Clustering Semi-Random Mixtures of Gaussians

Pranjal Awasthi [1] **Aravindan Vijayaraghavan** [2]

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 robust model for k-means clustering that generalizes the Gaussian mixture model, and that we believe will be useful in identifying robust algorithms. 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 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 (Williamson & Shmoys, 2011) (in fact, it is also NP-hard to find near-optimal solutions (Awasthi et al., 2015; Lee et al., 2017)). 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 (Teicher, 1961; Pearson, 1894; Dasgupta, 1999; Arora & Kannan, 2001; Vempala & Wang, 2004; Dasgupta & Schulman, 2007).

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 (Dasgupta, 1999), there have been a variety of algorithms to provably cluster data from a GMM model. Algorithms based on PCA and distance-based clustering (Arora & Kannan, 2001; Vempala & Wang, 2004; Achlioptas & McSherry, 2005; Kannan et al., 2008) provably recover the clustering when there is adequate separation between every pair of components (parameters)[2]. Other algorithmic approaches include the method-of-moments (Moitra & Valiant, 2010; Belkin & Sinha, 2010), and algebraic methods based on tensor decompositions (Hsu & Kakade, 2012).(Please see Section 1 in Supplementary material for a more detailed comparison of related work like (Tang & Monteleoni, 2016; Dutta et al., 2017) and guarantees).

On the other hand, the method-of-choice in practice are iterative algorithms like the Lloyd's algorithm (also called k-means algorithm) (Lloyd, 1982) and the k-means++ algorithm (Lloyd's algorithm initialized with centers from distance-based D^2-sampling). D^2-sampling based initialization schemes were first theoretically analyzed in (Ostrovsky et al., 2006; Arthur & Vassilvitskii, 2007). In the absence of good worst-case guarantees, a compelling direction is to use

[*]Equal contribution [1]Department of Computer Science, Rutgers University, USA. [2]EECS Department, Northwestern University, USA. Correspondence to: Pranjal Awasthi <pranjal.awasthi@rutgers.edu>, Aravindan Vijayaraghavan <aravindv@northwestern.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

[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.

[2]The best known guarantees along these lines for non-spherical Gaussians (Kumar & Kannan, 2010; Awasthi & Sheffet, 2012) 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.

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 (Ostrovsky et al., 2006; Dasgupta & Schulman, 2007; Kumar & Kannan, 2010; Awasthi & Sheffet, 2012).

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 robust generalization of the GMM model that we call *semi-random model*.

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 (Blum & Spencer, 1995; Feige & Kilian, 1998; Makarychev et al., 2012) (see Section 1 of Supplementary material). 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 (1998) 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.

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 like 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 model.

Definition 1.1 (Semi-random GMM model)**.** Given 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 $\mathcal{C} = (C_1, \ldots, C_k)$ of $\{1, \ldots, N\}$, let $N_i = |C_i| \; \forall i \in [k]$.
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 $\{\mu_i + \lambda(y^{(t)} - \mu_i) : \lambda \in [0,1]\}$ (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 $w_{\min} = \min_{i \in [k]} N_i/N$.

It is necessary that each point is moved closer to its mean *along the direction* of the mean – otherwise, one can move points closer to its own mean, but in other directions in such a way that the optimal k-means clustering of the perturbed instance is very different from the planted clustering. This is especially true in the separation range of interest (when $k \ll d$), where the inter-mean distance is smaller than the average radius of the clusters (see Supplementary material for details).

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 (Kumar & Kannan, 2010).

Informal Theorem 1.2. *Consider any semi-random instance $\mathcal{X}$ with N points generated by the semi-random GMM model (Def. 1.1) with planted clustering $C_1, \ldots, C_k$ and parameters $\mu_1, \ldots, \mu_k, \sigma^2$ satisfying $\forall i \neq j \in [k]$, $\|\mu_i - \mu_j\|_2 > \Delta\sigma$ where $\Delta \geq c_0\sqrt{\min\{k, d\}\log N}$ and $N \geq k^2d^2/w_{min}^2$. 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, the initialization procedure of Kumar and Kannan (2010) can be replaced by a simpler procedure based on the popular k-means++ algorithm (Ostrovsky et al., 2006; Arthur & Vassilvitskii, 2007). On the other hand, certain other algorithmic techniques like moment-based methods, distance-based clustering and tensor decompositions seem less robust to semi-random perturbations (please see related work in Section 1 of supplementary material for more details). The most closely related to our work is that of (Kumar & Kannan, 2010) and (Awasthi & Sheffet, 2012) 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 (Awasthi & Sheffet, 2012) for Lloyd's algorithm even in the absence of any semi-random errors or perturbations [3]. We also remark 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 see Section 1 of supplementary material 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 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. 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 of Supplementary material for details). This robustness of the Lloyd's algorithm suggests an explanation for its widely documented empirical success across different application domains.

Challenges and Overview of Techniques. Lloyd's algorithm has been analyzed in the context of clustering mixtures of Gaussians (Kumar & Kannan, 2010; Awasthi & Sheffet, 2012). 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 (2010) 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 (Awasthi & Sheffet, 2012) leads to state of the art results that perfectly recovers all the clusters under a separation of order $\sqrt{k\log N}\sigma$.

We analyze the variant of Lloyd's algorithm that was introduced by Kumar and Kannan (2010); further, our analysis also extends to a simpler initialization procedure based on the popular k-means++ algorithm. However, there are several challenges in extending the analysis of (Awasthi & Sheffet, 2012) to the semi-random setting. While the

[3] We note that for clustering GMMs, the work of Brubaker and Vempala (2008) give a qualitatively different separation condition that does not depend on the maximum variance. However this separation condition is incomparable to (Awasthi & Sheffet, 2012), because of the potentially worse dependence on k.

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 empirical success of Lloyd's algorithm on real-world data is widely documented. The main contribution of the paper is a robust theoretical 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 $\tilde{O}(kd\tau^2/\Delta^4)$. To analyze this, we need to understand the projection of points onto the direction of the line joining the centers. While the average projection is small i.e. $\widetilde{O}(1)$, there are certain directions where it is as large as $\Omega(\sqrt{d})$! However, we can still prove an upper bound (in Lemma 3.5) on the number of points x in a cluster C_i s.t. 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 (in Proposition 3.2 and Lemma 3.4).

Some Related Work. There has been a long line of algorithmic results on Gaussian mixture models starting from (Teicher, 1961; 1967; Pearson, 1894). Considering the vast and rich literature on mixtures of Gaussians, we defer much of the comparison to the related work section in the supplementary material (Section 1). Here we mention and compare to some recent related work on robust parameter estimation for Gaussians and related models. A recent exciting line of work concerns designing robust high-dimensional estimators of the mean and covariance of Gaussians (and models) when a small ε fraction of the points are adversarially corrupted (Brubaker, 2009; Diakonikolas et al., 2016; Lai et al., 2016; Charikar et al., 2017). However, this model and results are incomparable to our semi-random model — for example, they typically assume that only a $o(1/k)$ fraction of the points are corrupted, while potentially all the points could be perturbed in our semi-random model, while on the other hand, our work does not handle arbitrary outliers. Please see Section 1 of the supplementary material for additional related work.

2. Preliminaries

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 2.1. (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 2.2. (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 $\|x^{(t)} - \mu(C_i)\| \leq \|x^{(t)} - \mu(C_j)\|$. 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 (Lloyd, 1982) 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 (Arthur & Vassilvitskii, 2005). 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 (Arthur & Vassilvitskii, 2007). There also exist more sophisticated constant factor approximation algorithms for the k-means problem (Kanungo et al., 2002; Ahmadian et al., 2016) that can be used for seeding in our framework.

We now state several properties of semi-random mixtures that will be used throughout the analysis in the subsequent sections (please see supplementary material for proofs). We first start with a few simple properties that follow directly from the corresponding properties about high dimensional Gaussians.

Lemma 2.3. *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 w.h.p. $\forall i \in [k]$,

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

Further, for a fixed unit vector $u \in \mathbb{R}^d$, with probability at least $(1 - 1/(N^3))$

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

Finally, we have w.h.p. $\forall i \in [k]$,

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

We observe that (3) immediately follows from (2) after a union bound over the $k^2 < N^2$ directions given by the unit vectors along $(\mu_i - \mu_j)$ directions. The proofs of these statements are given in Section 2 of the Supplementary material. The next lemma gives an upper bound on how far the mean and variance of the points in a component of a semi-random GMM can move away from the true parameters.

Lemma 2.4. *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 $1 - \delta$,*

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

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

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 (Def. 1.1) with parameters $\mu_1, \ldots, \mu_k, \sigma^2$ s.t.*

$$\forall i \neq j \in [k], \ \|\mu_i - \mu_j\|_2 > \Delta\sigma, \quad \textit{where} \quad (6)$$

$\Delta > c_0\sqrt{\min\{k, d\}\log N}$ and $N \geq k^2 d^2 / w_{min}^2$, finds w.h.p. a clustering $C'_1, \ldots, C'_k$ such that

$$\min_{\pi} \sum_{i=1}^{k} |C_{\pi(i)} \triangle C'_i| \leq \frac{c_1 kd}{\Delta^4} \max\left\{1, \log\left(\tfrac{6(\sqrt{d} + \log^{\frac{1}{2}} N)}{\Delta^2}\right)\right\}$$

Algorithm 1 Lloyd's Algorithm

Input: 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.

Use these k-centers to seed a series of Lloyd-type iterations i.e.,

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 $\mu_r^{(i)} \leftarrow \frac{1}{|Z_i|} \sum_{A_j \in Z_i} A_j$.

end for

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 (2010). The initialization algorithm first performs a k-SVD of the data matrix followed by running the k-means++ algorithm that uses D^2-sampling to compute seed centers (Arthur & Vassilvitskii, 2007). One can also use any constant factor approximation algorithm for k-means clustering in the projected space to obtain the initial centers (Kanungo et al., 2002; Ahmadian et al., 2016). 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 (7). These equations are shown to hold w.h.p. in Section 2 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)-(7) 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)-(7) *and* (6) *and suppose we are given $\mu'_1, \dots, \mu'_k$ satisfying $\forall j \in [k],\ \|\mu'_j - \mu_j\|_2 \le \tau\sigma$ and $\tau < \Delta/24$. Then there exists a set $Z \subset \mathcal{X}$ such that for any $i \in [k]$ we have*

$$\forall x \in C_i \cap (\mathcal{X} \setminus Z),\ \ \|x - \mu'_i\|_2^2 \le \min_{j \neq i} \|x - \mu'_j\|_2^2 \text{ where}$$

$$|Z| = O\Big(\frac{d\tau^2}{\Delta^4} \cdot \max\Big\{1, \log\big(\tfrac{3\tau(\sqrt{d}+2\sqrt{\log N})}{\Delta^2}\big)\Big\}\Big).$$

The following lemma quantifies the improvement in each step of 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 $w_i N \ge \frac{d\sqrt{d}}{4\log(d)}$ for all i satisfying* (1)-(7)*. Furthermore, suppose we are given centers $\mu'_1, \dots, \mu'_k$ such that $\|\mu'_i - \mu_i\| \le \tau\sigma,\ \forall i \in [k]$ where $\tau < \Delta/24$. Then the centers $\mu''_1, \dots, \mu''_k$ obtained after one Lloyd's update satisfy $\|\mu''_i - \mu_i\| \le \max((6 + \frac{\tau}{4})\sigma, \frac{\tau}{2}\sigma))$ for all $i \in [k]$.*

We now present the proof of Theorem 3.1.

Proof of Theorem 3.1. Firstly, the eight deterministic conditions (1)-(7) are shown to hold for instance $\mathcal{X}$ w.h.p. in Section 2. The proof follows in a straightforward manner by combining Proposition 3.2, Lemma 3.4 and Lemma 3.3. Proposition 3.2 shows that $\|\mu_i^{(0)} - \mu_i\|_2 \le \Delta/(24)$ for all $i \in [k]$. Applying Lemma 3.4, we have that after $T = O(\log \Delta)$ iterations we get $\|\mu_i^{(T)} - \mu_i\|_2 \le 8\sigma$ 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

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 3.5 (Points in Bad Directions)**.** *Consider any semi-random instance $\mathcal{X} = \{x^{(1)}, \dots, x^{(N)}\}$ with N points having parameters $\mu_1, \dots, \mu_k, \sigma^2$ and planted clustering $C_1, \dots, C_k$, and denote by $\bar{x}^{(\ell)} = x^{(\ell)} - \mu_i\ \forall i \in [k], \ell \in C_i$. Then $\exists c > 0$ (universal constant) s.t. $\forall \lambda > 100\sqrt{\log N}$, w.p. at least $1 - 2^{-d}$*

$$\forall \widehat{e} : \|\widehat{e}\|_2 = 1, \Big|\{\, \ell \in [N] : |\langle \bar{x}^{(\ell)}, \widehat{e}\rangle| > \lambda\sigma \,\}\Big| \le \frac{cd}{\lambda^2} \cdot \max\Big\{1, \log\big(\tfrac{3(\sqrt{d}+\log^{\frac{1}{2}} N)}{\lambda}\big)\Big\} \quad (7)$$

Proof. Set $\eta := \min\{\lambda/(2\sqrt{d} + 2\sqrt{\log N}), \frac{1}{2}\}$ and $m := 512 d \log(3/\eta)/\lambda^2$. 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 $u = u(\widehat{e}) \in \mathcal{N}$

$$|\langle \bar{x}, u\rangle| > |\langle \bar{x}, \widehat{e}\rangle + \langle \bar{x}, \widehat{e} - u\rangle| \ge \sigma\lambda - \|\bar{x}\|_2 \|\widehat{e} - u\|_2 \ge \sigma\big(\lambda - \eta(\sqrt{d} + 2\sqrt{\log N})\big) \ge \frac{\lambda}{2}, \quad (8)$$

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

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

The probability that m points in $\{\, x^{(1)}, \dots, x^{(N)} \,\}$ satisfy (8) 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 (8).

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

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

□

Lemma 3.3 and Lemma 3.4 use the above lemma to analyze each iteration of the Lloyd's algorithm and show that we make progress in each step by misclassifying fewer points with successive iterations.

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 3.6. *In the above notation, for any $\delta > 0$ suppose we are given an instance $\mathcal{X}$ on N points satisfying satisfying (1)-(7), with components $C_1, \dots C_k$ 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 $\|\mu_i - \mu_i'\| \leq 20\sqrt{k\alpha}\frac{\|A-M^*\|}{\sqrt{N w_{min}}}$.*

The above proof 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 (Ahmadian et al., 2016). 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 (Arthur & Vassilvitskii, 2007). The above lemma has a bad dependence on $w_{\min}$. However, using the Boosting technique of (Kumar & Kannan, 2010) 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 $C_1, \dots, C_k$ generated by applying semi-random perturbations to points generated from a mixture of spherical Gaussians with means $\mu_1, \mu_2, \dots, \mu_k$, 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 $C_1', C_2', \dots, C_k'$ 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 $\mathbb{E}[h(y)] = \mu + \frac{1}{\sqrt{2\pi}}\sigma\widehat{e}$.

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 $\{ (C_1, C_2), (C_3, C_4), \dots, (C_{k-1}, C_k) \}$, and we will ensure that m points are misclassified in each of the $k/2$ clusters $C_1, C_3, \dots, C_{k-1}$. The parameters of the mixture of spherical Gaussians $\mathcal{G}$ are set up as follows. For each $i \in 1, 3, 5, \dots, k-1$, $\|\mu_i - \mu_{i+1}\|_2 = \Delta\sigma$, and all the other inter-mean distances (across different pairs) are at least $M\sigma$ which is arbitrarily large ($M \mapsto \infty$).

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

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

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 4.4 (using Lemma 4.3) and Lemma 4.5 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 4.6 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 $1 - 1/\mathrm{poly}(m, k)$).

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

Lemma 4.4. *There exists a constant $c_3 > 0$ s.t. for the semi-random instance $\mathcal{X}$ described above, we have w.h.p.*

for some z_i with $\|z_i\|_2 \le c_3\sigma\sqrt{dk/N}$

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

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

Lemma 4.5. *Consider any cluster C_i of the instance $\mathcal{X}$, and let C_i' satisfy $|C_i' \triangle C_i| \le m'$. Suppose $\tilde{\mu}_i, \mu_i'$ are the means of clusters C_i and C_i' respectively, then*

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

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

Lemma 4.6. *Let $\mu_1', \mu_2', \ldots, \mu_k'$ satisfy $\|\mu_i' - \tilde{\mu}_i\|_2 \le \sigma/(16\sqrt{m}(1 + 2\sqrt{\frac{\log N}{d}}))$, 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 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 4.5 we have

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

However, from Lemma 4.6, 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. A concrete open question left from our work is to extend our lower bound for locally optimal clusterings to a more general statistical lower bound – this would also imply a separation between recovery guarantees for the semi-random model and the pure GMM model. Robust analysis under semi-random adversaries for related heuristics such as the EM algorithm and studying semi-random variants for other popular statistical models in machine learning will further improve the gap between our theoretical understanding and observed practical performance of algorithms for such models.

Acknowledgements

Aravindan Vijayaraghavan is supported by the National Science Foundation (NSF) under Grant No. CCF-1652491 and CCF-1637585.

References

Achlioptas, D. and McSherry, F. On spectral learning of mixtures of distributions. In *Learning Theory*, pp. 458–469. Springer, 2005.

Ahmadian, S., Norouzi-Fard, A., Svensson, O., and Ward, J. Better guarantees for k-means and euclidean k-median by primal-dual algorithms. *CoRR*, abs/1612.07925, 2016. URL http://arxiv.org/abs/1612.07925.

Arora, S. and Kannan, R. Learning mixtures of arbitrary Gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pp. 247–257. ACM, 2001.

Arthur, D. and Vassilvitskii, S. On the worst case complexity of the k-means method. Technical report, Stanford, 2005.

Arthur, D. and Vassilvitskii, S. K-means++: The advantages of careful seeding. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pp. 1027–1035, 2007. ISBN 978-0-898716-24-5. URL http://dl.acm.org/citation.cfm?id=1283383.1283494.

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

Awasthi, P., Charikar, M., Krishnaswamy, R., and Sinop, A. K. The hardness of approximation of euclidean k-means. *arXiv preprint arXiv:1502.03316*, 2015.

Belkin, M. and Sinha, K. Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pp. 103–112. IEEE, 2010.

Blum, A. and Spencer, J. Coloring random and semi-random k-colorable graphs. *J. Algorithms*, 19:204–234, September 1995. ISSN 0196-6774. doi: http://dx.doi.org/10.1006/jagm.1995.1034. URL http://dx.doi.org/10.1006/jagm.1995.1034.

Brubaker, S. C. Robust PCA and clustering in noisy mixtures. In *Proceedings of the Symposium on Discrete Algorithms*, pp. 1078–1087, 2009.

Brubaker, S. C. and Vempala, S. Isotropic pca and affine-invariant clustering. In *Proceedings of the 2008 49th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '08, pp. 551–560, Washington, DC, USA, 2008. IEEE Computer Society. ISBN 978-0-7695-3436-7. doi: 10.1109/FOCS.2008.48. URL http://dx.doi.org/10.1109/FOCS.2008.48.

Charikar, M., Steinhardt, J., and Valiant, G. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pp. 47–60, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4528-6. doi: 10.1145/3055399.3055491. URL http://doi.acm.org/10.1145/3055399.3055491.

Dasgupta, S. Learning mixtures of Gaussians. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pp. 634–644. IEEE, 1999.

Dasgupta, S. and Schulman, L. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *The Journal of Machine Learning Research*, 8:203–226, 2007.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Robust estimators in high dimensions without the computational intractability. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 655–664, Oct 2016. doi: 10.1109/FOCS.2016.85.

Dutta, A., Vijayaraghavan, A., and Wang, A. Clustering stable instances of euclidean k-means. *Proceedings of Neural Information Processing Systems (NIPS).*, 2017.

Feige, U. and Kilian, J. Heuristics for finding large independent sets, with applications to coloring semi-random graphs. In *Foundations of Computer Science, 1998. Proceedings.39th Annual Symposium on*, pp. 674 –683, nov 1998. doi: 10.1109/SFCS.1998.743518.

Hsu, D. and Kakade, S. M. Learning Gaussian mixture models: Moment methods and spectral decompositions. *arXiv preprint arXiv:1206.5766*, 2012.

Kannan, R., Salmasian, H., and Vempala, S. The spectral method for general mixture models. *SIAM J. Comput.*, 38(3):1141–1156, 2008. doi: 10.1137/S0097539704445925. URL http://dx.doi.org/10.1137/S0097539704445925.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. A local search approximation algorithm for k-means clustering. In *Proceedings of the eighteenth annual symposium on Computational geometry*, pp. 10–18. ACM, 2002.

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

Lai, K. A., Rao, A. B., and Vempala, S. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 665–674, Oct 2016. doi: 10.1109/FOCS.2016.76.

Lee, E., Schmidt, M., and Wright, J. Improved and simplified inapproximability for k-means. *Information Processing Letters*, 120:40–43, 2017.

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

Makarychev, K., Makarychev, Y., and Vijayaraghavan, A. Approximation algorithms for semi-random partitioning problems. In *Proceedings of the 44th Symposium on Theory of Computing (STOC)*, pp. 367–384. ACM, 2012.

Moitra, A. and Valiant, G. Settling the polynomial learnability of mixtures of Gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pp. 93–102. IEEE, 2010.

Ostrovsky, R., Rabani, Y., Schulman, L. J., and Swamy, C. The effectiveness of lloyd-type methods for the k-means problem. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pp. 165–176. IEEE, 2006.

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

Tang, C. and Monteleoni, C. On lloyd's algorithm: New theoretical insights for clustering in practice. In *Artificial Intelligence and Statistics*, pp. 1280–1289, 2016.

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

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

Vempala, S. and Wang, G. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

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