)$ with the planted clustering $C_1, \ldots, C_k$. We will denote by $w_{\min} = \min_{i \in [k]} N_i/N$.

Data generated by mixtures of high-dimensional Gaussians have certain properties that are often not exhibited by real-world instances. High-dimensional Gaussians have strong concentration properties; for example, all the points generated from a high-dimensional Gaussian are concentrated at a reasonably far distance from the mean (they are $\approx \sqrt{d}\sigma$ far away w.h.p.). In many real-world datasets on the other hand, clusters in the ground-truth often contain dense "cores" that are close to the mean. Our semi-random model admits such instances by allowing points in a cluster to move arbitrarily close to the mean.

Our Results. Our first result studies the Lloyd's algorithm on the semi-random GMM model and gives an upper bound on the clustering error achieved by the Lloyd's algorithm with the initialization procedure used in [KK10].

Informal Theorem 1.2. *Consider any semi-random instance $\mathcal{X}$ with N points generated by the semi-random GMM model (Definition 1.1) with planted clustering $C_1, \ldots, C_k$ and parameters $\mu_1, \ldots, \mu_k, \sigma^2$ satisfying*

$$\forall i \neq j \in [k],\ \|\mu_i - \mu_j\|_2 > \Delta\sigma, \quad \textit{where } \Delta \geq c_0\sqrt{\min\{k, d\} \log N},$$

and $N \geq k^2 d^2 / w_{min}^2$. There is polynomial time algorithm based on the Lloyd's iterative algorithm that recovers the cluster memberships of all but $\tilde{O}(kd/\Delta^4)$ points.

The $\tilde{O}$ in the above statement hides a $\log(\log N/\Delta^4)$ and $\log(d/\Delta^2)$ factor. Please see Theorem 3.1 for a formal statement. Furthermore, we show that in the above result the Lloyd's iterations can be initialized using the popular k-mean++ algorithm that uses D^2-sampling [AV07]. The most closely related to our work is that of [KK10] and [AS12] who provided deterministic data conditions under which the Lloyd's algorithm converges to the optimal clustering. Along these lines, our work provides further theoretical justification for the enormous empirical success that the Lloyd's algorithm enjoys.

It is also worth noting that in spite of being robust to semi-random perturbations, the separation requirement of $\sigma\sqrt{k \log N}$ in our upper bound matches the separation requirement in the best guarantees [AS12] for Lloyd's algorithm even in the absence of any semi-random errors or perturbations (see Section 1.1 for a comparison) [2]. We also remark

[2] We note that for clustering GMMs, the work of Brubaker and Vempala [BV08] give a qualitatively different separation condition that does not depend on the maximum variance, and can model Gaussian mixtures that look like "parallel pancakes". However this separation condition is incomparable to [AS12], because of the potentially worse dependence on k.

that while the algorithm recovers a clustering of the given data that is very close to the planted clustering, this does not necessarily estimate the means of the original Gaussian components up to inverse polynomial accuracy (in fact the centers of the planted clustering after the semi-random perturbation may be $\Omega(\sigma)$ far from the original means). This differs from the recent body of work on parameter estimation in the presence of some adversarial noise (please refer to Section 1.1 for a comparison).

While the monotone changes allowed in the semi-random model should only make the clustering task easier, our next result shows that the error achieved by the Lloyd's algorithm is in fact near optimal. More specifically, we provide a lower bound on the number of points that will be misclassified by any k-means optimal solution for the instance.

Informal Theorem 1.3. *Given any N (that is sufficiently large polynomial in d, k) and Δ such that $\sqrt{\log N} \leq \Delta \leq d/(4 \log d)$, there exists an instance $\mathcal{X}$ on N points in d dimensions generated from the semi-random GMM model 1.1 with parameters $\mu_1, \ldots, \mu_k, \sigma^2$, and planted clustering $C_1, \ldots, C_k$ having separation $\forall i \neq j \in [k], \|\mu_i - \mu_j\|_2 \geq \Delta\sigma$ s.t. any optimal k-means clustering solution $C'_1, C'_2, \ldots, C'_k$ of $\mathcal{X}$ misclassifies at least $\Omega(kd/\Delta^4)$ points with high probability.*

The above lower bound also holds when the semi-random (monotone) perturbations are applied to points generated from a mixture of k spherical Gaussians each with covariance $\sigma^2 I$ and weight $1/k$. Further, the lower bound holds not just for the optimal k-means solution, but also for any "locally optimal" clustering solution. Please see Theorem 4.1 for a formal statement. These two results together show that the Lloyd's algorithm essentially recovers the planted clustering up to the optimal error possible for any k-means clustering based algorithm.

Unlike algorithmic results for other semi-random models, an appealing aspect of our algorithmic result is that it gives provable robust guarantees in the semi-random model for a simple, popular algorithm that is used in practice (Lloyd's algorithm). Further, other approaches for clustering like distance-based clustering, method-of-moments and tensor decompositions seem inherently non-robust to these semi-random perturbations (see Section 1.1 for details). This robustness of the Lloyd's algorithm suggests an explanation for its widely documented empirical success across different application domains.

Considerations in the choice of the Semi-random GMM model. Here we briefly discuss different semi-random models, and considerations involved in favoring Definition 1.1. Another semi-random model that comes to mind is one that can move each point closer to the mean of its own cluster (closer just in terms of distance, regardless of direction). Intuitively this seems appealing since this improves the cost of the planted clustering. However, in this model the optimal k-means clustering of the perturbed instance can be vastly different from the planted solution. This is because one can move many points x in cluster C_i in such a way that x becomes closer to a different mean rather than μ_i. For high dimensional Gaussians it is easy to see that the distance of each point to its own mean will be on the order of $(\sqrt{d} + 2\sqrt{\log N})\sigma$. Hence, in our regime of interest, the inter mean separation of $\sqrt{k \log N}\sigma$ could be much smaller than the radius of any cluster (when $d \gg k$). Consider an adversary that moves a large fraction of the points in a given cluster to the mean of another cluster. While the distance of these points to their cluster mean has only decreased from roughly $(\sqrt{d} + 2\sqrt{\log N})\sigma$ to around $\sqrt{k \log N}\sigma$, these points now become closer to the mean of a different cluster! In the semi-random GMM model on the other hand, the adversary is only allowed to move the point x along the direction of $x - \mu$; hence, each point x becomes closer to its own mean than to the means of other clusters. Our results show that in such a model, the optimal clustering solution can change by at most $\tilde{O}(d/\Delta^4)$ points.

Challenges in the Semi-random GMM model and Overview of Techniques. Lloyd's algorithm has been analyzed in the context of clustering mixtures of Gaussians [KK10, AS12]. Any variant of the Lloyd's algorithm consists of two steps — an initialization stage where a set of k initial centers are computed, and the iterative algorithm which successively improves the clustering in each step. Kumar and Kannan [KK10] considered a variant of the Lloyd's method where the initialization is given by using PCA along with a $O(1)$ factor approximation to the k-means optimization problem. The improved analysis of this algorithm in [AS12] leads to state of the art results that perfectly recovers all the clusters when the separation is of the order $\sqrt{k \log N}\sigma$.

We analyze the variant of Lloyd's algorithm that was introduced by Kumar and Kannan [KK10]. However, there are several challenges in extending the analysis of [AS12] to the semi-random setting. While the semi-random perturbations in the model only move points in a cluster C_i closer to the mean μ_i, these perturbations can be co-ordinated in a way that can move the empirical mean of the cluster significantly. For instance, Lemma 4.3 gives a simple semi-random perturbation to the points in C_i that moves the empirical mean of the points in C_i to $\tilde{\mu}_i$ s.t. $\tilde{\mu}_i \approx \mu_i + \Omega(\sigma)\widehat{e}$, for any desired direction $\widehat{e}$. This shift in the empirical means may now cause some of the points in cluster C_i to become closer to $\tilde{\mu}_j$ (in particular points that have a relatively large projection onto $\widehat{e}$) and vice-versa. In fact, the lower bound instance in Theorem 4.1 is constructed by applying such a semi-random perturbation given by Lemma 4.3 to the points in a cluster, along a carefully picked direction so that $m = \Omega(d/\Delta^4)$ points are misclassified per cluster.

The main algorithmic contribution of the paper is an analysis of the Lloyd's iterative algorithm when the points come from the semi-random GMM model. The key is to understand the number of points that can be misclassified in an intermediate step of the Lloyd's iteration. We show in Lemma 3.3 that if in the current iteration of the Lloyd's algorithm, each of the current estimates of the means μ_i' is within $\tau\sigma$ from μ_i, then the number of misclassified points by the current iteration of Lloyd's iteration is at most $\tilde{O}(kd\tau^2/\Delta^4)$. This relies crucially on Lemma 2.11 which upper bounds the number of points x in a cluster C_i s.t. $(x - \mu_i)$ has a large inner product along any (potentially bad) direction $\widehat{e}$.

The effect of these bad points has to be carefully accounted for when analyzing both stages of the algorithm – the initialization phase, and the iterative algorithm. Proposition 3.2 argues about the closeness of the initial centers to the true means. As in [KK10], these initial centers are obtained via a boosting technique that first maps the points to an expanded feature space and then uses the (k-SVD + k-means approximation) to get initial centers. When using this approach for semi-random data one needs to carefully argue about how the set of bad points behave in the expanded feature space. This is done in Lemmas B.1 and B.2. Given the initial centers, it is not hard to see that the analysis of [AS12] can be carried out to argue about the improvements made in the Lloyd's iterative step; however, this leads to a bound that is sub-optimal by the factor of $O(k^2)$. Instead, we perform a much finer analysis for the semi-random model to control the effect of the bad points and achieve nearly optimal error bounds. This is done in Lemma 3.4.

1.1 Related Work

There has been a long line of algorithmic results on Gaussian mixture models starting from [Tei61, Tei67, Pea94]. These results fall into two broad categories: *(1) Clustering algorithms*, which aim to recover the component/cluster memberships of the points and *(2) Parameter estimation*, where the goal is to estimate the parameters of the Gaussian components. When the components of the mixture are sufficiently well-separated, i.e., $\|\mu_i - \mu_j\|_2 \geq \sigma\sqrt{\log(Nk)}$, then the Gaussians do not overlap w.h.p., and then the two tasks become equivalent w.h.p. We now review the different algorithms that have been designed

for these two tasks, and comment on their robustness to semi-random perturbations.

Clustering Algorithms. The first polynomial time algorithmic guarantees were given by Dasgupta [Das99], who showed how to cluster a mixture of k Gaussians with identical covariance matrices when the separation between the cluster means is of the order $\Omega(\sigma\sqrt{d}\text{polylog}(N))$, where σ denotes the maximum variance of any cluster along any direction[3]. Distance-based clustering algorithms that are based on strong distance-concentration properties of high-dimensional Gaussians improved the separation requirement between means μ_i and μ_j to be $\Omega(d^{1/4}\text{polylog}(N))(\sigma_i + \sigma_j)$ [AK01, DS07], where σ_i denotes the maximum variance of points in cluster i along any direction. Vempala and Wang [VW04] and subsequent results [KSV08, AM05] used PCA to project down to k dimensions (when $k \leq d$), and then used the above distance-based algorithms to get state-of-the-art guarantees for many settings: for spherical Gaussians a separation of roughly $\|\mu_i - \mu_j\|_2 \geq (\sigma_i + \sigma_j)\min\{k, d\}^{1/4}\text{polylog}(N)$ suffices [VW04]. For non-spherical Gaussians, a separation of $\|\mu_i - \mu_j\|_2 \geq (\sigma_i + \sigma_j)k^{3/2}\sqrt{\log N}$ is known to suffice [AM05, KSV08]. Brubaker and Vempala [BV08] gave a qualitative improvement on the separation requirement for non-spherical Gaussians by having a dependence only on the variance along the direction of the line joining the respective means, as opposed to the maximum variance along any direction.

Recent work has also focused on provable guarantees for heuristics such as the Lloyd's algorithm for clustering mixtures of Gaussians [KK10, AS12]. Iterative algorithms like the Lloyd's algorithm (also called k-means algorithm) [Llo82] and its variants like k-means++ [AV07] are the method-of-choice for clustering in practice. The best known guarantee [AS12] along these lines requires a separation of order $\sigma\sqrt{k\log N}$ between any pair of means, where σ is the maximum variance among all clusters along any direction. To summarize, for a mixture of k Gaussians in d dimensions with variance of each cluster being bounded by σ^2 in every direction, the state-of-the-art guarantees require a separation of roughly $\sigma\min\{k, d\}^{1/4}\text{polylog}(N)$ between the means of any two components [VW04] for spherical Gaussians, while a separation of $\sigma\sqrt{\min\{k, d\}\log N}$ is known to suffice for non-spherical Gaussians [AS12].

The techniques in many of the above works rely on strong distance concentration properties of high-dimensional Gaussians. For instance, the arguments of [AK01, VW04] that obtain a separation of order $\min\{k^{1/4}, d^{1/4}\}$ crucially rely on the tight concentration of the squared distance around $\sigma^2(d \pm c\sqrt{d})$, between any pair of points in the same cluster. These arguments do not seem to carry over to the semi-random model. Brubaker [Bru09] gave a robust algorithm for clustering a mixture of Gaussians when at most $o(1/k)$ fraction of the points are corrupted arbitrarily. However, it is unclear if the arguments can be modified to work under the semi-random model, since the perturbations can potentially affect all the points in the instance. On the other hand, our results show that the Lloyd's algorithm of Kumar and Kannan [KK10] is robust to these semi-random perturbations.

Parameter Estimation. A different approach is to design algorithms that estimate the parameters of the underlying Gaussian mixture model, and then assuming the means are well separated, accurate clustering can be performed. A very influential line of work focuses on the method-of-moments [KMV10, MV10, BS10] to learn the parameters of the model when the number of clusters $k = O(1)$. Moment methods (necessarily) require running time (and sample complexity) of roughly $d^{O(k^2)}$, but do not assume any explicit separation between the components of the mixture. Recent work [HK13, BCV14, GVX14, BCMV14, ABG+14, GHK15] uses uniqueness of tensor decompositions (of order 3 and

[3]The polylog(N) term involves a dependence of either $(\log N)^{1/4}$ or $(\log N)^{1/2}$.

above) to implement the method of moments and give polynomial time algorithms assuming the means are sufficiently high dimensional, and do not lie in certain degenerate configurations [HK12, GVX14, BCMV14, ABG$^+$14, GHK15].

Algorithmic approaches based on method-of-moments and tensor decompositions rely heavily on the exact parametric form of the Gaussian distribution and the exact algebraic expressions to express various moments of the distribution in terms of the parameters. These algebraic methods can be easily foiled by a monotone adversary, since the adversary can perturb any subset to alter the moments significantly (for example, even the first moment, i.e., the mean of a cluster, can change by $\Omega(\sigma)$).

Recent work has also focused on provable guarantees for heuristics such as Maximum Likelihood estimation and the Expectation Maximization (EM) algorithm for parameter estimation [DS07, BWY14, XHM16, DTZ16]. Very recently, [RV17] considered other iterative algorithms for parameter estimation, and studied the optimal order of separation required for parameter estimation. However, we are not aware of any existing analysis that shows that these iterative algorithms for parameter estimation are robust to modeling errors.

Another recent line of exciting work concerns designing robust high-dimensional estimators of the mean and covariance of a single Gaussian (and mixtures of k Gaussians) when an $\varepsilon = \Omega_k(1)$ fraction of the points are adversarially corrupted [DKK$^+$16, LRV16, CSV17]. However, these results and similar results on agnostic learning do not necessarily recover the ground-truth clustering. Further, they typically assume that only a $o(1/k)$ fraction of the points are corrupted, while potentially all the points could be perturbed in the semi-random model. On the other hand, our work does not necessarily give guarantees for estimating the means of the original Gaussians (in fact the centers given by the planted clustering in the semi-random instance can be $\Omega(\sigma)$ far from the original means). Hence, our semi-random model is incomparable to the model of robustness considered in these works.

Semi-random models for other optimization problems. There has been a long line of work on the study of semi-random models for various optimization problems. Blum and Spencer [BS95] initiated the study of semi-random models, and studied the problem of graph coloring. Feige and Kilian [FK98] considered semi-random models involving monotone adversaries for various problems including graph partitioning, independent set and clique. Makarychev et al. [MMV12, MMV14] designed algorithms for more general semi-random models for various graph partitioning problems. The work of [MPW15] studied the power of monotone adversaries in the context of community detection (stochastic block models), while [MMV16] considered the robustness of community detection to monotone adversaries and different kinds of errors and model misspecification. Semi-random models have also been studied for correlation clustering [MS10, MMV15], noisy sorting [MMV13] and coloring [DF16].

2 Preliminaries and Semi-random model

We first formally define the Gaussian mixture model.

Definition 2.1. (Gaussian Mixture Model). A Gaussian mixture model with k components is defined by the parameters $(\mu_1, \mu_2, \ldots \mu_k, \Sigma_1, \ldots, \Sigma_k, w_1, \ldots, w_k)$. Here $\mu_i \in \mathbb{R}^d$ is the mean for component i and $\Sigma_i \in \mathbb{S}^d_+$ is the corresponding $d \times d$ covariance matrix. $w_i \in [0, 1]$ is the mixing weight and we have that $\sum_{i=1}^k w_i = 1$. An instance $\mathcal{X} = \{ x^{(1)}, \ldots, x^{(N)} \}$ from the mixture is generated as follows: for each $t \in [N]$, sample a component $i \in [k]$ independently at random with probability w_i. Given the component,

sample $x^{(t)}$ from $\mathcal{N}(\mu_i, \Sigma_i)$. The N points can be naturally partitioned into k clusters $C_1, \ldots, C_k$ where cluster C_i corresponds to the points that are sampled from component i. We will refer to this as the *planted clustering* or *ground truth clustering*.

Clustering data from a mixture of Gaussians is a natural average-case model for the k-means clustering problem. Specifically, if the means of a Gaussian mixture model are well separated, then with high probability, the ground truth clustering of an instance sampled from the model corresponds to the k-means optimal clustering.

Definition 2.2. (k-means clustering). Given an instance $\mathcal{X} = \{ x^{(1)}, \ldots, x^{(N)} \}$ of N points in $\mathbb{R}^d$, the k-means problems is to find k points $\mu_1, \ldots, \mu_k$ such as to minimize $\sum_{t \in [N]} \min_{i \in [k]} \|x^{(t)} - \mu_i\|^2$.

The optimal means or centers $\mu_1, \ldots, \mu_k$ naturally define a clustering of the data where each point is assigned to its closest cluster. A key property of the k-means objective is that the optimal solution induces a locally optimal clustering.

Definition 2.3. (Locally Optimal Clustering). A clustering $C_1, \ldots, C_k$ of N data points in $\mathbb{R}^d$ is locally optimal if for each $i \in [k], x^{(t)} \in C_i$, and $j \neq i$ we have that $\|x^{(t)} - \mu(C_i)\| \leq \|x^{(t)} - \mu_j\|$. Here $\mu(C_i)$ is the average of the points in C_i.

Hence, given the optimal k-means clustering, the optimal centers can be recovered by simply computing the average of each cluster. This is the underlying principle behind the popular Lloyd's algorithm [Llo82] for k-means clustering. The algorithm starts with a choice of initial centers. It then repeatedly computes new centers to be the average of the clusters induced by the current centers. Hence the algorithm converges to a locally optimal clustering. Although popular in practice, the worst case performance of Lloyd's algorithm can be arbitrarily bad [AV05]. The choice of initial centers is very important in the success of the Lloyd's algorithm. We show that our theoretical guarantees hold when the initialization is done via the popular k-means++ algorithm [AV07]. There also exist more sophisticated constant factor approximation algorithms for the k-means problem [KMN$^+$02, ANSW16] that can be used for seeding in our framework.

While the clustering $C_1, C_2, \ldots, C_k$ typically represents a partition of the index set $[N]$, we will sometimes abuse notation and use C_i to also denote the set of points in $\mathcal{X}$ that correspond to these indices in C_i. Finally, many of the statements are probabilistic in nature depending on the randomness in the semi-random model. In the following section, w.h.p. will refer to a probability of at least $1 - o(1)$ (say $1 - 1/\text{poly}(N)$), unless specified otherwise.

2.1 Properties of Semi-random Gaussians

In this section we state and prove properties of semi-random mixtures that will be used throughout the analysis in the subsequent sections. We first start with a couple of simple lemmas that follow directly from the corresponding lemmas about high dimensional Gaussians.

Lemma 2.4. *Consider any semi-random instance $\mathcal{X} = \{ x^{(1)}, \ldots, x^{(N)} \}$ with parameters $\mu_1, \ldots, \mu_k, \sigma^2$ and clusters $C_1, \ldots, C_k$. Then with high probability we have*

$$\forall i \in [k],\ \forall \ell \in C_i, \quad \|x^{(t)} - \mu_i\|_2 \leq \sigma(\sqrt{d} + 2\sqrt{\log N}). \tag{1}$$

Proof. Let $y^{(t)}$ denote the point generated in the semi-random model in step 2 (Definition 1.1) before the semi-random perturbation was applied. Let $\bar{x}^{(t)} = x - \mu_i$, $\bar{y}^{(t)} = y - \mu_i$ where $t \in C_i$. We have

$$\forall i \in [k], \forall t \in C_i, \quad \|\bar{x}^{(t)}\|_2 \leq \|\bar{y}^{(t)}\|_2 \leq \sigma(\sqrt{d} + 2\sqrt{\log N}),$$

from Lemma A.3. ∎

Lemma 2.5. *Consider any semi-random instance $\mathcal{X} = \{ x^{(1)}, \ldots, x^{(N)} \}$ with parameters $\mu_1, \ldots, \mu_k, \sigma^2$ and clusters $C_1, \ldots, C_k$, and let u be a fixed unit vector in $\mathbb{R}^d$. Then with probability at least $(1 - 1/(N^3))$ we have*

$$\forall i \in [k], t \in C_i, \quad |\langle x^{(\ell)} - \mu_i, u \rangle| < 3\sigma\sqrt{\log N}. \tag{2}$$

Proof. Let $y^{(t)}$ denote the point generated in the semi-random model in step 2 (Definition 1.1) before the semi-random perturbation was applied. Let $\bar{x}^{(t)} = x - \mu_i$, $\bar{y}^{(t)} = y - \mu_i$ where $t \in C_i$.

Consider the sample $t \in C_i$. Let Σ_i be the covariance matrix of ith Gaussian component; hence $\|\Sigma_i\| \leq \sigma$. The projection $\langle \bar{y}^{(t)}, u \rangle$ is a Gaussian with mean 0 and variance $u^T \Sigma_i u \leq \sigma^2$. From Lemma A.1

$$\mathbb{P}\left[|\langle \bar{x}^{(t)}, u \rangle| \geq 3\sigma\sqrt{\log N} \right] \leq \mathbb{P}\left[|\langle \bar{y}^{(t)}, u \rangle| \geq 3\sigma\sqrt{\log N} \right] \leq \exp(-4 \log N) \leq N^{-4}.$$

Hence from a union bound over all N samples, the lemma follows. □

The above lemma immediately implies the following lemma after a union bound over the $k^2 < N^2$ directions given by the unit vectors along $(\mu_i - \mu_j)$ directions.

Lemma 2.6. *Consider any semi-random instance $\mathcal{X} = \{ x^{(1)}, \ldots, x^{(N)} \}$ with parameters $\mu_1, \ldots, \mu_k, \sigma^2$ and clusters $C_1, \ldots, C_k$. Then with high probability we have*

$$\forall i \in [k], t \in C_i, \quad \left| \left\langle x^{(\ell)} - \mu_i, \frac{\mu_i - \mu_j}{\|\mu_i - \mu_j\|_2} \right\rangle \right| < 3\sigma\sqrt{\log N}. \tag{3}$$

We next state a lemma about how far the mean of the points in a component of a semi-random GMM can move away from the true parameters.

Lemma 2.7. *Consider any semi-random instance $\mathcal{X}$ with N points generated with parameters $\mu_1, \ldots, \mu_k, C_1, \ldots, C_k$ such that $N_i \geq 4(d + \log(\frac{k}{\delta}))$ for all $i \in [k]$. Then with probability at least $1 - \delta$ we have that*

$$\forall i \in [k], \| \frac{1}{|C_i|} \sum_{x \in C_i} x - \mu_i \|_2 \leq 2\sigma. \tag{4}$$

Proof. For each point $x \in C_i$ in the semi-random GMM, let y_x be the original point in the GMM that is modified to produce x. Then, we know that $x - \mu_i = \lambda_x (y_x - \mu_i)$ where $\lambda_x \in [0, 1]$. Hence, $\frac{1}{|C_i|} \sum_{x \in C_i} (x - \mu_i) = \frac{1}{|C_i|} A_i D v$, where A_i is the matrix with columns as $(y_x - \mu_i)$ for $x \in C_i$, D is a diagonal matrix with values λ_x, and v is a unit length vector in the direction of $\frac{1}{|C_i|} \sum_{x \in C_i} (x - \mu_i)$. Then, we have that $\| \frac{1}{|C_i|} \sum_{x \in C_i} x - \mu_i \| = \| \frac{1}{|C_i|} A_i D v \| \leq \frac{1}{|C_i|} \|A\| \leq 2\sigma$ (from A.5). □

The next lemma argues about the variance of component i around μ_i in a semi-random GMM.

Lemma 2.8. *Consider any semi-random instance $\mathcal{X}$ with N points generated with parameters $\mu_1, \ldots, \mu_k, C_1, \ldots, C_k$ such that $N_i \geq 4(d + \log(\frac{k}{\delta}))$ for all $i \in [k]$. Then with probability at least $1 - \delta$ we have that*

$$\forall i \in [k], \max_{v : \|v\| = 1} \frac{1}{|C_i|} \sum_{x \in C_i} |\langle x - \mu_i, v \rangle|^2 \leq 4\sigma^2. \tag{5}$$

Proof. Exactly as in the proof of Lemma 2.7, we can write $\max_{v:\|v\|=1} \frac{1}{|C_i|} \sum_{x \in C_i} |\langle x - \mu_i, v \rangle|^2 = \max_{v:\|v\|=1} \frac{1}{|C_i|} \sum_{x \in C_i} |\lambda_x^2 \langle y_x - \mu_i, v \rangle|^2 \leq \max_{v:\|v\|=1} \frac{1}{|C_i|} \sum_{x \in C_i} |\langle y_x - \mu_i, v \rangle|^2$. Furthermore, since y_x are points from a Gaussian we know that with probability at least $1 - \delta$, for all $i \in [k]$, $\max_{v:\|v\|=1} \| \frac{1}{|C_i|} \sum_{x \in C_i} |\langle y_x - \mu_i, v \rangle|^2 \leq 4\sigma^2$. Hence, the claim follows. □

We would also need to argue about the mean of a large subset of points from a component of a semi-random GMM.

Lemma 2.9. *Consider any semi-random instance $\mathcal{X}$ with N points generated with parameters $\mu_1, \ldots, \mu_k$ and planted clustering $C_1, \ldots, C_k$ such that $N_i \geq 16(d + \log(\frac{k}{\delta}))$ for all $i \in [k]$. Let $G_i \subseteq C_i$ be such that $|G_i| \geq (1-\varepsilon)|C_i|$ where $\varepsilon < \frac{1}{2}$. Then, with probability at least $1-\delta$, we have that*

$$\forall i \in [k], \|\mu(G_i) - \mu_i\| \leq (4 + \frac{2}{\sqrt{1-\varepsilon}})\sigma. \tag{6}$$

Proof. Let C_i be the set of points in component i and let ν_i be the mean of the points in C_i. Notice that from Lemma 2.7 and the fact that the perturbation is semi-random, we have that with probability at least $1 - \frac{\delta}{2}$, $\|\nu_i - \mu_i\| \leq 2\sigma$. Also, because the component is a semi-random perturbation of a Gaussian, we have from Lemma 2.8 that $\frac{1}{|C_i|}\max_{v:\|v\|=1}\sum_{x\in C_i}[\langle x - \nu_i, v\rangle^2] \leq 4\sigma^2$ with probability at least $1 - \frac{\delta}{2}$.

Hence, with probability at least $1-\delta$ we have that $\|\mu(G_i) - \mu_i\| \leq \|\nu_i - \mu_i\| + \|\mu(G_i) - \nu_i\| \leq 4\sigma + \|\mu(G_i) - \nu_i\|$. To bound the second term notice that $\|\mu(G_i) - \nu_i\| = |(\frac{1}{|G_i|}\sum_{x\in G_i}\langle x - \nu_i, \hat{u}\rangle|$, where $\hat{u}$ is a unit vector in the direction of $(\mu(G_i) - \nu_i)$. Using Cauchy-Schwarz inequality, this is at most $\frac{1}{\sqrt{|G_i|}}\sqrt{\sum_{x\in C_i}\langle x - \nu_i, \hat{u}\rangle^2} \leq \frac{2\sigma}{\sqrt{1-\varepsilon}}$. Combining the two bounds gives us the result. □

Finally, we argue about the variance of the entire data matrix of a semi-random GMM.

Lemma 2.10. *Consider any semi-random instance $\mathcal{X}$ with N points generated with parameters $\mu_1, \ldots, \mu_k, C_1, \ldots, C_k$ such that $N_i \geq 4(d + \log(\frac{k}{\delta}))$ for all $i \in [k]$. Let $A \in \mathbb{R}^{d\times N}$ be the matrix of data points and let $M \in \mathbb{R}^{d\times N}$ be the matrix composed of the means of the corresponding clusters. Then, with probability at least $1-\delta$, we have that*

$$\|A - M\| \leq 4\sigma\sqrt{N}. \tag{7}$$

Proof. Let M^* be the matrix of true means corresponding to the cluster memberships. We can write $\|A - M\| \leq \|A - M^*\| + \|M^* - M\|$. Using Lemma 2.7, we know that with probability at least $1-\frac{\delta}{2}$, $\max_i \|M_i^* - M_i\| \leq 2\sigma$. Hence, $\|M^* - M\| \leq 2\sigma\sqrt{N}$. Furthermore, $\|A - M^*\|^2 = \max_{v:\|v\|=1}\sum_i\sum_{x\in C_i}|(x-\mu_i)\cdot v|^2$. From Lemma 2.8, with probability at least $1 - \frac{\delta}{2}$, we can bound the sum by at most $4\sigma^2 N$. Hence, $\|A - M^*\| \leq 2\sigma\sqrt{N}$. Combining the two bounds we get the claim. □

The following lemma is crucial in analyzing the performance of the Lloyd's algorithm. We would like to upper bound the inner product $|\langle x^{(\ell)} - \mu_i, \widehat{e}\rangle| < \lambda\sigma$ for every direction $\widehat{e}$ and sample $\ell \in [N]$, but this is impossible since $\widehat{e}$ can be aligned along $x^{(\ell)} - \mu_i$. The following lemma however upper bounds the total number of points in the dataset that can have a large projection of λ (or above) onto any direction $\widehat{e}$ by at most $\tilde{O}(d/\lambda^2)$. This involves a union bound over a net of all possible directions $\widehat{e}$.

Lemma 2.11 (Points in Bad Directions)**.** *Consider any semi-random instance $\mathcal{X} = \{x^{(1)}, \ldots, x^{(N)}\}$ with N points having parameters $\mu_1, \ldots, \mu_k, \sigma^2$ and planted clustering $C_1, \ldots, C_k$, and suppose $\forall i \in [k], \ell \in C_i$, $\bar{x}^{(\ell)} = x^{\ell} - \mu_i$. Then there exists a universal constant $c > 0$ s.t. for any $\lambda > 100\sqrt{\log N}$, with probability at least $1 - 2^{-d}$, we have that*

$$\forall \widehat{e} \in \mathbb{R}^d \text{ s.t. } \|\widehat{e}\|_2 = 1, \quad \Big|\{\ell \in [N] : |\langle \bar{x}^{(\ell)}, \widehat{e}\rangle| > \lambda\sigma\}\Big| \leq \frac{cd}{\lambda^2}\cdot\max\Big\{1, \log(\tfrac{3(\sqrt{d}+2\sqrt{\log N})}{\lambda})\Big\}. \tag{8}$$

Proof. Set $\eta := \min\{\lambda/(2\sqrt{d}+2\sqrt{\log N}), \frac{1}{2}\}$ and $m := 512d\log(3/\eta)/\lambda^2$. Consider an η-net $\mathcal{N} \subset \{u : \|u\|_2 = 1\}$ over unit vectors in $\mathbb{R}^d$. Hence

$$\forall u \in \mathbb{R}^d : \|u\|_2 = 1,\ \exists v \in \mathcal{N} \text{ s.t. } \|u-v\|_2 \le \eta \text{ and } |\mathcal{N}| \le \Big(\frac{2+\eta}{\eta}\Big)^d \le \exp\big(d\log(3/\eta)\big).$$

Further, since $|\langle \bar{x}, \widehat{e}\rangle| > \lambda$ and $\mathcal{N}$ is an η-net, there exists some unit vector $u = u(\widehat{e}) \in \mathcal{N}$

$$|\langle \bar{x}, u\rangle| > |\langle \bar{x}, \widehat{e}\rangle + \langle \bar{x}, \widehat{e}-u\rangle| \ge \sigma\lambda - \|\bar{x}\|_2\|\widehat{e}-u\|_2 \ge \sigma\big(\lambda - \eta(\sqrt{d}+2\sqrt{\log N})\big) \ge \frac{\lambda}{2}, \quad (9)$$

for our choice of η. Consider a fixed $x \in \{x^{(1)}, \ldots, x^{(N)}\}$ and a fixed direction $u \in \mathcal{N}$. Since the variance of y is at most σ^2 we have

$$\mathbb{P}\Big[|\langle \bar{x}, u\rangle| > \lambda\sigma/2\Big] \le \mathbb{P}\Big[|\langle \bar{y}, u\rangle| > \lambda\sigma/2\Big] \le \exp(-\lambda^2/8).$$

The probability that m points in $\{x^{(1)}, \ldots, x^{(N)}\}$ satisfy (9) for a fixed direction u is at most $\binom{N}{m} \cdot \exp(-m\gamma^2/2)$. Let E represent the bad event that there exists a direction in $\mathcal{N}$ such that more that m points satisfy the bad event given by (9). The probability of E is at most

$$\begin{aligned}\mathbb{P}[E] &\le |\mathcal{N}| \cdot \binom{N}{m} \exp(-m\lambda^2/8) \le \exp\Big(d\log(3/\eta) + m\log N - \frac{m\lambda^2}{8}\Big) \\ &\le \exp\Big(-d\log(1/\eta)\Big) \le \eta^d,\end{aligned}$$

since for our choice of parameters $\lambda^2 > 32\log N$, and $m\lambda^2 \ge 32d\log(3/\eta)$.

□

3 Upper Bounds for Semi-random GMMs

In this section we prove the following theorem that provides algorithmic guarantees for the Lloyd's algorithm with appropriate initialization, under the semi-random model for mixtures of Gaussians in Definition 1.1.

Theorem 3.1. *There exists a universal constant $c_0, c_1 > 0$ such that the following holds. There exists a polynomial time algorithm that for any semi-random instance $\mathcal{X}$ on N points with planted clustering $C_1, \ldots, C_k$ generated by the semi-random GMM model (Definition 1.1) with parameters $\mu_1, \ldots, \mu_k, \sigma^2$ s.t.*

$$\forall i \ne j \in [k],\ \|\mu_i - \mu_j\|_2 > \Delta\sigma, \quad \text{where } \Delta > c_0\sqrt{\min\{k,d\}\log N}, \quad (10)$$

and $N \ge k^2d^2/w_{min}^2$ finds w.h.p. a clustering $C_1', C_2', \ldots, C_k'$ such that

$$\min_{\pi \in Perm_k} \sum_{i=1}^{k} \Big|C_{\pi(i)} \triangle C_i'\Big| \le \frac{c_1 kd}{\Delta^4} \cdot \max\Big\{1, \log\big(\tfrac{3(\sqrt{d}+2\sqrt{\log N})}{\Delta^2}\big)\Big\}.$$

In Section 4 we show that the above error bound is close to the information theoretically optimal bound (up to the logarithmic factor). The Lloyd's algorithm as described in Figure 1 consists of two stages, the initialization stage and an iterative improvement stage.

The initialization follows the same scheme as proposed by Kumar and Kannan in [KK10]. The initialization algorithm first performs a k-SVD of the data matrix followed by running the k-means++ algorithm [AV07] that uses D^2-sampling to compute seed centers.

One can also use any constant factor approximation algorithm for k-means clustering in the projected space to obtain the initial centers [KMN+02, ANSW16]. This approach works for clusters that are nearly balanced in size. However, when the cluster sizes are arbitrary, an appropriate transformation of the data is performed first that amplifies the separation between the centers. Following this transformation, the (k-SVD + k-means++) is used to get the initial centers. The formal guarantee of the initialization procedure is encapsulated in the following proposition, whose proof is given in Section 3.2.

The main algorithmic contribution of this paper is an analysis of the Lloyd's algorithm when the points come from the semi-random GMM model. For the rest of the analysis we will assume that the instance $\mathcal{X}$ generated from the semi-random GMM model satisfies (1) to (8). These eight equations are shown to hold w.h.p. in Section 2.1 for instances generated from the model. Our analysis will in fact hold for any deterministic data set satisfying these equations. This helps to gracefully argue about performing many iterations of Lloyd's on the same data set without the need to draw fresh samples at each step.

Proposition 3.2. *In the above notation for any $\delta > 0$, suppose we are given an instance $\mathcal{X}$ on N points satisfying* (1)-(8) *such that $|C_i| \geq \Omega(d + \log(\frac{k}{\delta}))$ and assume that $\Delta \geq 125\sqrt{\min\{k, d\}\log N}$. Then after the initialization step, for every μ_i there exists μ_i' such that $\|\mu_i - \mu_i'\| \leq \tau\sigma$, where $\tau < \Delta/24$.*

The analysis of the Lloyd's iterations crucially relies on the following lemma that upper bounds the number of misclassified points when the current Lloyd's iterative is relatively close to the true means.

Lemma 3.3 (Projection condition)**.** *In the above notation, consider an instance $\mathcal{X}$ satisfying* (1)-(8) *and* (10) *and suppose we are given $\mu_1', \ldots, \mu_k'$ satisfying $\forall j \in [k],\ \|\mu_j' - \mu_j\|_2 \leq \tau\sigma$ and $\tau < \Delta/24$. Then there exists a set $Z \subset \mathcal{X}$ such that for any $i \in [k]$ we have*

$$\forall x \in C_i \cap (\mathcal{X}\backslash Z),\ \ \|x-\mu_i'\|_2^2 \leq \min_{j\neq i}\|x-\mu_j'\|_2^2 \quad \text{where } |Z| = O\Big(\frac{d\tau^2}{\Delta^4}\cdot\max\Big\{1, \log\big(\tfrac{3\tau(\sqrt{d}+2\sqrt{\log N})}{\Delta^2}\big)\Big\}\Big).$$

The following lemma quantifies the improvement in each step of the Lloyd's algorithm. The proof uses Lemma 3.3 along with properties of semi-random Gaussians.

Lemma 3.4. *In the above notation, suppose we are given an instance $\mathcal{X}$ on N points with $w_i N \geq \frac{d\sqrt{d}}{4\log(d)}$ for all i satisfying* (1)-(8). *Furthermore, suppose we are given centers $\mu_1', \ldots, \mu_k'$ such that $\|\mu_i' - \mu_i\| \leq \tau\sigma,\ \forall i \in [k]$ where $\tau < \Delta/24$. Then the centers $\mu_1'', \ldots, \mu_k''$ obtained after one Lloyd's update satisfy $\|\mu_i'' - \mu_i\| \leq \max((6 + \frac{\tau}{4})\sigma, \frac{\tau}{2}\sigma))$ for all $i \in [k]$.*

1. Let A be the $N \times d$ data matrix with rows A_i for $i \in [N]$. Use A to compute initial centers $\mu_0^{(1)}, \mu_0^{(2)}, \ldots \mu_0^{(k)}$ as detailed in Proposition 3.2.
2. Use these k-centers to seed a series of Lloyd-type iterations. That is, for $r = 1, 2, \ldots$ do:
   - Set Z_i be the set of points for which the closest center among $\mu_{r-1}^{(1)}, \mu_{r-1}^{(2)}, \ldots, \mu_{r-1}^{(k)}$ is $\mu_{r-1}^{(i)}$.
   - Set $\mu_r^{(i)} \leftarrow \frac{1}{|Z_i|}\sum_{A_j \in Z_i} A_j$.

Figure 1: Lloyd's Algorithm

We now present the proof of Theorem 3.1.

Proof of Theorem 3.1. Firstly, the eight deterministic conditions (1)-(8) are shown to hold for instance $\mathcal{X}$ w.h.p. in Section 2.1. The proof follows in a straightforward manner by combining Proposition 3.2, Lemma 3.4 and Lemma 3.3. Proposition 3.2 shows that $\|\mu_i^{(0)} - \mu_i\|_2 \leq \Delta/(24)$ for all $i \in [k]$. Applying Lemma 3.4, we have that after $T = O(\log \Delta)$ iterations we get $\|\mu_i^{(T)} - \mu_i\|_2 \leq 8\sigma$ for all $i \in [k]$ w.h.p. Finally using Lemma 3.3 with $\tau = 1$, the theorem follows. □

3.1 Analyzing Lloyd's Algorithm

We now analyze each iteration of the Lloyd's algorithm and show that we make progress in each step by misclassifying fewer points with successive iterations. As a first step we begin with the proof of Lemma 3.3.

Proof of Lemma 3.3. Set $m := 512d\log(3/\eta)\tau^2/\Delta^4$ where $\eta = \min\{\, \Delta^2/(\tau(2\sqrt{d} + 2\sqrt{\log N})), \frac{1}{2}\,\}$.

Fix a sample $x \in \{\, x^{(1)}, \ldots, x^{(N)}\,\}$ and suppose $x \in C_i$ and let $y := y(x)$ be the corresponding point before the semi-random perturbation, and let $\bar{x} = x - \mu_i$, $\bar{y} = y - \mu_i$. For each $i \in [k]$, let $\widehat{e}_i$ be the unit vector along $(\mu_i - \mu_i')$.

We first observe that by projecting the Gaussians around μ_i, μ_j onto the direction along $\widehat{e}_{ij} = (\mu_i - \mu_j)/\|\mu_i - \mu_j\|_2$, we have that

$$\begin{aligned}\|x - \mu_j\|_2^2 - \|x - \mu_i\|_2^2 &= \langle x - \mu_j, \widehat{e}_{ij}\rangle^2 - \langle x - \mu_i, \widehat{e}_{ij}\rangle^2 \geq (|\langle x - \mu_j, \widehat{e}_{ij}\rangle| - |\langle x - \mu_i, \widehat{e}_{ij}\rangle|)^2 \\ &\geq (|\langle \mu_i - \mu_j, \widehat{e}_{ij}\rangle| - 2|\langle x - \mu_i, \widehat{e}_{ij}\rangle|)^2 \geq (\Delta\sigma - 2|\langle x - \mu_i, \widehat{e}_{ij}\rangle|)^2 \\ &\geq (\Delta\sigma - 6\sigma\sqrt{\log N})^2 \geq \tfrac{1}{4}\Delta^2\sigma^2, \end{aligned} \tag{11}$$

where the first inequality follows from (3), and the second inequality uses $\Delta > 12\sqrt{\log N}$.

Suppose $x \in C_i$ is misclassified i.e., $\|x - \mu_i'\|_2 \geq \|x - \mu_j'\|_2$ for some $j \in [k] \setminus \{\, i\,\}$. Then,

$$\begin{aligned} \|(x - \mu_i) + \mu_i - \mu_i'\|_2^2 &\geq \|(x - \mu_j) + (\mu_j - \mu_j')\|_2^2 \\ 2\langle x - \mu_i, \mu_i - \mu_i'\rangle - 2\langle x - \mu_j, \mu_j - \mu_j'\rangle &\geq \|x - \mu_j\|_2^2 - \|x - \mu_i\|_2^2 + \|\mu_j - \mu_j'\|_2^2 - \|\mu_i - \mu_i'\|_2^2 \\ 2\langle \bar{x}, \mu_i - \mu_i'\rangle - 2\langle \bar{x}, \mu_j - \mu_j'\rangle - 2\langle \mu_i - \mu_j, \mu_j - \mu_j'\rangle &\geq \tfrac{1}{4}\Delta^2\sigma^2 - \tau^2\sigma^2 \qquad \text{(from (11))} \\ 2|\langle \bar{x}, \mu_i - \mu_i'\rangle| + 2|\langle \bar{x}, \mu_j - \mu_j'\rangle| &\geq (\tfrac{1}{4}\Delta^2 - \tau^2)\sigma^2 - 2|\langle \mu_i - \mu_j, \mu_j - \mu_j'\rangle| \\ \left|\left\langle \bar{x}, \frac{\mu_i - \mu_i'}{\|\mu_i - \mu_i'\|_2}\right\rangle\right| + \left|\left\langle \bar{x}, \frac{\mu_j - \mu_j'}{\|\mu_j - \mu_j'\|_2}\right\rangle\right| &\geq \frac{(\frac{\Delta^2}{8} - \frac{\tau^2}{2} - \tau\Delta)\sigma^2}{\tau\sigma} \geq \frac{\Delta^2}{16\tau}\sigma, \end{aligned}$$

since $\tau < \Delta/(24)$. Hence, we have that if $x \in C_i$ is misclassified by $\mu_1', \ldots, \mu_k'$ then

$$|\langle \bar{x}, \widehat{e}\rangle| > \sigma\Delta^2/(32\tau) \text{ for some unit vector } \widehat{e} \in \mathbb{R}^d. \tag{12}$$

From (8) with $\lambda = \Delta^2/(32\tau)$, we get from (8) that at most m points in C_i can satisfy (12). Hence the lemma follows. □

Next we prove Lemma 3.4, which quantifies the improvement in every iteration of the Lloyd's algorithm.

Proof of Lemma 3.4. Let $C_1, C_2, \ldots, C_k$ be the partitioning of the indices according to the ground truth clustering of the semi-random instance $\mathcal{X}$ and $S_1, S_2, \ldots, S_k$ be the indices of the clustering obtained by using the centers μ_i'. Then $\mu_i'' = \frac{1}{|S_i|}\sum_{t \in S_i} x^{(t)}$. Partition S_i into two sets G_i and B_i where $G_i = S_i \cap C_i$ and $B_i = S_i \setminus G_i$. Let $\mu(G_i)$ and $\mu(B_i)$ be the means of the two partitions respectively. Let $\gamma = O(\frac{d\tau^2}{\Delta^4}\max\{\, 1, \log(\frac{3\tau(\sqrt{d}+2\sqrt{\log N})}{\Delta^2})\,\})$.

From Lemma 3.3 we know that $|G_i| \geq |C_i| - \gamma$ and $|B_i| \leq k\gamma$. Then we have that $\mu_i'' = \frac{|G_i|}{|S_i|}\mu(G_i) + \frac{|B_i|}{|S_i|}\mu(B_i)$. Hence, $\|\mu_i'' - \mu_i\| \leq \frac{|G_i|}{|S_i|}\|\mu(G_i) - \mu_i\| + \frac{|B_i|}{|S_i|}\|\mu(B_i) - \mu_i\|$.

We have $\frac{|G_i|}{|C_i|} \geq 1 - \frac{\gamma}{|C_i|} \geq 1 - \frac{\tau}{64\sqrt{k}\sqrt{d}}$ using the bound on Δ and $|C_i| = w_i N \geq \frac{d\sqrt{d}}{4\log(d)}$. Using (6) we get that

$$\frac{|G_i|}{|S_i|}\|\mu(G_i) - \mu_i\| \leq \Big(4 + \frac{2}{\sqrt{1 - \frac{\tau}{64\sqrt{k}\sqrt{d}}}}\Big)\sigma \leq \Big(6 + \frac{\tau}{128\sqrt{k}\sqrt{d}}\Big)\sigma \leq 6\sigma + \frac{\tau}{8}\sigma.$$

To bound the second term we first show that for each point $x^{(t)} \in B_i$, $\|x^{(t)} - \mu_i\| \leq (\sqrt{d} + 2\sqrt{\log N} + 2\tau)\sigma$. Let C_j be the cluster that point $x^{(t)}$ belongs to. Then

$$\|x^{(t)} - \mu_i\| \leq \|x^{(t)} - \mu_i'\| + \tau\sigma \leq \|x^{(t)} - \mu_j'\| + \tau\sigma \leq \|x^{(t)} - \mu_j\| + 2\tau\sigma \leq (\sqrt{d} + 2\sqrt{\log N} + 2\tau)\sigma,$$

using (1). Hence,

$$\frac{|B_i|}{|S_i|}\|\mu(B_i) - \mu_i\| \leq \frac{|B_i|}{|S_i|}(\sigma\sqrt{d} + \sigma\sqrt{\log N} + 2\tau\sigma) \leq \frac{2k\gamma}{|C_i|}(\sigma\sqrt{d} + \sigma\sqrt{\log N} + 2\tau\sigma) < \frac{\tau}{8}\sigma.$$

Combining, we get that $\|\mu_i'' - \mu_i\| \leq (6 + \frac{\tau}{4})\sigma \leq \max(6\sigma + \frac{\tau}{4}, \frac{\tau}{2}\sigma)$. □

3.2 Initialization

In this section we describe how to obtain the initial centers satisfying the condition in Lemma 3.4. The final initialization procedure relies on the following subroutine that provides a good initializer if the mean separation is much larger than that in Theorem 3.1. Let A denote the $N \times d$ matrix of data points and M^* be the $N \times d$ matrix where each row of C is equal to one of the means μ_is of the component to which the corresponding row of A belongs to.

Lemma 3.5. *In the above notation, for any $\delta > 0$ suppose we are given an instance $\mathcal{X}$ on N points satisfying satisfying* (1)-(8)*, with components $C_1, \ldots C_k$ such that $|C_i| \geq \Omega(d + \log(\frac{k}{\delta}))$. Let A be the $N \times d$ matrix of data points and $\hat{A}$ be the matrix obtained by projecting points onto the best k-dimensional subspace obtained by SVD of A. Let μ_i' be the centers obtained by running an α factor k-means approximation algorithm on $\hat{A}$. Then for every μ_i there exists μ_i' such that $\|\mu_i - \mu_i'\| \leq 20\sqrt{k\alpha}\frac{\|A - M^*\|}{\sqrt{N w_{min}}}$.*

Proof. Let $\hat{A}$ denote the matrix obtained by projecting A onto the span of its top k right singular vectors. Furthermore, let $\nu_1, \ldots \nu_k$ be the centers obtained by running a 9-approximation algorithm for k-means on the instance $\hat{A}$. We know that the optimal k-means solution for $\hat{A}$ is at most $\|\hat{A} - M^*\|_F^2$. Since both $\hat{A}$ and M^* are rank k matrices, we get that $\|\hat{A} - M^*\|_F^2 \leq 2k\|\hat{A} - M^*\|_2^2 \leq 2k(\|\hat{A} - A\|_2^2 + \|A - M^*\|_2^2)$. Since $\hat{A}$ is the best rank k approximation to A we also have that $\|\hat{A} - A\|_2^2 \leq \|A - M^*\|_2^2$. Hence, $\|\hat{A} - M^*\|_F^2 \leq 4k\|A - M^*\|_2^2$. Hence, the cost of the solution using centers ν_is must be at most $36k\sigma^2 N$ (using 7).

Next, suppose that there exists μ_i such that for all j, $\|\mu_i - \nu_j\| > 20\sqrt{k\alpha}\frac{\|A - M^*\|}{\sqrt{N w_{\min}}}$. let's compute the cost paid by the points in component C_i in the clustering obtained via the approximation algorithm. For any $x \in C_i$ let ν_x be the center that it is closest to. Then the cost is at least $\sum_{x \in C_i}\|x - \nu_x\|^2 \geq \sum_{x \in C_i}\frac{1}{2}\|\mu_i - \nu_x\|^2 - \|x - \mu_i\|^2$. The first summation is at least $\frac{1}{2}|N w_{\min}|(400\alpha k\frac{\|A - M^*\|^2}{N w_{\min}}) > 200k\alpha\|A - M^*\|^2$. The second summation is at most $\sum_{x \in C_i}\|x - \mu_i\|^2 \leq \sum_i \sum_{x \in C_i}\|x - \mu_i\|^2 = \|\hat{A} - M^*\|_F^2 \leq 4k\|A - M^*\|^2$. Hence, we reach a contradiction to the fact that the solution obtained via ν_is is an α-approximation to the optimal cost. □

The proof of the above theorem already provides a good initializer provided Δ is larger than $\sqrt{k\frac{\log N}{w_{\min}}}$ and one uses a constant factor approximation algorithm for k-means [ANSW16]. Furthermore, if Δ is larger than $\sqrt{k\log k\frac{\log N}{w_{\min}}}$, then one can instead use the simpler and faster k-means++ approximation algorithm [AV07]. The above lemma has a bad dependence on $w_{\min}$. However, using the Boosting technique of [KK10] we can reduce the dependence to $\Delta > 25\sqrt{k\log N}$ and hence prove Proposition 3.2. We provide a proof of this in the Appendix.

4 Lower Bounds for Semi-random GMMs

We prove the following theorem.

Theorem 4.1. *For any $d, k \in Z_+$, there exists $N_0 = \mathrm{poly}(d,k)$ and a universal constant $c_1 > 0$ such that the following holds for all $N \geq N_0$ and Δ such that $\sqrt{\log N} \leq \Delta \leq d/(64\log d)$. There exists an instance $\mathcal{X}$ on N points in d dimensions with planted clustering $C_1, \ldots, C_k$ generated by applying semi-random perturbations to points generated from a mixture of spherical Gaussians with means $\mu_1, \mu_2, \ldots, \mu_k$, covariance $\sigma^2 I$ and weights being $1/k$ each, with separation $\forall i \neq j \in [k], \|\mu_i - \mu_j\|_2 \geq \Delta\sigma$, such that any locally optimal k-means clustering solution $C_1', C_2', \ldots, C_k'$ of $\mathcal{X}$ satisfies w.h.p.*

$$\min_{\pi \in Perm_k} \sum_{i=1}^{k} |C'_{\pi(i)} \triangle C_i| \geq \frac{c_1 kd}{\Delta^4}.$$

It suffices to set $N_0(d,k) := c_0 k^2 d^{3/2}\log^2(kd)$, where $c_0 > 0$ is a sufficiently large universal constant.

Remark 4.2. Note that the lower bound also applies in particular to the more general semi-random model in Definition 1.1; in this instance, the points are drawn i.i.d. from the mixture of spherical Gaussians, before applying semi-random perturbations. Further, this lower bound holds for any *locally optimal solution*, and not just the optimal solution.

The lower bound construction will pick an arbitrary $\Omega(d/\Delta^4)$ points from $k/2$ clusters, and carefully choose a semi-random perturbation to all the points so that these $\Omega(kd/\Delta^4)$ points are misclassified. We start with a simple lemma that shows that an appropriate semi-random perturbation can move the mean of a cluster by an amount $O(\sigma)$ along any fixed direction.

Lemma 4.3. *Consider a spherical Gaussian in d dimensions with mean μ and covariance $\sigma^2 I$, and let $\widehat{e}$ be a fixed unit vector. Consider the semi-random perturbation given by*

$$\forall y \in \mathbb{R}^d, h(y) = \begin{cases} \mu & \text{if } \langle y - \mu, \widehat{e}\rangle < 0 \\ y & \text{otherwise} \end{cases}.$$

Then we have $\mathbb{E}[h(y)] = \mu + \frac{1}{\sqrt{2\pi}}\sigma\widehat{e}$.

Proof. We assume without loss of generality that $\mu = 0, \sigma = 1$ (by shifting and scaling) and $\widehat{e} = (1, 0, 0, \ldots, 0) \in \mathbb{R}^d$ (by the rotational symmetry of a spherical Gaussian). Let γ be the p.d.f. of the standard Gaussian in d dimensions with mean 0, and $\gamma'(y)$ be the distribution on y conditioned on the event $[y(1) = \langle y, \widehat{e}\rangle > 0]$. First, we observe that $\mathbb{E}[h(y)|y_1 < 0] = 0$ from construction, and $\mathbb{E}[h(y)|y_1 > 0] = \mathbb{E}_{y\sim\gamma'(y)}[y]$. Further, since the $(d-1)$ co-ordinates of y orthogonal to $\widehat{e}$ are independent of y_1,

$$\mathbb{E}[h(y)] = \mathbb{P}[y_1 < 0]\,\mathbb{E}[h(y)|y_1 < 0] + \mathbb{P}[y_1 > 0]\,\mathbb{E}[h(y)|y_1 > 0] = \frac{1}{2}\,\mathbb{E}[y_1|y_1 > 0]\widehat{e}$$

$$\mathbb{E}[h(y)] - \mu\,\mathbb{E}[h(y)] = \Big(\frac{1}{2\sqrt{2\pi}}\int_{-\infty}^{\infty} |y_1|\exp(-y_1^2/2)\; dy_1\Big)\widehat{e} = \frac{\sigma}{\sqrt{2\pi}}\widehat{e}.$$

□

Construction. Set $m := c_1 d/\Delta^4$ for some appropriately small constant $c_1 \in (0,1)$. We assume without loss of generality that k is even (the following construction also works for odd k by leaving the last cluster unchanged). We pair up the clusters into $k/2$ pairs $\{\,(C_1, C_2), (C_3, C_4), \ldots, (C_{k-1}, C_k)\,\}$, and we will ensure that m points are misclassified in each of the $k/2$ clusters $C_1, C_3, \ldots, C_{k-1}$. The parameters of the mixture of spherical Gaussians $\mathcal{G}$ are set up as follows. For each $i \in 1, 3, 5, \ldots, k-1$, $\|\mu_i - \mu_{i+1}\|_2 = \Delta\sigma$, and all the other inter-mean distances (across different pairs) are at least $M\sigma$ which is arbitrarily large (think of $M \mapsto \infty$).

- Let for any $i \in \{\,1, 3, \ldots, k-1\,\}$, $Z_i \subset C_i$ be the first m points in cluster C_i respectively among the samples $y^{(1)}, \ldots, y^{(N)}$ drawn from $\mathcal{G}$ (these m points inside the clusters can be chosen arbitrarily). Set $Z_i = \emptyset$ for $i \in \{\,2, 4, \ldots, k\,\}$.
- For each $i \in \{\,1, 3, \ldots, k-1\,\}$, set $\widehat{e}_i$ to be the unit vector along $u_i = \frac{1}{\sigma\sqrt{md}} \sum_{y \in Z_i} (y - \mu_i)$.
- For each $i \in \{\,1, 3, \ldots, k-1\,\}$ apply the following semi-random perturbation given by Lemma 4.3 to points in cluster C_{i+1} along $\widehat{e}_i$, i.e., each point $y^{(t)} \in C_{i+1}$

$$x^{(t)} = h(y^{(t)}) = \begin{cases} \mu_{i+1} & \text{if } \langle y^{(t)} - \mu_{i+1}, \widehat{e}_i \rangle < 0 \\ y^{(t)} & \text{otherwise} \end{cases}.$$

Note that the semi-random perturbations are only made to points in the even clusters (based on a few points in its respective odd cluster). The lower bound proof proceeds in two parts. Lemma 4.5 (using Lemma 4.3) and Lemma 4.6 shows that in any k-means optimal clustering the means of each even cluster C_i moves by roughly $\Omega(\sigma) \cdot \widehat{e}_{i-1}$. Lemma 4.7 then shows that these means will classify all the m points in Z_{i-1} *incorrectly* w.h.p. In this proof w.h.p. will refer to a probability of at least $1 - o(1)$ unless specified otherwise (this can be made $1 - 1/\mathrm{poly}(m, k)$ by choosing suitable constants).

We start with two simple concentration statements about the points in Z_i (from Lemma A.2 and Lemma 2.6). We have with probability at least $1 - 1/(mk)$,

$$\forall i \in \{\,1, 3, k-1\,\},\ \forall t \in Z_i\ \ \|x^{(t)} - \mu_i\|_2 \le \sigma(\sqrt{d} + 2\sqrt{\log(mk)}) \tag{13}$$

$$\forall i \in \{\,1, 3, k-1\,\},\ \forall t \in Z_i\ \ |\langle x^{(t)} - \mu_i, \mu_i - \mu_{i+1} \rangle| \le 2\sqrt{\log(mk)}\Delta\sigma^2 \tag{14}$$

We start with a couple of simple claims about the unit vectors $\widehat{e}_1, \widehat{e}_3, \ldots, \widehat{e}_{k-1}$.

Lemma 4.4. *In the above construction, for every $i \in \{\,1, 3, \ldots, k-1\,\}$ we have w.h.p. $\|\widehat{e}_i - u_i\|_2^2 \le 6\sqrt{m \log(mk)/d}$. Further, for each $x \in Z_i$, we have $\langle x - \mu_i, \widehat{e}_i \rangle \ge \frac{1}{2}\sigma\sqrt{d/m}$.*

Proof. Let us fix an $i \in \{\,1, 3, \ldots, k-1\,\}$. Let $y^{(1)}, y^{(2)}, \ldots, y^{(m)} \in Z_i$ and $\bar{y}^{(t)} = y^{(t)} - \mu_i$.

From (13), we know that w.h.p., $\|\bar{y}^{(t)}\|_2 \le \sigma(\sqrt{d} + 2\sqrt{\log m})$ $\forall t \in [m]$. Fix $t \in [m]$, and let $Q(t) = \sum_{t' \in [m] \setminus \{\,t\,\}} \langle \bar{y}^{(t)}, \bar{y}^{(t')} \rangle$. For $t' \neq t$, due to independence and spherical symmetry, $\frac{1}{\|\bar{y}^{(t)}\|_2} \langle \bar{y}^{(t)}, \bar{y}^{(t')} \rangle$ is distributed as a normal r.v. with mean 0 and variance σ^2. Further, $Q(t)/\|y^{(t)}\|_2$ is distributed as a normal r.v. with mean 0 and variance $\sigma^2 m$. Hence,

$$Q(t) = \|y^{(t)}\|_2 \cdot \sum_{t' \in [m] \setminus \{\,t\,\}} \langle \bar{y}^{(t')}, \frac{\bar{y}^{(t)}}{\|\bar{y}^{(t)}\|_2} \rangle \le \sigma^2(\sqrt{d} + \sqrt{\log(mk)}) \cdot 2\sqrt{m \log(mk)}, \tag{15}$$

with probability at least $1 - 1/(mk)^2$. Hence, w.h.p. $Q(t) \le 4\sigma^2\sqrt{dm \log(mk)}$ for all $t \in [m]$.

For the first part, we see that

$$\|u_i\|_2^2 = \frac{1}{\sigma^2 md}\Big(\sum_{t\in[m]}\|\bar{y}^{(t)}\|_2^2 + 2\sum_{t\neq t'\in[m]}\langle \bar{y}^{(t)}, \bar{y}^{(t')}\rangle\Big) = \frac{1}{\sigma^2 md}\Big(\sum_{t\in[m]}\|\bar{y}^{(t)}\|_2^2 + 2\sum_{t\in[m]} Q(t)\Big).$$

Along with (13), the bound on $Q(t)$ and $\mathbb{E}[\|y^{(t)}\|_2^2] = d\sigma^2$, this implies

$$|\|u_i\|_2^2 - 1| \leq \frac{1}{md}(4m\sqrt{d\log(mk)} + 4m\log(mk) + 4m\sqrt{dm\log(mk)}) \text{ w.p. at least } 1 - 1/(mk)$$

$$|\|u_i\|_2^2 - 1| \leq 6\sqrt{\frac{m\log(mk)}{d}} \text{ with probability at least } 1 - 1/(mk).$$

Since $\widehat{e}_i$ is the unit vector along u_i, and performing a union bound over all i we have that w.h.p., $\|\widehat{e}_i - u_i\|_2^2 \leq 6\sqrt{m\log(mk)/d}$.

For the furthermore part, suppose $x = y^{(t)}$ for some $t \in [m]$ then

$$\begin{aligned}\langle \bar{x}, \widehat{e}_i\rangle &= \frac{1}{\|u_i\|_2\sqrt{md}}\sum_{t'\in[m]}\langle y^{(t)}, y^{(t')}\rangle \geq \frac{1}{\|u_i\|_2\sqrt{md}}\Big(\|\bar{y}^{(t)}\|_2^2 - \sum_{t'\neq t}|\langle \bar{y}^{(t)}, \bar{y}^{(t')}\rangle|\Big)\\ &\geq \frac{\sigma^2}{\|u_i\|_2\sqrt{md}}\Big((d - \sqrt{d\log(mk)}) - Q(t)\Big) \geq \frac{\sigma^2}{\|u_i\|_2\sqrt{md}}(d - 4\sqrt{dm\log(mk)})\\ &\geq \frac{\sigma^2}{4}\sqrt{d}m,\end{aligned}$$

since $64m\log m \leq d$ and $\|u_i\|_2 \leq 2$ w.h.p. □

Let $\tilde{\mu}_1, \ldots, \tilde{\mu}_k$ be the (empirical) means of the clusters in the planted clustering $C_1, C_2, \ldots, C_k$ after the semi-random perturbations. The following lemma shows that $\|\tilde{\mu}_i - \mu_i\|_2 \leq \sigma$.

Lemma 4.5. *There exists a universal constant $c_3 > 0$ s.t. for the semi-random instance $\mathcal{X}$ described above, we have that w.h.p.*

$$\forall i \in [k],\ \tilde{\mu}_i = \begin{cases}\mu_i + \frac{1}{\sqrt{2\pi}}\sigma\widehat{e}_{i-1} + z_i & \text{if } i \text{ is even}\\ \mu_i + z_i & \text{if } i \text{ is odd}\end{cases}, \quad \text{where } \|z_i\|_2 \leq c_3\sigma\sqrt{\frac{dk}{N}}.$$

Proof. The lemma follows in a straightforward way from Lemma 4.3 and by standard concentration bounds. Firstly, the clusters C_i for odd i are unaffected by the perturbation. Hence, $\mathbb{E}_{x\in C_i}[x] = \mu_i$ and from Lemma A.4, the empirical mean of the points in C_i (there are at least $N/(2k)$ of them w.h.p.) gives the above lemma. Consider any even i. From Lemma 4.3, the semi-random perturbation applied to the points in C_i along the direction $\widehat{e}_{i-1}$ ensures that $\mathbb{E}_{x\in C_i}[x] = \mu_i + \frac{\sigma}{\sqrt{2\pi}}\widehat{e}_{i-1}$. Again by Lemma A.4 applied to the points from C_i, the lemma follows. □

The following lemma shows that if C_i, C_i' are close, then the empirical means are also close.

Lemma 4.6. *Consider any cluster C_i of the instance $\mathcal{X}$, and let C_i' satisfy $|C_i' \triangle C_i| \leq m'$. Suppose $\tilde{\mu}_i$ and μ_i' are the means of clusters C_i and C_i' respectively, then*

$$\|\mu_i' - \tilde{\mu}_i\|_2 \leq 4\sigma \cdot \frac{m'}{|C_i|}(\sqrt{d} + 2\sqrt{\log N} + \Delta).$$

Proof. Let i be even (an even cluster). First, we note that from our construction, all the points in $C_i' \setminus C_i \in C_{i-1}$ w.h.p., since the distance between the means $\|\mu_i - \mu_j\|_2 \geq M\sigma$ when $j \notin \{ i-1, i \}$, for M that is chosen to be appropriately large enough. Further, $\|\mu_i - \mu_{i-1}\|_2 = \Delta\sigma$. Let $\bar{x} = x - \mu_i$ if $i \in C_i$ and $\bar{x} = x - \mu_j$ if $x \in C_j$. Hence w.h.p.,

$$\forall x \in C_i' \cup C_i, \quad \|x - \mu_i\|_2 \leq \Delta\sigma + \|\bar{x}\|_2 \leq \Delta\sigma + (\sqrt{d} + 2\sqrt{\log N})\sigma.$$

Further, $\tilde{\mu}_i$ is the empirical mean of all the points in C_i. Let $\delta = m'/|C_i|$.

$$\mu_i' - \mu_i = \frac{\sum_{x \in C_i}(x - \mu_i)}{|C_i'|} - \frac{\sum_{x \in C_i \setminus C_i'}(x - \mu_i)}{|C_i'|} + \frac{\sum_{x \in C_i' \setminus C_i}(x - \mu_i)}{|C_i'|}$$

$$\mu_i' - \tilde{\mu}_i = (\mu_i' - \mu_i) - (\tilde{\mu}_i - \mu_i) = (\mu_i' - \mu_i) + \frac{\sum_{x \in C_i}(x - \mu_i)}{|C_i|}$$

$$\text{Hence, } \mu_i' - \tilde{\mu}_i = (\tfrac{|C_i|}{|C_i'|} - 1)(\tilde{\mu}_i - \mu_i) - \frac{1}{|C_i'|}\sum_{x \in C_i \setminus C_i'}(x - \mu_i) + \frac{1}{|C_i'|}\sum_{x \in C_i' \setminus C_i}(x - \mu_i)$$

$$\begin{aligned}\|\mu_i' - \tilde{\mu}_i\|_2 &\leq (\frac{\delta}{1-\delta})\|\tilde{\mu}_i - \mu_i\|_2 + (\frac{2\delta}{1-\delta})\max_{x \in C_i \cup C_i'}\|x - \mu_i\| \\ &\leq (\frac{2\delta\sigma}{1-\delta})(1 + \Delta + \sqrt{d} + 2\sqrt{\log N}) \leq 4\delta\sigma(\Delta + \sqrt{d} + 2\sqrt{\log N}),\end{aligned}$$

where $\|\mu_i - \tilde{\mu}_i\|_2$ is bounded because of Lemma 4.5. A similar argument follows when i is odd. □

The following lemma shows that the Voronoi partition about $\tilde{\mu}_1, \ldots, \tilde{\mu}_k$ (or points close to it) incorrectly classify all points in Z_i for each $i \in [k]$.

Lemma 4.7. *Let $\mu_1', \mu_2', \ldots, \mu_k'$ satisfy $\|\mu_i' - \tilde{\mu}_i\|_2 \leq \sigma/(16\sqrt{m}(1 + 2\sqrt{\frac{\log N}{d}}))$, where $\tilde{\mu}_i$ is the empirical mean of the points in C_i. Then, we have w.h.p. that for each $i \in \{ 1, 3, \ldots, k-1 \}$, $\|x - \mu_i'\|_2^2 > \|x - \mu_{i+1}'\|_2^2$, i.e., every point $x \in Z_i$ is misclassified.*

Proof. Let i be odd, and consider a point x in Z_i, and let $\bar{x} = x - \mu_i$.

$$\begin{aligned}&\|x - \mu_i'\|_2^2 - \|x - \mu_{i+1}'\|_2^2 = \|(x - \mu_i) + \mu_i - \mu_i'\|_2^2 - \|(x - \mu_{i+1}) + (\mu_{i+1} - \mu_{i+1}')\|_2^2 \\ &\quad = \|x - \mu_i\|_2^2 - \|x - \mu_i + (\mu_i - \mu_{i+1})\|_2^2 + 2\langle x - \mu_i, \mu_i - \mu_i'\rangle - 2\langle x - \mu_{i+1}, \mu_{i+1} - \mu_{i+1}'\rangle \\ &\qquad + \|\mu_i - \mu_i'\|_2^2 - \|\mu_{i+1} - \mu_{i+1}'\|_2^2 \\ &\quad \geq 2\langle \bar{x}, \mu_i - \mu_i'\rangle + 2\langle \bar{x}, \mu_{i+1}' - \mu_{i+1}\rangle + 2\langle \bar{x}, \mu_{i+1} - \mu_i\rangle - \Delta^2\sigma^2 - 2\langle \mu_i - \mu_{i+1}, \mu_{i+1} - \mu_{i+1}'\rangle - \sigma^2 \\ &\quad \geq 2\langle \bar{x}, \mu_i - \mu_i'\rangle + 2\langle \bar{x}, \mu_{i+1}' - \mu_{i+1}\rangle - 4\Delta\sqrt{\log(mk)}\sigma^2 - \Delta^2\sigma^2 - 2\Delta\sigma^2 - \sigma^2,\end{aligned}$$

where the last inequality follows from (14). From Lemma 4.5, we have

$$\mu_{i+1}' - \mu_{i+1} = (\tilde{\mu}_{i+1} - \mu_{i+1}) + (\mu_{i+1}' - \tilde{\mu}_{i+1}) = \tfrac{1}{\sqrt{2\pi}}\sigma\widehat{e}_i + z_{i+1}',$$

$$\text{where } \|z_{i+1}'\|_2 \leq \|z_{i+1}\|_2 + \|\mu_{i+1}' - \tilde{\mu}_{i+1}\|_2 \leq \sigma \cdot \frac{1}{(12\sqrt{m}(1 + 2\sqrt{\log N/d})}},$$

since $N/\sqrt{\log N} \geq Cd^{3/2}km$ for some appropriately large constant $C > 0$. Similarly $\mu_i' - \mu_i = z_i'$, where $\|z_i'\|_2 \leq \sigma/(12\sqrt{m}(1 + \sqrt{\log N/d}))$. Hence, simplifying and applying Lemma 4.4 we get

$$\begin{aligned}\|x - \mu_i'\|_2^2 - \|x - \mu_{i+1}'\|_2^2 &\geq \tfrac{2}{\sqrt{2\pi}}\langle \bar{x}, \widehat{e}\rangle\sigma - 2|\langle \bar{x}, z_i'\rangle| - 2|\langle \bar{x}, z_{i+1}'\rangle| - 4\Delta\sqrt{\log(mk)}\sigma^2 - \Delta^2\sigma^2 - \sigma^2 \\ &\geq \sqrt{\frac{d}{2\pi m}} \cdot \sigma^2 - 2\|x\|_2(\|z_i'\|_2 + \|z_{i+1}'\|_2) - 4\Delta^2\sigma^2 \\ &\geq \sigma^2\sqrt{\frac{d}{2\pi m}} - \sigma^2\sqrt{\frac{d}{9m}} - 4\sigma^2\Delta^2 > 0,\end{aligned}$$

since $m \leq cd/\Delta^4$ for some appropriate constant c (say $c = 16\pi$). □

Proof of Theorem 4.1. Let $C'_1, \ldots, C'_k$ be a locally optimal k-means clustering of $\mathcal{X}$, and suppose $\sum_i |C'_i \triangle C_i| < mk/2$ (for sake of contradiction). For each $i \in [k]$, let $\tilde{\mu}_i$ be the empirical mean of C_i and μ'_i be the empirical mean of C'_i. Since $C'_1, \ldots, C'_k$ is a locally optimal clustering, the Voronoi partition given by $\mu'_1, \ldots, \mu'_k$ classifies all the points in agreement with $C'_1, \ldots, C'_k$.

We will now contradict the local optimality of the clustering $C'_1, \ldots, C'_k$. Every cluster C_i has at least $N/(2k)$ points w.h.p. Hence, for each $i \in [k]$, from Lemma 4.6 we have

$$\begin{aligned}\|\mu'_i - \tilde{\mu}_i\|_2 &\leq \sigma(\sqrt{d} + 2\sqrt{\log N} + \Delta) \cdot \frac{4|C_i \triangle C'_i|}{\frac{N}{2k}} \leq \sigma \cdot \frac{8k^2 m(\sqrt{d} + 2\sqrt{\log N} + \Delta)}{N} \\ &\leq \frac{\sigma}{16\sqrt{m}(1 + \sqrt{(\log N)/d})}.\end{aligned}$$

However, from Lemma 4.7, every point in $\cup_{i\in[k]} Z_i$ is misclassified by $\mu'_1, \mu'_2, \ldots, \mu'_k$, i.e., the clustering given the Voronoi partition around $\mu'_1, \ldots, \mu'_k$ differs from $C_1, \ldots, C_k$ on at least $mk/2$ points in total. But $\sum_{i\in[k]} |C'_i \triangle C_i| < mk/2$. Hence, this contradicts the local optimality of the clustering $C'_1, \ldots, C'_k$. □

5 Conclusion

In this work we initiated the study of clustering data from a semi-random mixture of Gaussians. We proved that the popular Lloyd's algorithm achieves near optimal error. The robustness of the Lloyd's algorithm for the semi-random model suggests a theoretical justification for its widely documented success in practice. Similar robust analysis under stronger adversaries or for related heuristics such as the EM algorithm will significantly improve the gap between our theoretical understanding and observed practical performance of these algorithms. It would also be interesting to extend our results to study semi-random variants for other statistical models that are popular in machine learning.

References

[ABG+14] Joseph Anderson, Mikhail Belkin, Navin Goyal, Luis Rademacher, and James R. Voss. The more, the merrier: the blessing of dimensionality for learning large Gaussian mixtures. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 1135–1164, 2014.

[ACKS15] Pranjal Awasthi, Moses Charikar, Ravishankar Krishnaswamy, and Ali Kemal Sinop. The hardness of approximation of euclidean k-means. *arXiv preprint arXiv:1502.03316*, 2015.

[AK01] Sanjeev Arora and Ravi Kannan. Learning mixtures of arbitrary Gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257. ACM, 2001.

[AM05] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *Learning Theory*, pages 458–469. Springer, 2005.

[ANSW16] Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for k-means and euclidean k-median by primal-dual algorithms. *CoRR*, abs/1612.07925, 2016.

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

[AV05] David Arthur and Sergei Vassilvitskii. On the worst case complexity of the k-means method. Technical report, Stanford, 2005.

[AV07] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, 2007.

[BCMV14] Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Proceedings of the 46th Symposium on Theory of Computing (STOC)*. ACM, 2014.

[BCV14] Aditya Bhaskara, Moses Charikar, and Aravindan Vijayaraghavan. Uniqueness of tensor decompositions with applications to polynomial identifiability. *Proceedings of the Conference on Learning Theory (COLT).*, 2014.

[Bru09] S. Charles Brubaker. Robust PCA and clustering in noisy mixtures. In *Proceedings of the Symposium on Discrete Algorithms*, pages 1078–1087, 2009.

[BS95] Avrim Blum and Joel Spencer. Coloring random and semi-random k-colorable graphs. *J. Algorithms*, 19:204–234, September 1995.

[BS10] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 103–112. IEEE, 2010.

[BV08] Spencer Charles Brubaker and Santosh Vempala. Isotropic pca and affine-invariant clustering. In *Proceedings of the 2008 49th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '08, pages 551–560, Washington, DC, USA, 2008. IEEE Computer Society.

[BWY14] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *CoRR*, abs/1408.2156, 2014.

[CSV17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pages 47–60, New York, NY, USA, 2017. ACM.

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

[DF16] Roee David and Uriel Feige. On the effect of randomness on planted 3-coloring models. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2016, pages 77–90, New York, NY, USA, 2016. ACM.

[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 *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 655–664, Oct 2016.

[DS07] Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *The Journal of Machine Learning Research*, 8:203–226, 2007.

[DTZ16] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of EM suffice for mixtures of two Gaussians. *CoRR*, abs/1609.00368, 2016.

[FK98] U. Feige and J. Kilian. Heuristics for finding large independent sets, with applications to coloring semi-random graphs. In *Foundations of Computer Science, 1998. Proceedings.39th Annual Symposium on*, pages 674 –683, nov 1998.

[GHK15] Rong Ge, Qingqing Huang, and Sham M. Kakade. Learning mixtures of Gaussians in high dimensions. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 761–770, 2015.

[GVX14] Navin Goyal, Santosh Vempala, and Ying Xiao. Fourier PCA and robust tensor decomposition. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 584–593, 2014.

[HK12] Daniel Hsu and Sham M Kakade. Learning Gaussian mixture models: Moment methods and spectral decompositions. *arXiv preprint arXiv:1206.5766*, 2012.

[HK13] Daniel Hsu and Sham M Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20. ACM, 2013.

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

[KMM11] Alexandra Kolla, Konstantin Makarychev, and Yury Makarychev. How to play unique games against a semi-random adversary. In *Proceedings of 52nd IEEE symposium on Foundations of Computer Science*, FOCS '11, 2011.

[KMN+02] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. A local search approximation algorithm for k-means clustering. In *Proceedings of the eighteenth annual symposium on Computational geometry*, pages 10–18. ACM, 2002.

[KMV10] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two Gaussians. In *Proceedings of the 42nd ACM symposium on Theory of computing*, pages 553–562. ACM, 2010.

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

[Llo82] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[LM00] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 10 2000.

[LRV16] K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674, Oct 2016.

[LSW17] Euiwoong Lee, Melanie Schmidt, and John Wright. Improved and simplified inapproximability for k-means. *Information Processing Letters*, 120:40–43, 2017.

[MMV12] Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Approximation algorithms for semi-random partitioning problems. In *Proceedings of the 44th Symposium on Theory of Computing (STOC)*, pages 367–384. ACM, 2012.

[MMV13] Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Sorting noisy data with partial information. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 515–528. ACM, 2013.

[MMV14] Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Constant factor approximations for balanced cut in the random pie model. In *Proceedings of the 46th Symposium on Theory of Computing (STOC)*. ACM, 2014.

[MMV15] Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Correlation clustering with noisy partial information. *Proceedings of the Conference on Learning Theory (COLT)*, 2015.

[MMV16] Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Learning communities in the presence of errors. *Proceedings of the Conference on Learning Theory (COLT)*, 2016.

[MPW15] Ankur Moitra, William Perry, and Alexander S. Wein. How robust are reconstruction thresholds for community detection. *CoRR*, abs/1511.01473, 2015.

[MS10] Claire Mathieu and Warren Schudy. Correlation clustering with noisy input. In *Proceedings of the Twenty-first Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10, pages 712–728, Philadelphia, PA, USA, 2010. Society for Industrial and Applied Mathematics.

[MV10] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.

[Pea94] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[RV17] Oded Regev and Aravindan Vijayaraghavan. Learning mixtures of well-separated gaussians. In *Proceedings of the 58th Annual IEEE Foundations of Computer Science (FOCS)*. IEEE, 2017.

[Tei61] Henry Teicher. Identifiability of mixtures. *The annals of Mathematical statistics*, 32(1):244–248, 1961.

[Tei67] Henry Teicher. Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics*, 38(4):1300–1302, 1967.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[VW04] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

[WS11] David P. Williamson and David B. Shmoys. *The Design of Approximation Algorithms*. Cambridge University Press, New York, NY, USA, 1st edition, 2011.

[XHM16] Ji Xu, Daniel J. Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two Gaussians. In *NIPS*, 2016.

A Standard Properties of Gaussians

Lemma A.1. *Suppose $x \in \mathbb{R}$ be generated according to $N(0, \sigma^2)$, let $\Phi(t)$ represent the probability that $x > t$, and let $\Phi^{-1}y$ represent the quantile t at which $\Phi(t) \leq y$. Then*

$$\frac{\frac{t}{\sigma}}{(\frac{t^2}{\sigma^2}+1)} e^{-\frac{t^2}{2\sigma^2}} \leq \Phi(t) \leq \frac{\sigma}{t} e^{-\frac{t^2}{2\sigma^2}}. \tag{16}$$

Further, there exists a universal constant $c \in (1, 4)$ such that

$$\frac{1}{c}\sqrt{\log(1/y)} \leq \frac{t}{\sigma} \leq c\sqrt{\log(1/y)}. \tag{17}$$

Let γ_d be the Gaussian measure associated with a standard Gaussian with mean 0 and variance 1 in each direction. We start with a simple fact about the probability mass of high-dimensional spherical Gaussians being concentrated at around $\sqrt{d}\sigma$.

Using concentration bounds for the χ^2 random variables, we have the following bounds for the lengths of vectors picked according to a standard Gaussian in d dimensions (see (4.3) in [LM00]).

Lemma A.2. *For a standard Gaussian in d dimensions (mean 0 and variance σ^2 in each direction), and any $t > 0$*

$$\mathop{\mathbb{P}}_{x\sim\gamma_d}\left[\|x\|^2 \geq \sigma^2(d + 2\sqrt{dt} + 2t)\right] \leq e^{-t}.$$
$$\mathop{\mathbb{P}}_{x\sim\gamma_d}\left[\|x\|^2 \leq \sigma^2(d - 2\sqrt{dt})\right] \leq e^{-t}.$$

The following lemma follows from Lemma A.2 and a simple coupling to a spherical Gaussian with variance $\sigma^2 I$.

Lemma A.3. *Consider any points $y^{(1)}, \ldots, y^{(N)}$ drawn from a Gaussian with mean 0 and variance at most σ^2 in each direction. Then with high probability we have*

$$\forall \ell \in [N], \|y^{(\ell)}\|_2 \leq \sigma(\sqrt{d} + 2\sqrt{\log N}).$$

Proof. Consider a random vector $z \in \mathbb{R}^d$ generated from a Gaussian with mean 0 and variance σ^2 in each direction. From Lemma A.2,

$$Pr[\|z\|_2 \geq \sigma(\sqrt{d}+2\sqrt{\log N})] = Pr[\|z\|_2^2 \geq \sigma^2(d+4\sqrt{d\log N}+4\log N)] \leq \exp(-2\log N) < N^{-2}.$$

Fix $\ell \in [N]$. By a simple coupling to the spherical Gaussian random variable z we have

$$Pr[\|y^{(\ell)}\|_2 \geq \sigma(\sqrt{d} + 2\sqrt{\log N})] \leq Pr[\|z\| \geq \sigma(\sqrt{d} + 2\sqrt{\log N})] < N^{-2}.$$

By a union bound over all $\ell \in [N]$, the lemma follows. ∎

Lemma A.4 ([Ver10], Proposition 5.10)**.** *Let $Y_i \sim N(\mu, \sigma^2 I_{d\times d})$ for $i = 1, 2, \ldots N$ where $N = \Omega(\frac{d+\log(\frac{1}{\delta})}{\varepsilon^2})$. Then, with probability at least $1-\delta$ we have that*

$$\|\frac{1}{N}\sum_{i=1}^{N} Y_i - \mu\|_2 \leq \sigma\varepsilon.$$

Lemma A.5 ([Ver10], Corollary 5.50)**.** *Let $Y_i \sim N(\mu, \sigma^2 I_{d\times d})$ for $i = 1, 2, \ldots N$ where $N = \Omega(\frac{d+\log(\frac{1}{\delta})}{\varepsilon^2})$. Then, with probability at least $1-\delta$ we have that*

$$\|\frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu)(Y_i - \mu)^T - \sigma^2 I\| \leq \sigma\varepsilon.$$

B Proof of Proposition 3.2

The proof will follow the outline in [KK10]. Given N points from a semi random mixture $\mathcal{X}$, we first randomly partition them into two sets S_1 and S_2 of equal size. Let $T_1, \ldots, T_k$ be the partition induced by the true clustering over S_1 and $T'_1, \ldots, T'_k$ be the partition induced over S_2. Furthermore, let A be the $\frac{N}{2} \times d$ matrix consisting of points in S_1 as rows and C be the $\frac{N}{2} \times d$ matrix of the corresponding true centers. It is easy to see that with probability at least $1-\delta$, we will have that

$$\forall r \in [k], \min(|T_r|, |T'_r|) \geq \frac{|C_r|}{4}. \tag{18}$$

Assuming that equations (1) to (8) hold with high probability, we next prove that the following conditions will also hold with high probability

$$\max_{v:\|v\|=1} \frac{1}{|T_r|}\sum_{x\in T_r}[(x-\mu_r)\cdot v]^2 \leq 4\sigma^2, \forall r \in [k] \tag{19}$$

$$\max_{v:\|v\|=1} \frac{1}{|T'_r|}\sum_{x\in T'_r}[(x-\mu_r)\cdot v]^2 \leq 4\sigma^2, \forall r \in [k] \tag{20}$$

$$\|\frac{1}{|T_r|}\sum_{x\in T_r} x - \mu_i\| \leq 8\sigma, \forall r \in [k] \tag{21}$$

$$\|A - C\|^2 \leq 4\sigma^2 N \tag{22}$$

To prove (19) notice that $\frac{1}{|T_r|}\sum_{x\in T_r}[(x-\mu_r)\cdot v]^2 \leq \frac{4}{|C_r|}\sum_{x\in C_r}[(x-\mu_r)\cdot v]^2 \leq 16\sigma^2$ (using 5). Similarly, (20) follows. The proof of (21) follows directly from (6). Finally notice that $\|A-C\|^2 = \max_{v:\|v\|=1}\sum_r\sum_{x\in T_r}[(x-\mu_r)\cdot v]^2 \leq \max_{v:\|v\|=1}\sum_r\sum_{x\in C_r}[(x-\mu_r)\cdot v]^2 \leq 4\sigma^2 N$ (using (7)).

In the analysis below we will assume that the above equations are satisfied by the random partition. Define a graph $G = (A \cup B, E)$ where the edge set consists of any pair of points that have a distance of at most $\gamma = 4\sigma(\sqrt{d} + \sqrt{\log N})$. Notice that from the definition of γ, any two points from the same true cluster C_r will be connected by an edge in G (using 1). Next we map the points in A to a new $\frac{N}{2}$ dimensional space as follows. For any row A_i of A define $A'_{i,j} = (A_i - \mu)\cdot(B_j - \mu)$ if A_i and B_j are in the same connected component of G. Otherwise, define $A'_{i,j} = L$ where L is a large quantity. Here μ denotes the mean of the points in the component in G to which A_i belongs to. Let θ_r denote the mean of the points in T_r in the new space. We will show that the new mapping amplifies the mean separation.

Lemma B.1. *For all $r \neq s$, $\|\theta_r - \theta_s\| \geq \Omega(\sqrt{|Nw_{min}|}k\log N)\sigma^2$.*

Proof. We can assume that points in T_r, T'_r and T_s, T'_s belong to the same connected component in G. Otherwise, $\|\theta_r - \theta_S\| > L$. Let Q be the component to which T_r and T_s belong with μ being the mean of the points in Q. Then, $\|\theta_r - \theta_s\|^2 \geq \sum_{B_j \in Q}[(\mu_r - \mu_s) \cdot (B_j - \mu)]^2$. Notice that $(\mu_r - \mu_s) \cdot (\mu_r - \mu_s) = (\mu_r - \mu) \cdot (\mu_r - \mu_s) - (\mu_s - \mu) \cdot (\mu_r - \mu_s)$. Hence, one of the two terms is at least $\frac{1}{2}\|\mu_r - \mu_s\|^2$ in magnitude. Without loss of generality assume that $|(\mu_r - \mu) \cdot (\mu_r - \mu_s)| \geq \frac{1}{2}\|\mu_r - \mu_s\|^2 \geq \frac{125^2}{2} k \log N$.

Now, $\|\theta_r - \theta_s\|^2 \geq \sum_{B_j \in T'_r}[(\mu_r - \mu_s) \cdot (B_j - \mu)]^2 = \sum_{B_j \in T'_r}[(\mu_r - \mu_s) \cdot (\mu_r - \mu) - (\mu_r - \mu_s) \cdot (\mu_r - B_j)]^2 \geq \frac{1}{2}|B_j|[(\mu_r - \mu_s) \cdot (\mu_r - \mu)]^2 - \sum_{B_j \in T'_r}[(\mu_r - \mu_s) \cdot (\mu_r - B_j)]^2$. The first term is at least $\frac{|T'_r|}{8}\|\mu_r - \mu_s\|^4$ and the second term (in magnitude) is at most $4|T'_r|\|\mu_r - \mu_s\|^2\sigma^2$ (using 20). Substituting the bound on $\|\mu_r - \mu_s\|$ and using 18, we get that $\|\theta_r - \theta_s\| = \Omega(\sqrt{|N w_{\min}|})(k \log N)\sigma^2$. ☐

Let A' be the matrix of points in the new space and C' be the matrix of the corresponding centers. We next bound $\|A' - C'\|$.

Lemma B.2. $\|A' - C'\| \leq 24\sigma^2 k(\sqrt{d} + 2\sqrt{\log N})\sqrt{N}$.

Proof. Let $Y = A' - C'$. Then we have that $\|Y\|^2 \leq \|Y^T Y\| = \max_{v:\|v\|=1} \sum_r \sum_{x \in T_r}[(x - \theta_r) \cdot v]^2$. Let Q_r be the connected component in G that the points in T_r belong to. Then we can write $\|Y^T Y\| = \max_{v:\|v\|=1} \sum_r \sum_{x \in T_r} \sum_{B_j \in Q_r} v_j^2[(x - \mu_r) \cdot (B_j - \mu)]^2 \leq \sum_r \sum_{B_j \in Q_r} v_j^2 \sum_{x \in T_r}[(x - \mu_r) \cdot (B_j - \mu)]^2$. Using 19, we can bound the inner term as $\sum_{x \in T_r}[(x - \mu_r) \cdot (B_j - \mu)]^2 \leq 4|T_r|\|B_j - \mu\|^2\sigma^2$.

Next notice that because of the way G is constructed, points within the same connected component have distance at most $k\gamma$. Hence, $\|B_j - \mu\| \leq k\gamma$. Hence, $\|Y^T Y\| \leq \sum_r \sum_{B_j \in Q_r} v_j^2 4|T_r|(k^2\gamma^2)\sigma^2 \leq 4Nk^2\gamma^2\sigma^2$. This gives the desired bound on $\|Y\| = \|A' - C'\|$. ☐

Combining the previous two lemmas we get that $\|\theta_r - \theta_s\| \geq \Omega(\sqrt{\frac{|N w_{\min}|}{d}})\frac{\|A' - C'\|}{\sqrt{N}}$. We next run the initialization procedure from Section 3.2 by projecting A' onto the top k subspace and running a k-means approximation algorithm. Let $\phi_1, \ldots \phi_k$ be the means obtained. Using Lemma 3.5 with $M^* = C'$, we get that for all r, $\|\phi_r - \theta_r\| \leq 20\sqrt{k\alpha}\frac{\|A' - C'\|}{\sqrt{|N w_{\min}|}}$, where α is the approximation guarantee of the k-means approximation used. If $\Delta > c_0\sqrt{\min\{k, d\} \log N}$, then we use a constant factor approximation algorithm [ANSW16]. If $\Delta > c_0\sqrt{\min\{k, d\} \log k \log N}$, then we can use the simpler k-means++ algorithm [AV07]

Proof of Proposition 3.2. Assuming $N = \Omega(\frac{k^2 d^2}{w_{\min}^2})$ we get that for all $r \neq s$, $\|\phi_r - \phi_s\| \geq 10\sqrt{kd}\frac{\|A' - C'\|}{\sqrt{N w_{\min}}}$. Let $P_1, \ldots P_k$ be the clustering of points in A' obtained by using centers $\phi_1, \ldots \phi_k$. Then we have that for each r, $|T_r \triangle P_r| \leq \frac{N w_{\min}}{10\sqrt{d}]}$, since otherwise the total cost paid by the misclassified points will be more than $4k\|A' - C'\|^2$. Next we use the clustering $P_1, \ldots P_k$ to compute means for the original set of points in A. Let $\nu_1, \ldots \nu_k$ be the obtained means. We will show that for all r, $\|\nu_r - \mu_r\| \leq \tau\sigma$, where $\tau < \frac{\Delta}{4}$.

Consider a particular partition P_r that is uniquely identified with T_r. Let $n_{r,r}$ be the number of points that belong to both P_r and T_r and $\mu_{r,r}$ be the mean of those points. Similarly, let $n_{r,s}$ be the number of points that belong to T_s originally but belong to P_r in the current clustering, and let $\mu_{r,s}$ be their mean. Then, $\|\mu_r - \nu_r\| \leq \frac{n_{r,r}}{|P_r|}\|\mu_{r,r} - \mu_r\| + \sum_{s \neq r} \frac{n_{r,s}}{|P_r|}\|\mu_{r,s} - \mu_r\|$. We can bound $\|\mu_{r,r} - \mu_r\|$ by $O(\sigma)$ using 6 and $\|\mu_{r,s} - \mu_s\|$ by $O(k(\sqrt{d} + 2\sqrt{\log N})$ using 1 and the fact that points in r and s must belong to the same component in G. Combining we get the claim. ☐