

STATISTICAL AND COMPUTATIONAL GUARANTEES OF LLOYD'S ALGORITHM AND ITS VARIANTS

BY YU LU AND HARRISON H. ZHOU

Yale University

Clustering is a fundamental problem in statistics and machine learning. Lloyd's algorithm, proposed in 1957, is still possibly the most widely used clustering algorithm in practice due to its simplicity and empirical performance. However, there has been little theoretical investigation on the statistical and computational guarantees of Lloyd's algorithm. This paper is an attempt to bridge this gap between practice and theory. We investigate the performance of Lloyd's algorithm on clustering sub-Gaussian mixtures. Under an appropriate initialization for labels or centers, we show that Lloyd's algorithm converges to an exponentially small clustering error after an order of $\log n$ iterations, where n is the sample size. The error rate is shown to be minimax optimal. For the two-mixture case, we only require the initializer to be slightly better than random guess.

In addition, we extend the Lloyd's algorithm and its analysis to community detection and crowdsourcing, two problems that have received a lot of attention recently in statistics and machine learning. Two variants of Lloyd's algorithm are proposed respectively for community detection and crowdsourcing. On the theoretical side, we provide statistical and computational guarantees of the two algorithms, and the results improve upon some previous signal-to-noise ratio conditions in literature for both problems. Experimental results on simulated and real data sets demonstrate competitive performance of our algorithms to the state-of-the-art methods.

1. Introduction. Lloyd's algorithm, proposed in 1957 by Stuart Lloyd at Bell Labs [40], is still one of the most popular clustering algorithms used by practitioners, with a wide range of applications from computer vision [3], to astronomy [46] and to biology [26]. Although considerable innovations have been made on developing new provable and efficient clustering algorithms in the past six decades, Lloyd's algorithm has been consistently listed as one of the top ten data mining algorithms in several recent surveys [56].

Lloyd's algorithm is very simple and easy to implement. It starts with an initial estimate of centers or labels and then iteratively updates the labels

MSC 2010 subject classifications: Primary 62C20

Keywords and phrases: Lloyd's algorithm, nonconvex optimization, mixture of Gaussians, community detection, crowdsourcing

and the centers until convergence. Despite its simplicity and a wide range of successful applications, surprisingly, there is little theoretical analysis on explaining the effectiveness of Lloyd’s algorithm. It is well known that there are two issues with Lloyd’s algorithm under the worst case analysis. First, as a greedy algorithm, Lloyd’s algorithm is only guaranteed to converge to a local minimum [44]. The k -means objective function that Lloyd’s algorithm attempts to minimize is NP-hard [14, 42]. Second, the convergence rate of Lloyd’s algorithm can be very slow. Arthur and Vassilvitskii [6] construct a worst-case showing that Lloyd’s algorithm can require a superpolynomial running time.

A main goal of this paper is trying to bridge this gap between theory and practice of Lloyd’s algorithm. We analyze its performance on the Gaussian mixture model [48, 53], a standard model for clustering, and consider the generalization to sub-Gaussian mixtures, which includes binary observations as a special case. Specifically, we attempt to address following questions to help understand Lloyd’s algorithm: How good does the initializer need to be? How fast does the algorithm converge? What separation conditions do we need? What is the clustering error rate and how it is compared with the optimal statistical accuracy?

Despite the popularity of Lloyd’s algorithm as a standard procedure for the k -means problem, to the best of our knowledge, there is little work in statistics to understand the algorithm. Some efforts have been made by computer scientists to develop effective initialization techniques for Lloyd’s algorithm [7, 47, 4]. Their main focus is to find polynomial-time approximation scheme of the k -means objective function rather than to identify the cluster labels of data points, which is often the primary interest for many applications. It is worthwhile to emphasize that when the signal-to-noise ratio is not sufficiently large, a small k -means objective function value does not necessarily guarantee a small clustering error. Furthermore, the error rate is different from the optimal error of an exponential form. Recently, some fascinating results by Kumar and Kannan [37] and Awasthi and Sheffet [8] show that under certain strong separation conditions, the Lloyd’s algorithm initialized by spectral clustering correctly classifies all data points with high probability. However, they focus on the strong consistency results and the clustering error rate of the Lloyd’s algorithm remains unclear. It is desirable to have a systematic study of Lloyd’s algorithm under various separation conditions such that the strong consistency can be included as a special case.

Lloyd’s algorithm is an iterative procedure. Its analysis can be challenging due to dependence between iterative steps. In the statistics literature, vari-

ous two-stage estimators, or more precisely, two-step estimators, have been proposed to successfully solve some very important non-convex problems, for example, sparse principle analysis [10, 55], community detection [22, 23], mixture of linear regression [11, 59] and crowdsourcing [61, 21]. For all those problems, under a strong assumption that the initial estimator is consistent, one-step update in the second stage usually leads us to a minimax optimal estimator. However, as observed in various simulation or real data studies in [61] and [22], the initial estimator may perform poorly, and more iterations in the second stage keeps driving down the clustering error. Unfortunately, due to some technical difficulties, theoretical analyses in [61] and [22] restrict to one-step iteration.

1.1. *Our contributions.* In this paper, we give a considerably weak initialization condition under which Lloyd's algorithm converges to the optimal label estimators of sub-Gaussian mixture model. While previous results [37, 8] focus on exact recovery (strong consistency) of the labels, we obtain the clustering error rates of Lloyd's algorithm under various signal-to-noise levels. As a special case, we obtain exact recovery with high probability when the signal-to-noise level is bigger than $4 \log n$. The signal-to-noise ratio condition for exact recovery is weaker than the state-of-the-art result [8]. In contrast to previous two-stage (two-step) estimators, our analyses go beyond one-step update. We are able to show a linear convergence to the statistical optimal error rate for Lloyd's algorithms and its two variants for community detection and crowdsourcing.

We illustrate our contributions here by considering the problem of clustering two-component spherical Gaussian mixtures, with symmetric centers θ^* and $-\theta^* \in \mathbb{R}^d$ and variance σ^2 . See Section 2 for more details. Let n be the sample size and $r = \|\theta^*\|/(\sigma\sqrt{1+9d/n})$ be the normalized signal-to-noise ratio. We establish the following basin of attractions of Lloyd's algorithm.

THEOREM 1.1. *Assume $r \geq C$ and $n \geq C$ for a sufficiently large constant C . For symmetric, two-component spherical Gaussian mixtures, given any initial estimator of labels with clustering error*

$$A_0 < \frac{1}{2} - \frac{2.56 + \sqrt{\log r}}{r} - \frac{1}{\sqrt{n}}, \quad w.h.p.$$

Lloyd's algorithm converges linearly to an exponentially small rate after $\lceil 3 \log n \rceil$ iterations, which is the minimax rate as $r \rightarrow \infty$ w.h.p.

The results above are extended to general number of clusters k and to (non-spherical) sub-Gaussian distributions under an appropriate initializa-

tion condition and a signal-to-noise ratio condition, which, to the best of our knowledge, are the weakest conditions in literature.

Now we discuss the contributions of this paper in detail. Our contributions are three folds. First, we give statistical guarantees of Lloyd’s algorithm. Starting with constant clustering error, by an appropriate initializer, such as spectral clustering, we show an exponentially small clustering error rate of Lloyd’s algorithms, under a weak signal-to-noise ratio condition. We also provide a rate-matching lower bound to show that Lloyd’s algorithm initialized by spectral clustering is minimax optimal. When the clusters sizes are of the same order and the distance between different centers are of the same order, our signal-to-noise condition reduces to $\Delta \gtrsim \sigma k \sqrt{1 + kd/n}$, where Δ is the minimum Euclidean distances between two different cluster centers. Previous results on Lloyd’s algorithms focus on exact recovery of the labels and they assume sample size $n \gg kd$ [37, 8]. The best known signal-to-noise ratio condition is $\Delta \gtrsim \sigma k \text{polylog } n$ [8]. Our condition is weaker by a polylog n factor. Moreover, our results also hold for the high-dimensional case where d could be larger than n .

Second, we provide computational guarantees of Lloyd’s algorithm. We prove a linear convergence rate of Lloyd’s iterations given a label initializer of constant clustering error. The linear convergence rate depends on the signal-to-noise ratio. Larger signal-to-noise ratio leads to faster convergence rate. Counterexamples are constructed in the supplementary material to show that the basin of attractions we establish in Theorem 1.1 and in the general k case are almost necessary. It is worthwhile to point out that the initialization condition in Theorem 1.1 is just slightly stronger than random guess. As implied by the minimax lower bound (Theorem 3.3), a necessary condition for consistently estimating the labels is $\frac{\Delta}{\sigma} \rightarrow \infty$. In addition, when $\Delta \gtrsim \sigma d^{1/4}$, we are able to prove a random initialization scheme works with high probability.

Third, we develop new proof techniques for analyzing two-stage estimators. Previous analyses usually require the first stage’s estimator to be consistent to achieve the minimax optimal rates [61, 22]. In contrast, we only need a constant clustering error of the first stage estimator by introducing a new technique to analyze random indicator functions. In addition, we are able to improve previous signal-to-noise ratio conditions in community detection and crowdsourcing by considering two variants of Lloyd’s algorithms. Simulated and real data experiments show our algorithms are competitive as compared to the state-of-the-art algorithms for both problems.

1.2. *Related work on Gaussian mixture models.* The study of Gaussian mixture model [48] has a long and rich history. We give a very brief review here. In the original paper of Pearson [48], methods of moments estimators were first proposed, followed by the work of [18, 39]. They all involve solving high-order moments equations, which is computationally challenging as the sample size or the dimensionality grows. Recently, third-order moments was proved to be sufficient by using tensor decomposition techniques [12, 5, 28]. However, their sample size requirement is in high-order polynomial of d and k thus can not be extended to the high-dimensional case. Another line of research focuses on spectral projections and their variants. By using different dimension reduction technique, they keep improving the separation condition for general Gaussian mixture models [43, 1, 33, 45, 25] and spherical Gaussian mixture models [54]. These works almost all focus on estimating the centers. Little is known about the convergence rate of estimating the cluster labels.

Another popular algorithm for Gaussian mixture model is the EM algorithm [19]. [58] analyzed the local convergence of EM for well-separated Gaussian mixtures. [15] showed that a two-round variant of EM algorithm with a special initializer is consistent under a strong separation condition that $\Delta \gtrsim \sigma\sqrt{d}$. Recently, a lot of attention has been gained on the computational guarantees of EM algorithm. Two component, symmetric Gaussian mixture is most widely studied due to its simple structure. For this model, [9] first proves the linear convergence of EM if the center is initialized in a small neighborhood of the true parameters, which implies an exponentially small error rate of labels after one-step label update. [36] extends their basin of attraction to be the intersection of a half space and a ball near the origin. More recently, [57, 16] prove the global convergence of the EM algorithm given infinite samples. While these results are encouraging, it is unclear whether their technique can be generalized to the k -mixtures or non-Gaussian cases. [32] provides some interesting negative results of EM algorithms. When $k \geq 3$, counterexamples are constructed to show there is no general global convergence of EM algorithm by uniformly initializing the centers from data points.

1.3. *Organization and Notation.* The rest of this paper is organized as follows. We introduce the sub-Gaussian mixture problem and Lloyd’s algorithm in Section 2. Section 3 gives statistical and computational guarantees of Lloyd’s algorithm, and discusses their implications when using spectral initialization. In Section 4, we consider two variants of Lloyd’s algorithms for community detection and crowdsourcing and establish their theoretical

properties. A number of experiments are conducted in Section 5 that confirm the sharpness of our theoretical findings and demonstrate the effectiveness of our algorithms on real data. In Section 6, we discuss our results on random initialization and the error of estimating centers. Finally, we present the main proofs in Section 7, with more technical part of the proofs deferred to the supplement.

Throughout the paper, let $[m] = \{1, 2, \dots, m\}$ for any positive integer m . For any vector a , $\|a\|$ is the ℓ_2 norm of a . $\mathbb{I}\{\cdot\}$ is the indicator function. Given two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n \gtrsim b_n$ means there is a universal constant C such that $a_n \geq Cb_n$ for all n , and define $a_n \lesssim b_n$ vice versa. We write $a_n \asymp b_n$ if $a_n \gtrsim b_n$ and $a_n \lesssim b_n$. Denote $a_n = o(b_n)$ if $a_n/b_n \rightarrow 0$ as $n \rightarrow \infty$.

2. Model and Lloyd’s algorithm. In this section, we first introduce the mixture of sub-Gaussians, a standard model for k -means, then briefly review Lloyd’s algorithm, followed by introducing spectral clustering algorithms as a way of initialization.

2.1. *Mixture of sub-Gaussians.* Suppose we observe independent samples $y_1, y_2, \dots, y_n \in \mathbb{R}^d$ from a mixture of k sub-Gaussian distributions,

$$(2.1) \quad y_i = \theta_{z_i} + w_i \quad \text{for } i \in [n],$$

where $z_1, z_2, \dots, z_n \in [k]$ are the underlying labels, and $\theta_1, \dots, \theta_k \in \mathbb{R}^d$ are unknown centers of those k distributions. We assume the noise $\{w_i, i \in [n]\}$ are independent zero mean sub-Gaussian vectors with parameter $\sigma > 0$, i.e.

$$(2.2) \quad \mathbb{E}e^{\langle a, w_i \rangle} \leq e^{\frac{\sigma^2 \|a\|^2}{2}}, \quad \text{for all } i \in [n] \text{ and } a \in \mathbb{R}^d.$$

A special case of model (2.1) is the symmetric, two-component mixture model, in which the two centers are θ^* and $-\theta^*$. We observe y_1, y_2, \dots, y_n from the following generative model,

$$(2.3) \quad y_i = z_i \theta^* + \xi_i,$$

where $z_i \in \{-1, 1\}$ and $\{\xi_i, i \in [n]\}$ are independent Gaussian noise with covariance matrix $\sigma^2 I_d$. Here with a little abuse of notation, we also use $z_i \in \{-1, 1\}$ to denote the underlying labels. As arguably the simplest mixture model, this special model has recently gained some attention in studying the convergence of EM algorithm [9, 36, 57]. Other examples of model (2.1) are planted partition model [43] for random graphs, stochastic block model [27]

for network data analysis and Dawid-Skene model [17] for crowdsourcing (see Section 4 for more tails).

For these clustering problems, our main goal is to recover the unknown labels z_i rather than to estimate the centers $\{\theta_j\}$. Note that the cluster structure is invariant to the permutations of label symbols. We define the mis-clustering rate of estimated labels $\hat{z}_1, \dots, \hat{z}_n$ as

$$(2.4) \quad L(\hat{z}, z) = \inf_{\pi \in \mathcal{S}_k} \left[\frac{1}{n} \sum_{i=1}^n \mathbb{I}\{\pi(\hat{z}_i) \neq z_i\} \right],$$

where \mathcal{S}_k is the collection of all the mappings from $[k]$ to $[k]$.

2.2. Lloyd's Algorithm. Lloyd's algorithm was originally proposed to solve the following k -means problem. Given n vectors $y_1, y_2, \dots, y_n \in \mathbb{R}^d$ and an integer k , the goal is to find k points $\theta_1, \theta_2, \dots, \theta_k \in \mathbb{R}^d$ to minimize the following objective function

$$(2.5) \quad \sum_{i \in [n]} \min_{j \in [k]} \|y_i - \theta_j\|^2.$$

The problem above is equivalent to find $\hat{\theta}_1, \dots, \hat{\theta}_k \in \mathbb{R}^d$ and $\hat{z}_1, \dots, \hat{z}_n \in [k]$ such that

$$(2.6) \quad (\hat{\theta}, \hat{z}) = \operatorname{argmin}_{(\theta, z)} \sum_{i \in [n]} \left\| y_i - \sum_{j=1}^k \theta_j \mathbb{I}\{z_i = j\} \right\|^2.$$

From a statistical point of view, (2.6) is the maximum likelihood estimator of model (2.1) with spherical Gaussian noise. It has been proved by Pollard [51, 49] the strong consistency and central limit theorem of using (2.5) to estimate the centers θ .

Lloyd's algorithm is simply motivated by the following observation of (2.6). If we fix θ , \hat{z}_i is the index of the center that y_i is closest to, and if we fix z , $\hat{\theta}_j$ is the sample mean of those y_i with $z_i = j$. If we start with an initial estimate of centers or labels, we can iteratively update the labels and centers. Therefore, we have the following Lloyd's algorithm.

1. Get an initial estimate of the centers or the labels.
2. Repeat the following iteration until convergence.
 - 2a. For $h = 1, 2, \dots, k$,

$$(2.7) \quad \hat{\theta}_h^{(s)} = \frac{\sum_{i=1}^n y_i \mathbb{I}\{\hat{z}_i^{(s)} = h\}}{\sum_{i=1}^n \mathbb{I}\{\hat{z}_i^{(s)} = h\}}.$$

2b. For $i = 1, 2, \dots, n$,

$$(2.8) \quad \hat{z}_i^{(s+1)} = \operatorname{argmin}_{h \in [k]} \|y_i - \hat{\theta}_h^{(s)}\|^2.$$

For model (2.3), since the centers are parametrized by a single vector θ^* , the center update step (2.7) can be simplified to

$$(2.9) \quad \hat{\theta}^{(s)} = \frac{1}{n} \sum_{i=1}^n \hat{z}_i^{(s)} y_i.$$

Lloyd's algorithm is guaranteed to converge because it is a greedy algorithm that keeps decreasing the value of objective function (2.6). However, it may only converge to a local minimum and thus a good initializer is needed. One possible way of initializing Lloyd's algorithm is to use spectral methods. Note that the signal matrix θ in model (2.1) is low rank. Spectral methods first do a de-noising step by projecting the data onto the subspace spanned by its top singular vectors, which approximately preserves the clustering structure. Since solving the k -means problem is NP-hard in general [14], we run an approximated k -means algorithm [52] on the projected data matrix, whose running time is polynomial in n, k and d . There are many versions of spectral clustering algorithms and here is the pseudo code of the one used in this paper.

1. Compute the SVD of the data matrix $Y = [y_1, \dots, y_n] = UDV'$. Let U_k be the first k columns of U .
2. Project y_1, \dots, y_n onto U_k , i.e. let $\hat{y}_i = U_k U_k' y_i$ for $i \in [n]$.
3. Run a $O(1)$ -approximation algorithm [52] for k -means problem on the columns of projected matrix $\hat{Y} = [\hat{y}_1, \dots, \hat{y}_n]$.

We refer this algorithm as spectral clustering algorithm henceforth. A lot of progress has been made in studying theoretical properties of this spectral clustering algorithm. For more details, we refer to [34, 8] and references therein.

3. Clustering sub-Gaussian mixtures. In this section, we present the results of using Lloyd's algorithm to clustering sub-Gaussian mixtures. In Section 3.1, we give the convergence results of Lloyd's algorithm to the symmetric, two component mixture model (2.3). Then we extend it to the general k mixture model (2.1) in Section 3.2. A minimax lower bound is established in Section 3.3.

3.1. *Two mixtures.* A key quantity in determining the basin of attraction of Lloyd's algorithm is the following normalized signal-to-noise ratio

$$r = \frac{\|\theta^*\|}{\sigma\sqrt{1+\eta}},$$

where $\eta = 9d/n$. Here we normalize the signal-to-noise ratio by $\sqrt{1+\eta}$ because the statistical precision of estimating θ^* is at the order of $\sigma\sqrt{\eta}$. If $\|\theta^*\| \lesssim \sigma\sqrt{\eta}$, information theoretically we could not distinguish between two centers with positive probability, even when the labels are known. Let $A_s = \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{\hat{z}_i^{(s)} \neq z_i\}$ be the mis-clustering rate at the iteration s of Lloyd's algorithm, and $\hat{\theta}^{(s)}$ be the estimated centers at step s , $s = 0, 1, 2, \dots$. We have the following theorem that characterizes the behavior of A_s .

THEOREM 3.1. *Assume $n \geq C$ and $r \geq C$ for a sufficiently large constant C . For any given (data dependent) initializer satisfying*

$$(3.1) \quad A_0 \leq \frac{1}{2} - \frac{2.56 + \sqrt{\log r}}{r} - \frac{1}{\sqrt{n}} \quad \text{or} \quad \|\hat{\theta}^{(0)} - \theta^*\| \leq \left(1 - \frac{4}{r}\right) \|\theta^*\|,$$

with probability $1 - \nu$, we have

$$(3.2) \quad A_{s+1} \leq \left(A_s + \frac{8}{r^2}\right) A_s + \frac{2}{r^2} + \sqrt{\frac{4 \log n}{n}}, \quad \text{for all } s \geq 0$$

with probability greater than $1 - \nu - n^{-3} - 2 \exp\left(-\frac{\|\theta^*\|^2}{3\sigma^2}\right)$, and

$$(3.3) \quad A_s \leq \exp\left(-\frac{\|\theta^*\|^2}{16\sigma^2}\right), \quad \text{for all } s \geq 3 \log n$$

with probability greater than $1 - \nu - 5n^{-1} - 8 \exp\left(-\frac{\|\theta^*\|^2}{16\sigma^2}\right)$. Moreover, if $r \rightarrow \infty$ as $n \rightarrow \infty$, the error rate in (3.3) can be improved to $\exp\left(-(1 + o(1))\frac{\|\theta^*\|^2}{2\sigma^2}\right)$.

The proof of Theorem 3.1 is given in Section 7. We will first show that after a few iterations, the mis-clustering ratio A_s is sufficiently small (for example, smaller than $1/8$) by proving (3.2), then prove mis-clustering rate of Lloyd's algorithm is exponentially small with an exponent determined by the signal-to-noise ratio. Note that the mis-clustering rate A_s only takes discrete values in $\{0, n^{-1}, 2n^{-1}, \dots, 1\}$. If $\|\theta^*\| > 4\sigma\sqrt{\log n}$, (3.3) implies $A_s < 1/n$, and thus Theorem 3.1 guarantees a zero clustering error after $\lceil 3 \log n \rceil$ Lloyd's iterations with high probability.

By Theorem 3.1, the convergence of Lloyd's algorithm has three stages. In the first iteration, it escapes from a small neighborhood around $1/2$ and achieves an error rate $A_s \leq 1/2 - c$ for a small constant c . Then it has a linear convergence rate, which depends on the signal-to-noise ratio. Finally, similar to other two-stage estimators [61, 22], once the mis-clustering rate A_s is sufficiently small, it converges to the optimal statistical precision of the problem after one or more iterations.

As we shall see in Section 3.3, a necessary condition for consistently estimating the labels is $\|\theta^*\|/\sigma \rightarrow \infty$. Therefore, our condition on the initializer is just slightly stronger than the random initialization error $1/2$. Balakrishnan et al.[9] studied the convergence of EM algorithm under the same model. They require an initializer $\hat{\theta}^{(0)} \in \mathbb{B}(\theta^*, \|\theta^*\|/4)$ under the assumption that r is sufficiently large. Here we replace $1/4$ by a factor close to 1 in Theorem 3.1. More specifically, the factor in the initialization condition is determined by the signal-to-noise ratio. The stronger the signal-to-noise ratio is, the weaker initialization condition we need.

3.2. *k mixtures.* Now we consider general number of mixtures. To better present our results, we first introduce some notation. For all $h \in [k]$, let T_h^* be the true cluster h and $T_h^{(s)}$ be the estimated cluster h at iteration s . Define $n_h^* = |T_h^*|$, $n_h^{(s)} = |T_h^{(s)}|$ and $n_{hg}^{(s)} = |T_h^* \cap T_g^{(s)}|$. The mis-clustering rate at iteration s can be written as

$$A_s = \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{\hat{z}_i^{(s)} \neq z_i\} = \frac{1}{n} \sum_{g \neq h \in [k]^2} n_{gh}^{(s)}.$$

We define a cluster-wise mis-clustering rate at iteration s as

$$G_s = \max_{h \in [k]} \left\{ \frac{\sum_{g \neq h \in [k]} n_{gh}^{(s)}}{n_h^{(s)}}, \frac{\sum_{g \neq h \in [k]} n_{hg}^{(s)}}{n_h^*} \right\}.$$

The first term in the maximum operator of definition of G_s can be understood as the false positive rate of cluster h and the second term is the true negative rate of cluster h . It is easy to see the relationship that $A_s \leq G_s$.

Let $\Delta = \min_{g \neq h \in [k]} \|\theta_g - \theta_h\|$ be the signal strength. For $h \in [k]$, let $\hat{\theta}_h^{(s)}$ be the estimated center of cluster h at iteration s . Define our error rate of estimating centers at iteration s as

$$\Lambda_s = \max_{h \in [k]} \frac{1}{\Delta} \|\hat{\theta}_h^{(s)} - \theta_h\|.$$

Besides signal-to-noise ratio, there are other two factors that determine the convergence of Lloyd's algorithm, the maximum signal strength and the minimum cluster size. As argued by [32], when one cluster is very far away from other clusters, local search may fail to find the global optimum, which indicates initialization should depend on the maximum signal strength. Since the cluster sizes influence the accuracy to estimate centers, we have a lower bound on the size of the smallest clusters. Define $\lambda = \max_{g \neq h \in [k]} \|\theta_g - \theta_h\|/\Delta$ and $\alpha = \min_{h \in [k]} n_h^*/n$. Similar to the two-cluster case, we define a normalized signal-to-noise ratio

$$(3.4) \quad r_k = \frac{\Delta}{\sigma} \sqrt{\frac{\alpha}{1 + kd/n}}.$$

Now we are ready to present our results for the k mixtures.

THEOREM 3.2. *Assume $n\alpha^2 \geq Ck \log n$ and $r_k \geq C\sqrt{k}$ for a sufficiently large constant C . Given any (data dependent) initializer satisfying*

$$(3.5) \quad G_0 < \left(\frac{1}{2} - \frac{6}{\sqrt{r_k}}\right) \frac{1}{\lambda} \quad \text{or} \quad \Lambda_0 \leq \frac{1}{2} - \frac{4}{\sqrt{r_k}},$$

with probability $1 - \nu$, we have

$$(3.6) \quad G_{s+1} \leq \frac{C_1}{r_k^2} G_s + \frac{C_1}{r_k^2} + \sqrt{\frac{5k \log n}{\alpha^2 n}} \quad \text{for all } s \geq 1$$

for some constant $C_1 \leq C$ with probability greater than $1 - \nu - n^{-3}$, and

$$(3.7) \quad A_s \leq \exp\left(-\frac{\Delta^2}{16\sigma^2}\right) \quad \text{for all } s \geq 4 \log n$$

with probability greater than $1 - \nu - 4/n - 2 \exp(-\Delta/\sigma)$. Moreover, if $\sqrt{k} = o(r_k)$ and $k \log n = o(n\alpha^2)$ as $n \rightarrow \infty$, the exponent in (3.7) can be improved to $\exp\left(-(1 + o(1))\frac{\Delta^2}{8\sigma^2}\right)$.

In Theorem 3.2, we establish similar results as in the two-mixture case. The initialization condition (3.5) here is slightly stronger due to the asymmetry. Similarly, after initialization that satisfies (3.5), the convergence has three stages. It first escapes from a small neighborhood of $1/2$, then converges linearly, and finally it achieves an exponentially small mis-clustering rate after $\lceil 4 \log n \rceil$ iterations.

From Theorem 3.2, Lloyd's algorithm does not require to know or to estimate the covariance structure of Gaussian mixtures. Likelihood-based

algorithms, such as EM algorithm and methods of moments, need to estimate the covariance structure. Lloyd's algorithm works for any general sub-Gaussian mixtures satisfying (2.2), and thus is robust to different models. This demonstrates a key advantage of using Lloyd's algorithm than other likelihood-based algorithms.

Theoretically, the initialization condition (3.5) in Theorem 3.2 is sharp in a sense that we give a counterexample in the supplementary material showing that Lloyd's algorithm may not converge when $G_0 = 1/(2\lambda)$ or $\Lambda = 1/2$.

Now let us give a sufficient condition for the spectral clustering initializer to fall into the basin of attraction (3.5). By Claim 1 in Section 3.2 of [34] and Lemma 7.1, for each center θ_h , there is a center μ_h returned by spectral clustering algorithm such that $\|\mu_h - \theta_h\| \lesssim (\sqrt{k}/r_k)\Delta$ with probability greater than $1 - \exp(-n/4)$, which implies the following corollary.

COROLLARY 3.1. *Assume $\sqrt{k} = o(r_k)$ and $k \log n = o(n\alpha^2)$ as $n \rightarrow \infty$. Let \hat{z} be the output of Lloyd's algorithm initialized by spectral clustering algorithm after $\lceil 4 \log n \rceil$ iterations. Then*

$$(3.8) \quad \ell(\hat{z}, z) \leq \exp\left(-\left(1 + o(1)\right)\frac{\Delta^2}{8\sigma^2}\right)$$

with probability greater than $1 - 5n^{-1} - 2\exp(-\Delta/\sigma)$.

This corollary gives a sufficient separation (signal-to-noise ratio) condition for clustering sub-Gaussian mixtures. When $n \geq kd$ and all the cluster sizes are of the same order, our separation condition simplifies to $\Delta \gtrsim k\sigma$. When there are finite number of clusters (k is finite), our separation condition further simplifies to $\Delta \gtrsim \sigma$. To our knowledge, this is the first result to give an explicit exponentially small error rate for estimating the labels. Previous studies mostly focus on exact recovery of the cluster labels and have no explicit statistical convergence rate. Furthermore, our results hold for all the range of d , while previous results all require $n \gg kd$.

Previously, the best known separation condition on the Lloyd-type algorithm is $r_k \gtrsim \sqrt{k} \text{poly log } n$, under the assumption that $n \gg kd/\alpha$ [8]. Corollary 3.1 improves it by a $\text{poly log } n$ factor. For Gaussian mixtures, among all algorithms including Lloyd's algorithm, the best known separation condition is $\Delta \gtrsim \sigma(\alpha^{-1/2} + \sqrt{k^2 + k \log(nk)})$ under the assumption that $\alpha n \gtrsim k(d + \log k)$ [1]. Our condition is weaker when the cluster sizes are of the same order. When k is finite, our condition is weaker by a $\sqrt{\log n}$ factor. As we shall see in Section 5, k is usually small in practice and can often be regarded as a constant.

3.3. *Minimax lower bound.* To show that the mis-clustering rate in Corollary 3.1 cannot be improved, we present a rate matching lower bound in this section. Define a parameter space as follows,

$$(3.9) \quad \Theta = \left\{ (\theta, z), \theta = [\theta_1, \dots, \theta_k] \in \mathbf{R}^{d \times k}, \Delta \leq \min_{g \neq h} \|\theta_g - \theta_h\|, \right. \\ \left. z : [n] \rightarrow [k], |\{i \in [n], z_i = u\}| \geq \alpha n, \forall u \in [k] \right\}.$$

We have the following minimax lower bound.

THEOREM 3.3. (*Lower Bound*) For model (2.1), assume independent gaussian noise $w_{ij} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, then when $\frac{\Delta}{\sigma \log(k/\alpha)} \rightarrow \infty$,

$$\inf_{\hat{z}} \sup_{(z, \theta) \in \Theta} \mathbb{E} \ell(\hat{z}, z) \geq \exp\left(- (1 + o(1)) \frac{\Delta^2}{8\sigma^2}\right).$$

If $\frac{\Delta}{\sigma} + \log(k/\alpha) = O(1)$, $\inf_{\hat{z}} \sup_{(z, \theta) \in \Theta} \mathbb{E} \ell(\hat{z}, z) \geq c$ for some constant $c > 0$.

This lower bound result shows that if the signal-to-noise ratio Δ/σ is at the constant order, the worst case mis-clustering rate is lower bounded by a constant. In other words, a necessary condition for consistently estimating the labels is $\Delta/\sigma \rightarrow \infty$. To achieve strong consistency, we need $\Delta/\sigma \gtrsim \sqrt{\log n}$. Theorem 3.3 indicates the necessity of separation conditions in estimating the labels. Previous results that use methods of moments to estimate the cluster centers [28, 5], on the contrary, do not require separation conditions. This reveals the difference between estimating the cluster centers and the labels.

Together with Corollary 3.1, Theorem 3.3 gives us the minimax rate of estimating the underlying labels of Gaussian mixtures.

4. Applications. In this section, we generalize the results in Section 3 to community detection and crowdsourcing by considering two variants of Lloyd's algorithm.

4.1. *Community Detection.* We have a network of k communities. There are n nodes in the network and $\{z_i \in [k], i \in [n]\}$ is the community assignment. We observe a symmetric adjacency matrix $A \in \{0, 1\}^{n \times n}$ with zero diagonals, which is generated by the following stochastic block model [27] (SBM thereafter),

$$A_{ij} \sim \begin{cases} \text{Ber}(\frac{a}{n}), & \text{if } z_i = z_j \\ \text{Ber}(\frac{b}{n}), & \text{otherwise,} \end{cases}$$

for $1 \leq i < j \leq n$, where $0 < b < a < n$ and $\{z_i\}$ are unknown parameters. The SBM is a special case of the Planted Partition Model [43] in the theoretical computer science literature. As in [22], we use a single parameter β to control the community sizes, namely, there is a $\beta > 0$ such that $\frac{n}{\beta k} \leq |\{i \in [n], z_i = g\}|$ for all $g \in [k]$. The goal of community detection is to estimate community structure z using the adjacency matrix A . Note that $\mathbb{E}A_i$ only takes k different values, and the SBM can be viewed as a special case of model (2.1) with $d = n$ and $\sigma^2 = \frac{a}{n}$, ignoring the fact that A is symmetric with zero diagonals.

When a and b are of an order of n , the SBM satisfies the sub-Gaussian noise condition (2.2) and we have $\Delta^2 \geq \frac{(a-b)^2}{n^2} \times \frac{2n}{\beta k} = \frac{2(a-b)^2}{\beta kn}$ and $\sigma^2 \leq a/n$. Initialized by the spectral clustering algorithm, let \hat{z} be the labels returned by running Lloyd's algorithm on the rows of adjacency matrix A with $\lceil 4 \log n \rceil$ iterations. By considering the fact that A is symmetric with zero diagonals, we can slightly modify the proof of Corollary 3.1 to obtain the following result.

COROLLARY 4.1. *Assume $\frac{(a-b)^2}{\beta^3 a k^4} \rightarrow \infty$ and $\frac{n}{\beta^2 k^3 \log n} \rightarrow \infty$ as $n \rightarrow \infty$, then*

$$(4.1) \quad \ell(z, \hat{z}) \leq \exp\left(-\frac{(1+o(1))(a-b)^2}{4a\beta k}\right)$$

with probability tending to 1.

We omit the proof of Corollary 4.1. A more interesting case of the SBM is the sparse case when $a = o(n)$ and $b = o(n)$, in which the sub-Gaussian assumption (2.2) does not hold. By utilizing the community structure, we consider the following variants of Lloyd's algorithm.

As pointed out by [13], under the sparse setting of $a \lesssim \log n$, the trimmed adjacency matrix A^τ is a better estimator of $\mathbb{E}A$ than A under the spectral norm. Therefore, we run spectral clustering on the A^τ to get an initial estimator of z . Then we update the labels node by node. For the i -th node in the network, we estimate its connecting probability to community j (defined as $B_{z_i j}$) based on our previous estimated labels. Then we assign its label to be the j that maximizes $\hat{B}_{z_i j}$, which is expected to be close to a/n or b/n , depending on whether $z_i = j$. The following theorem gives a theoretical guarantee of the CommuLloyd algorithm.

THEOREM 4.1. *Assume $n \geq 6k^2 \log n$, $(a-b)^2 \geq C_0 a \beta^2 k^2 \log(\beta k)$ for a sufficiently large constant C_0 . Let A_s be the mis-clustering rate at iteration*

Algorithm 1: CommuLloyd

Input: Adjacency matrix A . Number of communities k . Trimming threshold τ .

Output: Estimated labels $\hat{z}_1, \dots, \hat{z}_n$.

1. Trim the adjacency matrix:

1a. calculate the degree of each node $d_i = \sum_{j=1}^n A_{ij}, \forall i \in [n]$.

1b. trim the adjacency matrix $A_{ij}^\tau = A_{ij} \mathbb{I}\{d_i \leq \tau\}, \forall i \in [n], j \in [n]$.

2. Run spectral clustering algorithm on the trimmed matrix A_τ .

3. Run following iterations until converge,

$$(4.2) \quad \hat{B}_{ih} = \frac{\sum_{j=1}^n A_{ij} \mathbb{I}\{\hat{z}_j = h\}}{\sum_{j=1}^n \mathbb{I}\{\hat{z}_j = h\}}, \quad \forall h \in [k], i \in [n].$$

$$(4.3) \quad \hat{z}_i = \operatorname{argmax}_{h \in [k]} \hat{B}_{ih}, \quad \forall i \in [n].$$

s of the CommuLloyd algorithm. Then for any initializer satisfies $G_0 \leq 1/4$ with probability $1 - \nu$, we have

$$(4.4) \quad A_{s+1} \leq \exp\left(-\frac{(a-b)^2}{2C\beta ak}\right) + \frac{4}{5}A_s, \quad \forall 1 \leq s \leq 3 \log n$$

with probability greater than $1 - n^{-1} - \nu$.

The proof of Theorem 4.1 is given in Section B of supplemental material. We show the linear convergence of CommuLloyd algorithm given the first stage group-wise mis-clustering rate $G_0 \leq 1/4$. In fact, this initialization assumption can be relaxed to $G_0 \leq 1/2 - \sqrt{a \log(\beta k) \beta k} / (a - b)$. To better present our results, we simplify it to 1/4. By Lemma 7 in [22] and Theorem 3.1 in [8], we can guarantee a group-wise initialization error of 1/4 when $(a - b)^2 \gtrsim a\beta^2 k^3$. Therefore, we obtain an exponentially small error rate under the signal-to-noise ratio condition

$$(a - b)^2 \gtrsim a\beta^2 k^3 \log \beta.$$

Theorem 4.1 provides theoretical justifications of the phenomenon observed in the numerical experiments of [22]. More iterations achieve better mis-clustering rate. The Lloyd iterations enjoy similar theoretical performance as the likelihood based algorithm proposed in [22]. While they require a global initialization error to be $o(1/(k \log k))$, we require a cluster-wise initialization error to be smaller than 1/4. Moreover, the CommuLloyd algorithm is computationally much more efficient than Algorithm 1 in [22],

which requires obtaining n different initializers and hence running SVD on the network n times. Theoretically, we relax the assumption in [22] that $a \asymp b$ and β is a constant, and we improve the best known signal-to-noise ratio condition [22] by a $\log k$ factor. When β is a constant and $(a-b)^2/(ak^3) \rightarrow \infty$ as $n \rightarrow \infty$, we are able to match the minimax rate obtained in [60].

4.2. *Crowdsourcing.* Crowdsourcing is an efficient and inexpensive way to collect a large amount of labels for supervised machine learning problems. We refer to [61, 21] and references therein for more details. The following Dawid-Skene model is the most popular model considered in the crowdsourcing literature.

Suppose there are m workers to label n items. Each item belongs to one of the k categories. Denote by $z = (z_1, z_2, \dots, z_n) \in [k]^n$ the true labels of n items and by X_{ij} the label of worker i given to item j . Our goal is to estimate the true labels z using $\{X_{ij}\}$. Dawid-Skene model assumes that the workers are independent and that given $z_j = g$, the labels given by worker i are independently drawn from a multinomial distribution with parameter $\pi_{ig*} = (\pi_{ig1}, \dots, \pi_{igk})$, i.e.,

$$\mathbb{P}\{X_{ij} = h | z_j = g\} = \pi_{igh}$$

for all $g, h \in [k], i \in [m], j \in [n]$. Dawid-Skene model can be seen as a special case of the mixture model (2.1) via the following transformation. For $j \in [n]$, let

$$y_j = \left(\mathbb{I}\{X_{1j} = 1\}, \dots, \mathbb{I}\{X_{1j} = k\}, \dots, \mathbb{I}\{X_{mj} = 1\}, \dots, \mathbb{I}\{X_{mj} = k\} \right)'.$$

Then given $z_j = g$, we have $\mathbb{E}[y_j | z_j = g] = \theta_g$ with $\theta_g = (\pi_{1g1}, \pi_{1g2}, \dots, \pi_{mgk})$. By defining $w_{ijh} = \mathbb{I}\{X_{ij} = h\} - \pi_{iz_jh}$, our observations y_j can be decomposed as the following signal-plus-noise model (2.1) with

$$w_j = [w_{1j1}, w_{1j2}, \dots, w_{mj1}, w_{mj2}, \dots, w_{mjk}]'.$$

Therefore, we consider the following variants of the Lloyd's algorithm in Algorithm 2. We iteratively estimate workers' accuracy and items' labels. Iterations (4.6) and (4.7) are actually Lloyd's iteration (2.7) and (2.8) under the transformed Dawid-Skene model. Iterations (4.6) and (4.7) are also closely related to the EM update for Dawid-Skene model [61]. (4.6) is the same as the M-step. In (4.7), we first use least squares to approximate the log-likelihood function of multinomial distributions, then we do a hard labeling step instead of soft labeling in the E-step of EM algorithm.

Algorithm 2: CrowdLloyd

Input: $\{X_{ij}, i \in [m], j \in [n]\}$. Number of possible labels k .

Output: Estimated labels $\hat{z}_1, \dots, \hat{z}_n$.

1. Initialization via Majority Voting:

$$(4.5) \quad \hat{z}_j = \operatorname{argmax}_{g \in [k]} \sum_{i=1}^m \mathbb{I}\{X_{ij} = g\} \quad \forall j \in [n].$$

2. Run following iterations until converge.

$$(4.6) \quad \hat{\pi}_{igh} = \frac{\sum_{j=1}^n \mathbb{I}\{X_{ij} = h, \hat{z}_j = g\}}{\sum_{h=1}^k \sum_{j=1}^n \mathbb{I}\{X_{ij} = h, \hat{z}_j = g\}} \quad \forall i \in [m], g, h \in [k]^2.$$

$$(4.7) \quad \hat{z}_j = \operatorname{argmin}_{h \in [k]} \left[\sum_{i=1}^m \sum_{g=1}^k (\mathbb{I}\{X_{ij} = g\} - \hat{\pi}_{igh})^2 \right] \quad \forall j \in [n].$$

As suggested by Corollary 3.1, we could use spectral clustering to initialize the labels. But the majority voting initializer, as described in the first step of Algorithm 2, is a more natural and commonly used initializer of crowdsourcing. By regarding all the workers have the same accuracy, it simply estimates the labels by aggregating the results of every worker with equal weights. The majority voting initializer is computationally more efficient than spectral methods. We have the following result of the majority voting estimator.

THEOREM 4.2. *Under the Dawid-Skene model, the majority voting initializer satisfies*

$$(4.8) \quad \frac{1}{n} \sum_{j=1}^n \mathbb{I}\{\hat{z}_j \neq z_j\} \leq \exp\left(-\frac{mV(\pi)}{2}\right)$$

with probability greater than $1 - \exp(-mV(\pi)/2 + \log k)$, where

$$V(\pi) = \min_{g \neq h} \frac{1}{m} \left(\sqrt{\sum_{i=1}^m \pi_{igg}} - \sqrt{\sum_{i=1}^m \pi_{igh}} \right)^2.$$

The proof of Theorem 4.2 is given in Section C.2. To our knowledge, this theorem is the first theoretical characterization of the Majority Voting estimator under the general Dawid-Skene model. $mV(\pi)$ measures the collective accuracy of m workers on estimating one of the labels. If there are two labels

g and h that are confusing for most of the workers, the value of $V(\pi)$ is small and majority voting estimator may not have good performance. Under the special case of one-coin model, where $\pi_{igg} = p$ and $\pi_{igh} = \frac{1-p}{k-1}$ for $g \neq h$, we have

$$mV(\pi) = m(\sqrt{p} - \sqrt{(1-p)/(k-1)})^2 \gtrsim \log m$$

as long as $p \geq \frac{1}{k} + \sqrt{\frac{\log m}{m}}$. That is, for Majority Voting estimator to be consistent, the accuracy of workers giving true labels only needs to be $\sqrt{\log m/m}$ better than random guess.

With Theorem 4.2 and Theorem 3.2, we are able to obtain an error rate upper bound of Algorithm 2. It can be proved that the sub-Gaussian parameter of w_i is 2 (see the proof of Corollary 4.2 for more details). Define $\Delta^2 = \min_{g \neq h} \sum_{i=1}^m \sum_{u=1}^k (\pi_{igu} - \pi_{ihu})^2$. Then we have the following corollary of Theorem 3.2.

COROLLARY 4.2. *Assume $mV(\pi) \geq C \log k$, $\alpha \Delta^2 \geq Ck$ and $n\alpha \geq k\Delta^2 \log n$ for a sufficiently large constant C . Let $\hat{z} = (\hat{z}_1, \dots, \hat{z}_n)$ be the estimated labels returned by running Algorithm 2. We have*

$$(4.9) \quad \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{\hat{z}_i \neq z_i\} \leq \exp\left(-\frac{\Delta^2}{64}\right)$$

with probability greater than $1 - \exp(-mV(\pi)/4) - 4/n - 2 \exp(-\Delta/2)$.

The proof of Theorem 4.2 is given in Section C.2. We achieve an exponentially small error rate for estimating the labels under Dawid-Skene model. By Theorem 4.2 in [21], we can achieve the minimax optimal rate by an additional likelihood based Neyman-Pearson testing step followed by Algorithm 2. Previous provable results [61, 21] assume $\pi_{igh} \geq \rho > 0$ for all $(i, g, h) \in [m] \times [k]^2$, which is quite restrictive since a very good worker could have $\pi_{igh} = 0$ for some $h \neq g$. By considering least square update (4.7), we successfully get rid of this assumption. In contrast to algorithms that using spectral methods as initializer [61], Theorem 4.2 does not need an eigenvalue condition on the confusion matrix.

5. Numerical Analysis.

5.1. Gaussian Mixture Model.

5.1.1. *Simulation.* In this section, we provide simulation results that are designed to test our theoretical findings in Theorem 3.2. As predicted by Theorem 3.2, there are three stages of the convergence of Lloyd's algorithm. Given an initial estimator of the labels with error smaller than $1/(2\lambda)$, Lloyd's algorithm escapes from a small neighborhood of $1/(2\lambda)$ in the first iteration. Then it has a geometrically decay mis-clustering rate during the second stage. Finally, it converges to the statistical precision of the problem and achieves an exponentially small error. In all the results reported below, we plot the logarithm of the mis-clustering rate versus the iteration count. Each curve plotted is an average of 10 independent trails.

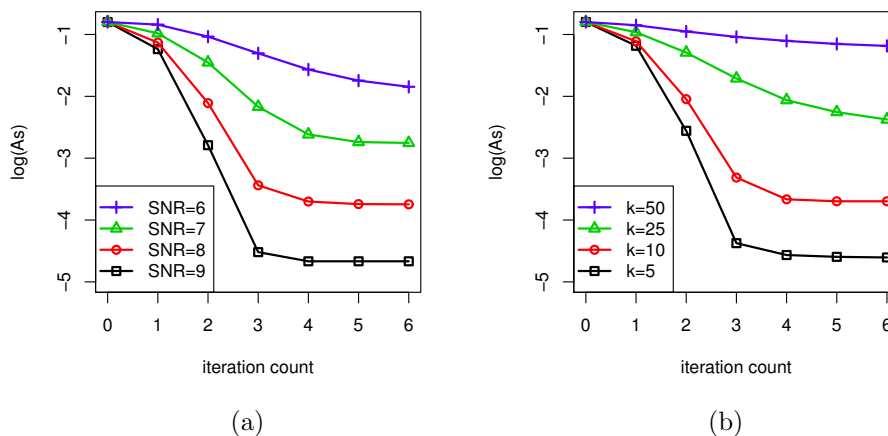


Fig 1: Plots of the log mis-clustering rate A_s versus iteration count.

In the first simulation, we independently generate $n = 1000$ samples from a mixture of $k = 10$ spherical Gaussians. Each cluster has 100 data points. The centers of those 10 gaussians are orthogonal unit vectors in \mathbb{R}^d with $d = 100$. The four curves in Figure 1a corresponds to four choices of the standard deviation $\sigma = 2/SNR$ with $SNR \in \{6, 7, 8, 9\}$. We use the same initializer with a group-wise mis-clustering rate of 0.454 for all the experiments. As shown by plots in Figure 1a, there is a linear convergence of the log mis-clustering rate at the first 2 iterations and then the convergence slows down. After four iterations, the log mis-clustering rate plateaus roughly at the level of $-\frac{1}{16}SNR^2$. Moreover, we observe in Figure 1a that the convergence rate increases as SNR increases, as expected from Theorem 3.2.

The second experiment investigates the role of k in the convergence of Lloyd's algorithm. We first randomly generate $k \in \{5, 10, 25, 50\}$ centers

Date set	k	n	d	IF-HCT-PCA	SpecLloyd	RandLloyd
Brain	5	42	5597	.262	.119	.262(.04)
Leukemia	2	72	3571	.069	.028	.028
Lung(1)	2	181	12533	.033	.083	.098
Lymphoma	3	62	4026	.069	.177	.016
Prostate	2	102	6033	.382	.422	.422
SRBCT	4	63	2308	.444	.492	.492

TABLE 1

Clustering error rate of three clustering methods on six gene microarray datasets.

that are orthogonal unit vectors in \mathbb{R}^{100} . Then we add independent gaussian noise $\mathcal{N}(0, 0.25^2)$ to each coordinate of those centers to get $1000/k$ samples for each cluster. In total, we have $n = 1000$ independent samples from a mixture of spherical Gaussian distributions. We set the same initializer as in the first simulation to make a group-wise mis-clustering rate of 0.37. As observed in Figure 1b, the larger the k , the slower Lloyd’s algorithm converges. This agrees with the phenomena predicted by Theorem 3.2. It is interesting to observe that when k is large, Lloyd’s algorithm does not converge to the optimal mis-clustering rate in Figure 1b. When the value of k is doubled, the convergence rate of $\log A_s$ slows down roughly by half, which matches the k -term in (3.6).

5.1.2. *Gene Microarray Data.* To demonstrate the effectiveness of Lloyd’s algorithm on high dimensional clustering problem, we use six gene microarray data sets that were preprocessed and analyzed in [20, 24]. In each data set, we are given measurements of the expression levels of d genes for n people. There are several underlying classes (normal or diseased) of those people. Our goal is to cluster them based on their gene expression levels, with the number of groups k given. One common feature of those six data sets is that the dimensionality d is much larger than the number of samples n . As shown in Table 1, while n is usually smaller than two hundreds, the number of genes can be as large as ten thousands.

We compare three different methods on these six data sets. RandLloyd is Lloyd’s algorithms initialized by randomly drawing k rows of the original data as k centers. SpecLloyd runs RandLloyd on the first k eigenvectors of the data, followed by Lloyd’s iterations on the original data. IF-HCT-PCA, proposed in [31], does a variable selection step first and then applies spectral clustering on selected variables. All three methods need random initializations at the k -means step and thus have algorithmic randomness. We randomly initialize centers for 30 times and use the initializer that give us the minimum k -means objective value.

In Table 1, we report the average mis-clustering rate rates over 30 independent experiments. We also report the standard deviation if it is larger than .01. As we can see, RandLloyd is comparable to other two methods on these gene microarray data. It achieves the best clustering error rate when the clustering task is relatively easy (signal-to-noise ratio is strong), for example, Leukemia and Lymphoma data sets. Note that both IF-HCT-PCA and SpecLloyd have a dimension reduction step. This agrees with the phenomena suggested by Theorem 3.2 that Lloyd’s algorithm also works on the high dimensional setting when the signal-to-noise ratio is strong enough.

5.2. *Community Detection.* In this section, we report the results of empirical studies to compare the CommuLloyd algorithm proposed in Section 4.1 with other methods. In particular, we compare to spectral clustering (Spec) and spectral clustering with one step refinement studied in [22]. Since the algorithm in [22] is motivated by the maximum likelihood estimator, we refer it to as Spec+MLE.

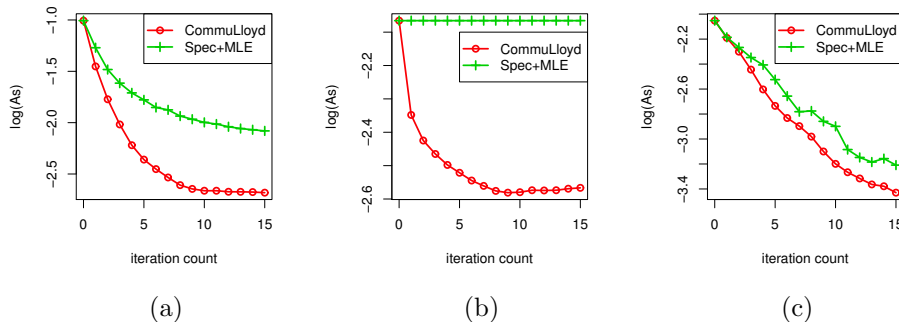
5.2.1. *Simulation.* Using the same settings as in [22], we generate data from SBM under three different scenarios: (1) dense network with equal size communities; (2) sparse network with equal size communities; and (3) dense network with unequal size communities. The detailed set-up of three simulations is described in Table 2. We conduct each experiment with 10 independent repetitions and report the average results of each algorithm in Figure 2.

For all the experiments, we use spectral clustering methods in [38] as our initializer and plot the logarithm mis-clustering rate versus the Lloyd iteration or MLE iteration counts. Iteration 0 represents the spectral clustering initializer. As shown by Figure 2, CommuLloyd significantly outperforms spectral clustering and spectral clustering with one step MLE refinement under all three settings. Surprisingly, it is even better than multiple steps MLE refinements. In panel (a), MLE refinements converge to a sub-optimal mis-clustering rate, and in panel (b), it does not improve the spectral clustering initializer. This phenomenon may be due to the stronger initialization condition for MLE refinements in [22], while in Theorem 4.1, CommuLloyd needs only a constant-error initializer to converge to the optimal mis-clustering rate rate. The convergence rate of CommuLloyd for the these two cases, however, is faster than the linear convergence predicted by Theorem 4.1. As for the unbalance case in panel (c), CommuLloyd and MLE has similar performance and they both exhibit a linear convergence rate.

Setting	n	k	Within	Between	Community Size
Balanced Case	2000	10	0.20	0.11	equal
Sparse Case	2000	4	0.019	0.005	equal
Unbalanced Case	1000	4	0.35	0.22	100,200,300,400

TABLE 2

The settings of three simulations on community detection. Within is the within-community connecting probability a/n . Between is the between-community connecting probability b/n . Equal means the communities are of equal size.

Fig 2: Plots of log mis-clustering rate A_s versus the iteration.

5.2.2. *Political Blog Data.* We now examine the performance of Algorithm 2 on a political blog dataset [2]. Each node in this data represents a web blog collected right after the 2004 US presidential election. Two nodes are connected if one contains a hyperlink to the other, and we ignore the direction of the hyperlink. After pre-processing [35], there are 1222 nodes and 16714 edges. Depending on whether liberal or conservative, the network naturally has two communities. We use the label given in [2] as our ground truth.

Table 3 reports the number of mis-labeled nodes of our algorithm on political blog dataset. Started from spectral clustering with an error rate of 437/1222, CommuLloyd achieves 56/1222 error rate after three iterations. After that, it fluctuates between 55 and 56 mis-labeled nodes. Our algorithm slightly outperforms the state-of-the-art algorithm SCORE [30], which reports an error rate of 58/1222.

5.3. *Crowdsourcing.* In this section, we compare CrowdLloyd algorithm with three other crowdsourcing methods: (1) Opt-D&S, which is Spectral methods with one step EM update proposed in [61], (2) MV, the majority

Methods	Spec	Spec+MLE(1st)	Spec+MLE(10th)	SCORE	CommuLloyd
No.mis-clustered	437	132	61	58	56

TABLE 3

Number of mis-labeled blogs for different community detection methods. Here *Spec* is the spectral clustering method described in [38]. We refer the one-step MLE refinement of *Spec* as *Spec+MLE(1st)*, and 10 steps MLE refinements as *Spec+MLE(10th)*. *SCORE* is a community detection algorithm for Degree-corrected SBM proposed in [30].

voting estimator and (3) MV+EM, which is majority voting estimator followed by EM updates. Among those four algorithms, Opt-D&S, MV and CrowdLloyd have provable error rates.

When there are ties, the majority voting estimator is not clearly defined. To avoid algorithmic randomness, we return the first label index when there are ties for CrowdLloyd and MV estimator. Since EM algorithm takes soft labels as input, we use weights rather than hard label of majority voting estimator in MV+EM. In practice, there are usually missing data. We denote by 0 as the category of missing data and modify the label update step in CrowdLloyd as

$$(5.1) \quad \hat{z}_j^{(s+1)} = \operatorname{argmin}_{h \in [k]} \left[\sum_{i, X_{ij} \neq 0} \sum_{g=1}^k \left(\mathbb{I}\{X_{ij} = g\} - \hat{\pi}_{igh}^{(s)} \right)^2 \right]$$

5.3.1. *Simulation.* We follow the same setting as in the experiments of [61], where there are $m = 100$ workers and $n = 1000$ items. Each item belongs to one of the $k = 2$ categories. For worker i , the diagonal entries of his or her confusion matrix π_{igg} are independently and uniformly drawn from the interval $[0.3, 0.9]$. We set $\pi_{igh} = 1 - \pi_{igg}$ for all $i \in [n]$ and $(h, g) \in \{(1, 2), (2, 1)\}$. To simulate missing entries, we assume each entry X_{ij} is observed independently with probability $p \in \{0.2, 0.5, 1\}$.

We replicate experiments independently for 10 times and report the average results in Table 4. Among three provable algorithms, CrowdLloyd is consistently better than Opt D&S (1st iteration) and MV under different missing levels. CrowdLloyd, MV+EM and Opt D&S (10th iteration) have almost identical performance. As expected, CrowdLloyd is slightly worse because it does not use likelihood function to update the confusion matrices. As discussed before, an additional Neyman-Pearson testing step followed by CrowdLloyd algorithm would give the same output as EM updates.

5.3.2. *Real Data.* We now compare CrowdLloyd algorithm to other three algorithms on five real crowdsourcing data sets: (1) bird dataset is a binary

Methods	$p = 1.0$	$p = 0.5$	$p = 0.2$
Opt D&S (1st iteration)	0.06	1.32	10.32
Opt D&S (10th iteration)	0.07	1.10	8.22
MV	2.14	8.56	18.97
MV + EM	0.07	1.10	8.17
CrowdLloyd	0.07	1.14	8.19

TABLE 4

mis-clustering rate rate (%) of different crowdsourcing algorithms on simulated data.

#	Date set	Opt D&S	MV	MV+EM	CrowdLloyd
1	Bird	10.09	24.07	11.11	12.04
2	RTE	7.12	8.13	7.34	6.88
3	TREC	29.80	34.86	30.02	29.19
4	Dog	16.89	19.58	16.66	15.99
5	Web	15.86	26.93	15.74	14.25

TABLE 5

mis-clustering rate rate (%) of different crowdsourcing algorithms on five real data sets

task of labeling two bird species, (2) recognizing textual entailment (RTE) dataset contains a set of sentence pairs and the task is to label whether the second sentence can be inferred from the first one, (3) TREC dataset is a binary task of assessing the quality of information retrieval in TREC 2011 crowdsourcing track, (4) dog dataset is a subset of Stanford dog dataset that requires labeling one of the four dog breeds, and (5) web dataset is a set of query-URL pairs for workers to label on a relevance scale from 1 to 5.

The mis-clustering rates of four methods are summarized in Table 5. CrowdLloyd slightly outperforms the other three on four of those five data sets. One possible explanation is that the least squares update (4.7) works for general sub-Gaussian distributions and thus is potentially more robust than the likelihood based EM update.

6. Discussion.

6.1. *Random Initialization.* In this section, we explore the condition under which random initialization converges to the global optimum. Here we consider the symmetric two-mixture model (2.3). Note that Theorem 3.1 holds for any initializer (could be data dependent) satisfying condition (3.1). If we have a data independent initializer, condition (3.1) can be weakened and we have the following result on random initialization of the labels.

THEOREM 6.1. *Assume*

$$\|\theta^*\| \geq C\sigma \left(\sqrt{\log(1/\delta)} + (d \log(1/\delta)(1 + d/n))^{1/4} \right)$$

for a sufficiently large constant C . Suppose we independently draw $3 \log(1/\delta)$ random initializers. Then Lloyd's algorithm initialized by one of them can correctly classify all the data points with probability greater than $1 - \delta$.

The proof of Theorem 6.1 is given in appendix. Klusowski and Brinda [36] showed similar results for EM algorithm under the condition that $\|\theta^*\| \gtrsim \sigma\sqrt{d \log d}$. We improve their condition to $\|\theta^*\| \gtrsim \sigma(d \log d)^{1/4}$ by choosing $\delta = 1/d$. Recently, Xu et al. [57] and Daskalakis et al. [16] proved the global convergence of EM algorithm on the population likelihood function. However, it is not clear what signal-to-noise ratio condition they need for the finite sample analysis. Moreover, our results can be generalized to k -mixtures and the sub-Gaussian case, but [36], [57] and [16] all restrict their analyses to the two mixture Gaussian case.

As for the general k mixture case, one popular way to initialize Lloyd's algorithm is to initialize k centers by sampling without replacement from the data. However, when $k \geq 3$, Jin et al. [32] show even for well-separated ($\Delta \gtrsim \sqrt{d}\sigma$) spherical Gaussian mixtures, EM algorithm, initialized by above strategy, converges to bad critical points with high probability. They prove it by constructing a bad case of two cluster of centers that are very far away with each other. We expect similar results for Lloyd's algorithm.

6.2. *Convergence of the centers.* As a byproduct of Theorem 3.2, we obtain the convergence rate of estimating Gaussian mixture's centers using Lloyd's algorithm.

THEOREM 6.2. *Assume $n\alpha^2 > Ck \log n$ and $r_k \geq C\sqrt{k}$ for a sufficiently large constant C . Then given any initializer satisfying*

$$(6.1) \quad G_0 < \left(\frac{1}{2} - \frac{6}{\sqrt{r_k}}\right) \frac{1}{\lambda} \quad \text{or} \quad \Lambda_0 \leq \frac{1}{2} - \frac{4}{\sqrt{r_k}}$$

with probability $1 - \nu$, we have

$$(6.2) \quad \max_{h \in [k]} \|\hat{\theta}_h^{(s)} - \theta_h\|^2 \leq \frac{C_2 \sigma^2 (d + \log n)}{n} + \sigma^2 \exp\left(-\frac{\Delta^2}{24\sigma^2}\right), \quad \text{for } s \geq 3 \log n + \log r_k$$

for some constant C_2 with probability greater than $1 - 2 \exp\left(-\frac{\Delta}{\sigma}\right) - \frac{4}{n} - \nu$.

We show a linear convergence rate of estimating the centers. When the labels are known, the minimax optimal rate of estimating the centers under Euclidean distance is $\sigma\sqrt{d/n}$. Thus, we obtain the optimal rate under the condition that $n\alpha^2 \gtrsim k \log n$, $r_k \gtrsim \sqrt{k}$, $d \gtrsim \log n$ and $\Delta \gtrsim \sigma \log(n/d)$.

7. Proofs. This section is devoted to prove Theorem 3.1. The proofs of other theorems are collected in the supplementary material.

7.1. *Preliminaries.* We rewrite the symmetric two-mixture model (2.3) as

$$(7.1) \quad z_i y_i = \theta^* + w_i, \quad i = 1, 2, \dots, n,$$

where $w_i = z_i \xi_i \sim \mathcal{N}(0, \sigma^2 I_d)$. To facilitate our proof, we present four lemmas on the concentration behavior of w_i . Their proofs are given in Section D of the appendix.

LEMMA 7.1. *For any $S \subseteq [n]$, define $W_S = \sum_{i \in S} w_i$. Then,*

$$(7.2) \quad \|W_S\| \leq \sigma \sqrt{2(n+9d)|S|} \quad \text{for all } S \subseteq [n]$$

with probability greater than $1 - \exp(-0.1n)$.

LEMMA 7.2. *Let $\bar{w} = \frac{1}{n} \sum_{i=1}^n w_i$. Then we have $\bar{w}'\theta^* \geq -\frac{\|\theta^*\|^2}{\sqrt{n}}$ and $\|\bar{w}\|^2 \leq \frac{3d\sigma^2}{n} + \frac{\|\theta^*\|^2}{n}$ with probability greater than $1 - 2\exp\left(-\frac{\|\theta^*\|^2}{3\sigma^2}\right)$.*

LEMMA 7.3. *The maximum eigenvalue of $\sum_{i=1}^n w_i w_i'$ is upper bounded by $1.62(n+4d)\sigma^2$ with probability greater than $1 - \exp(-0.1n)$.*

LEMMA 7.4. *For any fixed $i \in [n]$, $S \subseteq [n]$, $t > 0$ and $\delta > 0$, we have*

$$\mathbb{P} \left\{ \left\langle w_i, \frac{1}{|S|} \sum_{j \in S} w_j \right\rangle \geq \frac{3\sigma^2(t\sqrt{|S|} + d + \log(1/\delta))}{|S|} \right\} \leq \exp \left(-\min \left\{ \frac{t^2}{4d}, \frac{t}{4} \right\} \right) + \delta.$$

Let \mathcal{E} be the intersection of high probability events in Lemma 7.1, Lemma 7.2, Lemma 7.3 and label initialization condition (3.1). Then $\mathbb{P}\{\mathcal{E}^c\} \leq \nu + n^{-3} + \exp(-\|\theta^*\|^2/(3\sigma^2))$. For the center initialization condition, we refer to the proof of Theorem 3.2 in the supplement for more details. In the following, we deterministically analyze the error of estimating centers and labels on event \mathcal{E} .

Error of the centers. We first decompose the error of estimating θ^* into two parts, where the first part comes from the error of estimating labels z_i , and the second part is due to the stochastic error w_i ,

$$\hat{\theta}^{(s)} - \theta^* = \frac{1}{n} \sum_{i=1}^n \hat{z}_i^{(s)} y_i - \frac{1}{n} \sum_{i=1}^n z_i y_i + \frac{1}{n} \sum_{i=1}^n z_i y_i - \theta^*$$

Using (7.1) and the fact that $\hat{z}_i^{(s)} - z_i = -2\mathbb{I}\{\hat{z}_i^{(s)} \neq z_i\}z_i$, we obtain

$$(7.3) \quad \hat{\theta}^{(s)} - \theta^* = -\frac{2}{n} \sum_{i=1}^n \mathbb{I}\{\hat{z}_i^{(s)} \neq z_i\}(\theta^* + w_i) + \frac{1}{n} \sum_{i=1}^n w_i = -2A_s\theta^* - 2R + \bar{w}$$

where $c_i = \mathbb{I}\{\hat{z}_i^{(s)} \neq z_i\}$, $A_s = \frac{1}{n} \sum_{i=1}^n c_i$ and $R = \frac{1}{n} \sum_{i=1}^n c_i w_i$.

Error of the labels. Now we analyze the error of $\hat{z}_i^{(s+1)}$. The iteration of $\hat{z}_i^{(s+1)}$ is equivalent to

$$\hat{z}_i^{(s+1)} = \operatorname{argmin}_{g \in \{-1,1\}} \|gy_i - \hat{\theta}^{(s)}\|_2^2 = \operatorname{argmax}_{g \in \{-1,1\}} \langle gy_i, \hat{\theta}^{(s)} \rangle = \operatorname{argmax}_{g \in \{-1,1\}} gz_i \langle \theta^* + w_i, \hat{\theta}^{(s)} \rangle,$$

which gives us $\mathbb{I}\{\hat{z}_i^{(s+1)} \neq z_i\} = \mathbb{I}\left\{ \langle \theta^* + w_i, \hat{\theta}^{(s)} \rangle \leq 0 \right\}$. From (7.3),

$$\begin{aligned} \langle \theta^* + w_i, \hat{\theta}^{(s)} \rangle &= \langle \theta^* + w_i, (1 - 2A_s)\theta^* - 2R + \bar{w} \rangle \\ &= (1 - 2A_s)\|\theta^*\|^2 - (2R - \bar{w})'\theta^* + \langle w_i, \theta^* - 2A_s\theta^* + 2R - \bar{w} \rangle. \end{aligned}$$

Applying Lemma 7.1 with $S = \{i \in [n], c_i = 1\}$, we obtain

$$\|R\| \leq \frac{\sigma}{n} \sqrt{2(n+9d)nA_s} \leq \frac{\|\theta^*\|}{r} \sqrt{2A_s},$$

which, together with Lemma 7.2, implies

$$(2R - \bar{w})'\theta^* \leq 2\|R\|\|\theta^*\| - \bar{w}'\theta^* \leq \frac{2\sqrt{2A_s}}{r}\|\theta^*\|^2 + \frac{1}{\sqrt{n}}\|\theta^*\|^2.$$

Consequently, on event \mathcal{E} , we have

$$(7.4) \quad \mathbb{I}\{\hat{z}_i^{(s+1)} \neq z_i\} \leq \mathbb{I}\left\{ \beta_0 \|\theta^*\|^2 + \langle w_i, \theta^* \rangle + \langle w_i, -2A_s\theta^* + 2R - \bar{w} \rangle \leq 0 \right\},$$

where $\beta_0 = 1 - 2A_s - \frac{2\sqrt{2A_s}}{r} - \frac{1}{\sqrt{n}} \geq 1 - 2A_s - \frac{2}{r} - \frac{1}{\sqrt{n}}$. Now we are ready to prove Theorem 3.1. In the following, we are going to prove (3.2) and (3.3) separately based on two different decompositions of the RHS of (7.4).

7.2. *Proof of (3.2) in Theorem 3.1.* Using the fact that for any $a, b \in \mathbb{R}$ and $c > 0$, $\mathbb{I}\{a + b \leq 0\} \leq \mathbb{I}\{a \leq c\} + \mathbb{I}\{b \leq -c\} \leq \mathbb{I}\{a \leq c\} + \frac{b^2}{c^2}$, we obtain

$$\mathbb{I}\{\hat{z}_i^{(s+1)} \neq z_i\} \leq \mathbb{I}\left\{ \beta \|\theta^*\|^2 \leq -\langle w_i, \theta^* \rangle \right\} + \frac{(w_i'(2R - \bar{w} - 2A_s\theta^*))^2}{\delta^2 \|\theta^*\|^4},$$

with $\beta = \beta_0 - \delta$ and $\delta = 3.12/r$. Taking an average of the equation above over $i \in [n]$, we have $A_{s+1} \leq I_1 + I_2$, where

$$(7.5) \quad I_1 = \frac{1}{n} \sum_{i=1}^n \mathbb{I} \{ \langle w_i, \theta^* \rangle \leq -\beta \|\theta^*\|^2 \},$$

$$(7.6) \quad I_2 = \frac{1}{n\delta^2 \|\theta^*\|^4} \sum_{i=1}^n (w'_i(2R - \bar{w} - 2A_s \theta^*))^2.$$

Upper Bound I_1 . Note that nA_s only takes discrete values $\{1, 2, \dots, \lfloor \frac{n}{2} \rfloor\}$. Let us define $\gamma_a = 1 - 2a - \frac{5.12}{r} - \frac{1}{\sqrt{n}}$ and

$$T_i(a) = \mathbb{I} \{ \langle w_i, \theta^* \rangle \leq -\gamma_a \|\theta^*\|^2 \}$$

for $a \in \mathcal{D} \triangleq \{\frac{1}{n}, \frac{2}{n}, \dots, \frac{\lfloor \frac{n}{2} \rfloor}{n}\}$. For any fixed a , $\{T_i(a), i = 1, 2, \dots, n\}$ are independent Bernoulli random variables, then Hoeffding's inequality implies

$$\mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n (T_i(a) - \mathbb{E}T_i(a)) \geq \sqrt{\frac{4 \log(n/2)}{n}} \right\} \leq \frac{1}{n^4}.$$

Thus by union bound, we have

$$(7.7) \quad \mathbb{P} \left\{ \max_{a \in \mathcal{D}} \left[\frac{1}{n} \sum_{i=1}^n (T_i(a) - \mathbb{E}T_i(a)) \right] \geq \sqrt{\frac{4 \log(n/2)}{n}} \right\} \leq \frac{1}{n^3}.$$

Now it remains to upper bound the expectation of $T_i(a)$. Since $\langle w_i, \theta^* \rangle$ is sub-Gaussian with parameter $\sigma \|\theta^*\|$, Chernoff's bound yields

$$(7.8) \quad \mathbb{E}T_i(a) \leq \exp \left(-\frac{\gamma_a^2 \|\theta^*\|^2}{2\sigma^2} \right).$$

Consider the event

$$\mathcal{E}_1 = \left\{ \frac{1}{n} \sum_{i=1}^n T_i(a) \leq \exp \left(-\frac{\gamma_a^2 \|\theta^*\|^2}{2\sigma^2} \right) + \sqrt{\frac{4 \log(n/2)}{n}}, \forall a \in \mathcal{D} \right\}.$$

Equation (7.7) together with (7.8) implies $\mathbb{P}\{\mathcal{E}_1^c\} \leq n^{-3}$. Therefore,

$$(7.9) \quad I_1 \leq \exp \left(-\frac{\gamma_{A_s}^2 \|\theta^*\|^2}{2\sigma^2} \right) + \sqrt{\frac{4 \log(n/2)}{n}}$$

with probability greater than $1 - n^{-3}$.

Upper Bound I_2 . Lemma 7.3 implies

$$\begin{aligned} I_2 &\leq \frac{\|2R - \bar{w} - 2A_s\theta^*\|^4}{n\delta^2\|\theta^*\|^4} \lambda_{max} \left(\sum_{i=1}^n w_i w'_i \right) \\ &\leq \frac{1.62(1 + 4d/n)\sigma^2}{\delta^2\|\theta^*\|^4} \|2R - \bar{w} - 2A_s\theta^*\|^2, \end{aligned}$$

where $\lambda_{max}(B)$ stands for the maximum eigenvalue of the symmetric matrix B . By the Cauchy-Schwarz inequality, Lemma 7.1 and Lemma 7.2, we have

$$\begin{aligned} \|2R - \bar{w} - 2A_s\theta^*\|^2 &\leq (1 + 1 + 4) (4\|R\|^2 + \|\bar{w}\|^2 + A_s^2\|\theta^*\|^2) \\ &\leq 6\|\theta^*\|^2 \left(\frac{8}{r^2}A_s + \frac{1}{r^2} + \frac{1}{n} + A_s^2 \right). \end{aligned}$$

Since $\delta = 3.12/r$, we obtain

$$(7.10) \quad I_2 \leq \frac{8}{r^2}A_s + \frac{1}{r^2} + \frac{1}{n} + A_s^2.$$

Recall that $A_{s+1} \leq I_1 + I_2$. Combining (7.9) and (7.10), we have

$$A_{s+1} \leq \exp \left(-\frac{\gamma_{A_s}^2 \|\theta^*\|^2}{2\sigma^2} \right) + \sqrt{\frac{4 \log n}{n}} + \frac{8}{r^2}A_s + \frac{1}{r^2} + A_s^2$$

on event $\mathcal{E} \cap \mathcal{E}_1$. Let $\tau = \frac{1}{2} - \frac{1}{\sqrt{n}} - \frac{2.56 + \sqrt{\log r}}{r}$. From above inequality and the assumption that $A_0 \leq \tau$, it can be proved by induction that $A_s \leq \tau$ for all $s \geq 0$ when $r \geq 32$ and $n \geq 1600$. Consequently, we have $\gamma_{A_s} \geq \frac{2\sqrt{\log r}}{r}$ for all $s \geq 0$. Plugging this bound of γ_{A_s} into above inequality completes the proof.

7.3. *Proof of (3.3) in Theorem 3.1.* By (3.2) in Theorem 3.1 and the fact that $A_s \leq \tau$ when $r \geq 32$,

$$A_{s+1} \leq 0.5A_s + \frac{2}{r^2} + \sqrt{\frac{4 \log n}{n}},$$

which implies $A_s \leq 4r^{-2} + 5\sqrt{(\log n)/n}$ for all $s \geq \log n$. Then when $s \geq \log n$, (7.4) implies

$$\begin{aligned} &\mathbb{I}\{\hat{z}_i^{(s+1)} \neq z_i\} \\ &\leq \mathbb{I}\left\{ \left(1 - \frac{3}{r} - \frac{1 + 10\sqrt{\log n}}{\sqrt{n}} \right) \|\theta^*\|^2 + \langle w_i, \theta^* - 2A_s\theta^* + 2R - \bar{w} \rangle \leq 0 \right\}. \end{aligned}$$

To get a better upper bound for mis-clustering rate, we further decompose $\langle w_i, 2R - 2A_s\theta^* - \bar{w} \rangle$ as $\langle w_i, 2R - 2A_s\theta^* \rangle - \langle w_i, \bar{w} \rangle$. Following similar arguments as in the proof of Theorem 3.1(a), we obtain $A_{s+1} \leq J_1 + J_2 + J_3$, where J_1 is I_1 defined in (7.5) with $\beta = 1 - 6.72/r - (3 + 10\sqrt{\log n})/\sqrt{n}$, J_2 and J_3 are defined as follows,

$$(7.11) \quad J_2 = \frac{r^2}{8.1n\|\theta^*\|^4} \sum_{i=1}^n (w'_i(2R - 2A_s\theta^*))^2,$$

$$(7.12) \quad J_3 = \frac{1}{n} \sum_{i=1}^n \mathbb{I} \left\{ \left(\frac{1}{2r} + \frac{2}{\sqrt{n}} \right) \|\theta^*\|^2 \leq -\langle w_i, \bar{w} \rangle \right\}.$$

From (7.8), when $r \geq 32$ and $n \geq 40000$, $\mathbb{E}J_1 \leq \exp(-\frac{\|\theta^*\|^2}{8\sigma^2})$. Analogous to (7.10), we have

$$J_2 \leq \frac{8}{r^2} A_s + A_s^2 \leq \left(\frac{8}{r^2} + \frac{8}{r^2} \right) A_s \leq \frac{1}{r} A_s$$

on event \mathcal{E} . Choosing $S = [n]$, $t = \left(\sqrt{d} + \frac{\|\theta^*\|}{\sigma} \right) \frac{\|\theta^*\|}{\sigma}$ and $\delta = \exp\left(-\frac{\|\theta^*\|^2}{4\sigma^2}\right)$ in Lemma 7.4 yields

$$\mathbb{E}J_3 \leq 2 \exp\left(-\frac{\|\theta^*\|^2}{4\sigma^2}\right).$$

Combine the pieces and note that $A_{s+1} \leq 1$,

$$(7.13) \quad \begin{aligned} \mathbb{E}A_{s+1} &\leq \mathbb{E}[(J_1 + J_2 + J_3)\mathbb{I}\{\mathcal{E}\}] + \mathbb{E}[A_{s+1}\mathbb{I}\{\mathcal{E}^c\}] \\ &\leq \frac{1}{r} \mathbb{E}A_s + 3 \exp\left(-\frac{\|\theta^*\|^2}{8\sigma^2}\right). \end{aligned}$$

Then, by induction, we have

$$\mathbb{E}A_s \leq \frac{1}{r^{s-\lceil \log n \rceil}} + 4 \exp\left(-\frac{\|\theta^*\|^2}{8\sigma^2}\right),$$

when $r \geq 32$. Setting $s = \lceil 3 \log n \rceil$ and applying Markov's inequality, we obtain

$$\mathbb{P}\{A_s \geq t\} \leq 4 \exp\left(-\frac{\|\theta^*\|^2}{8\sigma^2} - \log t\right) + \frac{1}{n^2 t}.$$

If $\frac{\|\theta^*\|^2}{8\sigma^2} \leq 2 \log n$, a choice of $t = \exp\left(-\frac{\|\theta^*\|^2}{16\sigma^2}\right)$ gives us the desired result.

Otherwise, since A_s only takes discrete values $\{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\}$, choosing $t = \frac{1}{n}$, we have

$$\mathbb{P}\{A_s > 0\} = \mathbb{P}\left\{A_s \geq \frac{1}{n}\right\} \leq 4n \exp(-2 \log n) + \frac{1}{n} \leq \frac{5}{n}.$$

Therefore, the proof of (3.3) is complete. When $r \rightarrow \infty$, we can slightly modify the proof to get the improved error exponent. For more details, we refer to the proof of Theorem 3.2 in supplementary material.

SUPPLEMENTARY MATERIAL

Supplement to “Statistical and Computational Guarantees of Lloyd’s Algorithm and its Variants”

([url to be specified](#)). We provide detailed proofs in the supplement [41].

References.

- [1] ACHLIOPTAS, D. and MCSHERRY, F. (2005). On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory* 458–469. Springer.
- [2] ADAMIC, L. A. and GLANCE, N. (2005). The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery* 36–43. ACM.
- [3] AGARWAL, P. K. and MUSTAFA, N. H. (2004). k-Means projective clustering. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* 155–165. ACM.
- [4] AGGARWAL, A., DESHPANDE, A. and KANNAN, R. (2009). Adaptive sampling for k-means clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques* 15–28. Springer.
- [5] ANANDKUMAR, A., HSU, D. and KAKADE, S. M. (2012). A Method of Moments for Mixture Models and Hidden Markov Models. In *COLT* 1 4.
- [6] ARTHUR, D. and VASSILVITSKII, S. (2006). How Slow is the k-means Method? In *Proceedings of the twenty-second annual symposium on Computational geometry* 144–153. ACM.
- [7] ARTHUR, D. and VASSILVITSKII, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* 1027–1035. Society for Industrial and Applied Mathematics.
- [8] AWASTHI, P. and SHEFFET, O. (2012). Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques* 37–49. Springer.
- [9] BALAKRISHNAN, S., WAINWRIGHT, M. J. and YU, B. (2014). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *arXiv preprint arXiv:1408.2156*.
- [10] CAI, T. T., MA, Z., WU, Y. et al. (2013). Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics* **41** 3074–3110.
- [11] CHAGANTY, A. T. and LIANG, P. (2013). Spectral Experts for Estimating Mixtures of Linear Regressions. In *ICML (3)* 1040–1048.
- [12] CHANG, J. T. (1996). Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Mathematical biosciences* **137** 51–73.
- [13] CHIN, P., RAO, A. and VU, V. (2015). Stochastic block model and community detection in the sparse graphs: A spectral algorithm with optimal rate of recovery. *arXiv preprint arXiv:1501.05021* **2**.
- [14] DASGUPTA, S. (2008). *The hardness of k-means clustering*. Department of Computer Science and Engineering, University of California, San Diego.

- [15] DASGUPTA, S. and SCHULMAN, L. (2007). A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learning Research* **8** 203–226.
- [16] DASKALAKIS, C., TZAMOS, C. and ZAMPETAKIS, M. (2016). Ten Steps of EM Suffice for Mixtures of Two Gaussians. *arXiv preprint arXiv:1609.00368*.
- [17] DAWID, A. P. and SKENE, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics* 20–28.
- [18] DAY, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56** 463–474.
- [19] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)* 1–38.
- [20] DETTLING, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics* **20** 3583–3593.
- [21] GAO, C., LU, Y. and ZHOU, D. (2016). Exact Exponent in Optimal Rates for Crowdsourcing. *arXiv preprint arXiv:1605.07696*.
- [22] GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2015). Achieving optimal misclassification proportion in stochastic block model. *arXiv preprint arXiv:1505.03772*.
- [23] GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2016). Community Detection in Degree-Corrected Block Models. *arXiv preprint arXiv:1607.06993*.
- [24] GORDON, G. J., JENSEN, R. V., HSIAO, L.-L., GULLANS, S. R., BLUMENSTOCK, J. E., RAMASWAMY, S., RICHARDS, W. G., SUGARBAKER, D. J. and BUENO, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer research* **62** 4963–4967.
- [25] HARDT, M. and PRICE, E. (2015). Tight bounds for learning a mixture of two gaussians. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing* 753–760. ACM.
- [26] HERWIG, R., POUSTKA, A. J., MÜLLER, C., BULL, C., LEHRACH, H. and O’BRIEN, J. (1999). Large-scale clustering of cDNA-fingerprinting data. *Genome research* **9** 1093–1105.
- [27] HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social networks* **5** 109–137.
- [28] HSU, D. and KAKADE, S. M. (2013). Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science* 11–20. ACM.
- [29] HSU, D., KAKADE, S. M., ZHANG, T. et al. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab* **17** 1–6.
- [30] JIN, J. et al. (2015). Fast community detection by SCORE. *The Annals of Statistics* **43** 57–89.
- [31] JIN, J. and WANG, W. (2014). Important Feature PCA for high dimensional clustering. *arXiv preprint arXiv:1407.5241*.
- [32] JIN, C., ZHANG, Y., BALAKRISHNAN, S., WAINWRIGHT, M. J. and JORDAN, M. (2016). Local Maxima in the Likelihood of Gaussian Mixture Models: Structural Results and Algorithmic Consequences. *arXiv preprint arxiv:1609.00978*.
- [33] KANNAN, R., SALMASIAN, H. and VEMPALA, S. (2005). The spectral method for general mixture models. In *International Conference on Computational Learning Theory* 444–457. Springer.
- [34] KANNAN, R. and VEMPALA, S. (2009). Spectral Algorithms. *Found. Trends Theor. Comput. Sci.* 157–288.
- [35] KARRER, B. and NEWMAN, M. E. (2011). Stochastic blockmodels and community

- structure in networks. *Physical Review E* **83** 016107.
- [36] KLUSOWSKI, J. M. and BRINDA, W. (2016). Statistical Guarantees for Estimating the Centers of a Two-component Gaussian Mixture by EM. *arXiv preprint arXiv:1608.02280*.
- [37] KUMAR, A. and KANNAN, R. (2010). Clustering with spectral norm and the k-means algorithm. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on* 299–308. IEEE.
- [38] LEI, J. and RINALDO, A. (2013). Consistency of spectral clustering in sparse stochastic block models. *arXiv preprint arxiv:1312.2050*.
- [39] LINDSAY, B. G. and BASAK, P. (1993). Multivariate normal mixtures: a fast consistent method of moments. *Journal of the American Statistical Association* **88** 468–476.
- [40] LLOYD, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory* **28** 129–137.
- [41] LU, Y. and ZHOU, H. H. (2016). Supplement to “Statistical and Computational Guarantees of Lloyd’s Algorithm and its Variants”. *url to be specified*.
- [42] MAHAJAN, M., NIMBHORKAR, P. and VARADARAJAN, K. (2009). The planar k-means problem is NP-hard. In *International Workshop on Algorithms and Computation* 274–285. Springer.
- [43] MCSHERRY, F. (2001). Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on* 529–537. IEEE.
- [44] MILLIGAN, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika* **45** 325–342.
- [45] MOITRA, A. and VALIANT, G. (2010). Settling the polynomial learnability of mixtures of gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on* 93–102. IEEE.
- [46] ORDOVÁS-PASCUAL, I. and ALMEIDA, J. S. (2014). A fast version of the k-means classification algorithm for astronomical applications. *Astronomy & Astrophysics* **565** A53.
- [47] OSTROVSKY, R., RABANI, Y., SCHULMAN, L. J. and SWAMY, C. (2006). The effectiveness of Lloyd-type methods for the k-means problem. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS’06)* 165–176. IEEE.
- [48] PEARSON, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A* **185** 71–110.
- [49] POLLARD, D. (1982). A central limit theorem for k-means clustering. *The Annals of Probability* 919–926.
- [50] POLLARD, D. (1990). Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics* i–86. JSTOR.
- [51] POLLARD, D. et al. (1981). Strong consistency of k-means clustering. *The Annals of Statistics* **9** 135–140.
- [52] SONG, M. and RAJASEKARAN, S. (2010). Fast Algorithms for Constant Approximation k-Means Clustering. *Trans. MLDM* **3** 67–79.
- [53] TITTERINGTON, D. M., SMITH, A. F. and MAKOV, U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley.
- [54] VEMPALA, S. and WANG, G. (2004). A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences* **68** 841–860.
- [55] WANG, Z., LU, H. and LIU, H. (2014). Nonconvex statistical optimization: Minimax-optimal sparse pca in polynomial time. *arXiv preprint arXiv:1408.5352*.
- [56] WU, X., KUMAR, V., QUINLAN, J. R., GHOSH, J., YANG, Q., MOTODA, H., MCLACHLAN, G. J., NG, A., LIU, B., PHILIP, S. Y. et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems* **14** 1–37.

- [57] XU, J., HSU, D. and MALEKI, A. (2016). Global Analysis of Expectation Maximization for Mixtures of Two Gaussians. *arXiv preprint arXiv:1608.07630*.
- [58] XU, L. and JORDAN, M. I. (1996). On convergence properties of the EM algorithm for Gaussian mixtures. *Neural computation* **8** 129–151.
- [59] YI, X., CARAMANIS, C. and SANGHAVI, S. (2014). Alternating Minimization for Mixed Linear Regression. In *ICML* 613–621.
- [60] ZHANG, A. Y. and ZHOU, H. H. (2015). Minimax rates of community detection in stochastic block models. *arXiv preprint arXiv:1507.05313*.
- [61] ZHANG, Y., CHEN, X., ZHOU, D. and JORDAN, M. I. (2014). Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Advances in neural information processing systems* 1260–1268.

SUPPLEMENT TO “STATISTICAL AND COMPUTATIONAL
GUARANTEES OF LLOYD’S ALGORITHM AND ITS
VARIANTS”

BY Yu Lu and Harrison H. Zhou

Yale University

In this supplement, we present detailed proofs for general sub-Gaussian mixtures, random initializations for two component sub-Gaussian mixtures, statistical and computation guarantees of Algorithm 1 for the stochastic block model and Algorithm 2 for the Dawid-Skene model.

APPENDIX A: PROOF OF SUB-GAUSSIAN MIXTURES

A.1. Preliminaries. Let us begin by introducing some notation. For any $S \subseteq [n]$, define $W_S = \sum_{i \in S} w_i$. Recall that $T_g^* = \{i \in [n], z_i = g\}$ and $T_g^{(s)} = \{i \in [n], \hat{z}_i^{(s)} = g\}$, let us define

$$S_{gh}^{(s)} = \left\{ i \in [n], z_i = g, \hat{z}_i^{(s)} = h \right\} = T_g^* \cap T_h^{(s)}.$$

Then we have $n_h^{(s)} = \sum_{g \in [k]} n_{gh}^{(s)}$ and $n_h^* = \sum_{g \in [k]} n_{hg}^{(s)}$. In the rest of the proof, we will sometimes drop the upper index (s) of $n_{gh}^{(s)}$, $n_h^{(s)}$ and $S_{gh}^{(s)}$ when there is no ambiguity. Also, we suppress the dependence of k by writing r_k as r .

Analogous to Lemma 7.1 - Lemma 7.4, we have following technical lemmas on the concentration behavior of sub-Gaussian random vectors $\{w_i\}$.

LEMMA A.1.

$$(A.1) \quad \|W_S\| \leq \sigma \sqrt{3(n+d)|S|} \quad \text{for all } S \subseteq [n].$$

with probability greater than $1 - \exp(-0.3n)$.

LEMMA A.2.

$$(A.2) \quad \lambda_{max} \left(\sum_{i=1}^n w_i w_i' \right) \leq 6\sigma^2(n+d).$$

with probability greater than $1 - \exp(-0.5n)$.

LEMMA A.3. For any fixed $i \in [n]$, $S \subseteq [n]$, $t > 0$ and $\delta > 0$,

$$\mathbb{P} \left\{ \left\langle w_i, \frac{1}{|S|} \sum_{j \in S} w_j \right\rangle \geq \frac{3\sigma^2(t\sqrt{|S|} + d + \log(1/\delta))}{|S|} \right\} \leq \exp \left(- \min \left\{ \frac{t^2}{4d}, \frac{t}{4} \right\} \right) + \delta.$$

LEMMA A.4.

$$(A.3) \quad \|W_{T_h^*}\| \leq 3\sigma \sqrt{(d + \log n)|T_h^*|} \quad \text{for all } h \in [k]$$

with probability greater than $1 - n^{-3}$.

LEMMA A.5. For any fixed $\theta_1, \dots, \theta_k \in \mathbb{R}^d$ and $a > 0$, we have

$$(A.4) \quad \sum_{i \in T_g^*} \mathbb{I} \{ a \|\theta_h - \theta_g\|^2 \leq \langle w_i, \|\theta_h - \theta_g\| \rangle \} \leq n_g^* \exp \left(- \frac{a^2 \Delta^2}{2\sigma^2} \right) + \sqrt{5n_g^* \log n}$$

for all $g \neq h \in [k]^2$ with probability greater than $1 - n^{-3}$.

A.2. Two Key Lemmas. The following two lemmas give the iterative relationship between the error of estimating centers and the error of estimating labels. Let \mathcal{E} be the intersection of high probability events in Lemma A.1, Lemma A.2 Lemma A.4, Lemma A.5 and the initialization condition (3.5). Then we have $\mathbb{P}\{\mathcal{E}^c\} \leq 3n^{-3} + \nu$. In the rest part of the proof, if not otherwise stated, we all condition on the event \mathcal{E} and the following analysis are deterministic.

LEMMA A.6. On event \mathcal{E} , if $G_s \leq \frac{1}{2}$, then we have

$$(A.5) \quad \Lambda_s \leq \frac{3}{r} + \min \left\{ \frac{3}{r} \sqrt{kG_s} + 2G_s \Lambda_{s-1}, \lambda G_s \right\}.$$

LEMMA A.7. On event \mathcal{E} , if $\Lambda_s \leq \frac{1-\epsilon}{2}$ and $r \geq 36\epsilon^{-2}$, then

$$(A.6) \quad G_{s+1} \leq \frac{2}{\epsilon^4 r^2} + \left(\frac{28}{\epsilon^2 r} \Lambda_s \right)^2 + \sqrt{\frac{5k \log n}{\alpha^2 n}}.$$

PROOF OF LEMMA A.6. For any $B \subseteq [n]$, define $\bar{Y}_B = \frac{1}{|B|} \sum_{i \in B} y_i$. The error of estimated centers at step s can be written as

$$\begin{aligned} \hat{\theta}_h^{(s)} - \theta_h &= \frac{1}{n_h} \sum_{i \in S_{hh}} (y_i - \theta_h) + \frac{1}{n_h} \sum_{a \neq h} \sum_{i \in S_{ah}} (y_i - \theta_h) \\ &= \frac{1}{n_h} \sum_{i \in S_{hh}} w_i + \sum_{a \neq h} \frac{n_{ah}}{n_h} (\bar{Y}_{S_{ah}} - \theta_h) \end{aligned}$$

According to our label update step, we have $\|y_i - \hat{\theta}_h^{(s-1)}\| \leq \|y_i - \hat{\theta}_a^{(s-1)}\|$ for any $i \in S_{ah}$. This means for any $i \in S_{ah}$, y_i is closer to $\hat{\theta}_h^{(s-1)}$ than $\hat{\theta}_a^{(s-1)}$, so is the average of $\{y_i, i \in S_{ah}\}$. Thus, we have

$$\|\bar{Y}_{S_{ah}} - \hat{\theta}_h^{(s-1)}\| \leq \|\bar{Y}_{S_{ah}} - \hat{\theta}_a^{(s-1)}\|.$$

Consequently, triangle inequality gives us

$$\|\bar{Y}_{S_{ah}} - \theta_h\| \leq \|\bar{Y}_{S_{ah}} - \theta_a\| + \|\hat{\theta}_a^{(s-1)} - \theta_a\| + \|\hat{\theta}_h^{(s-1)} - \theta_h\|,$$

which, combined with Lemma A.1 and the definition of Λ_{s-1} , yields

$$\|\bar{Y}_{S_{ah}} - \theta_h\| \leq \sigma \sqrt{3(n+d)/n_{ah}} + 2\Lambda_{s-1}\Delta.$$

Taking a weighted sum over $a \neq h \in [k]$, we get

$$\begin{aligned} \sum_{a \neq h} \frac{n_{ah}}{n_h} \|\bar{Y}_{S_{ah}} - \theta_h\| &\leq \frac{\sigma \sqrt{3(n+d)}}{n_h} \sum_{a \neq h} \sqrt{n_{ah}} + 2\Lambda_{s-1}\Delta \sum_{a \neq h} \frac{n_{ah}}{n_h} \\ &\leq \frac{\sigma \sqrt{3(n+d)}}{\sqrt{n_h}} \sqrt{(k-1)G_s} + 2G_s\Lambda_{s-1}\Delta, \end{aligned}$$

where the Last inequality is due to Cauchy-Schwartz and the fact that $\sum_{a \neq h} n_{ah} \leq G_s n_h$. Note that $W_{S_{hh}} = W_{T_h^*} - \sum_{a \neq h} W_{S_{ha}}$. Triangle inequality and Lemma A.5 imply

$$\|W_{S_{hh}}\| \leq 3\sigma \sqrt{d + \log n} \sqrt{n_h^*} + \sigma \sqrt{3(n+d)} \sqrt{n_h^* - n_{hh}}.$$

Since $G_s \leq \frac{1}{2}$, we have

$$(A.7) \quad n_h \geq n_{hh} \geq n_h^*(1 - G_s) \geq \frac{1}{2}n_h^* \geq \frac{1}{2}\alpha n.$$

Combining the pieces, we obtain

$$\begin{aligned} \|\hat{\theta}_h^{(s)} - \theta_h\| &\leq 3\sigma \sqrt{\frac{d + \log n}{\alpha n}} + 3\sigma \sqrt{\frac{k(n+d)}{\alpha n}} G_s + 2G_s\Lambda_{s-1}\Delta \\ (A.8) \quad &\leq \left(\frac{3}{r} (1 + \sqrt{kG_s}) + 2G_s\Lambda_{s-1} \right) \Delta. \end{aligned}$$

Therefore, we get the first term in (A.5). To prove the second term, we

decompose $\hat{\theta}_h^{(s)}$ differently.

$$\begin{aligned}
\hat{\theta}_h^{(s)} &= \frac{1}{n_h} \sum_{i=1}^n (\theta_{z_i} + w_i) \mathbb{I} \left\{ \hat{z}_i^{(s)} = h \right\} \\
&= \frac{1}{n_h} \sum_{a=1}^k \sum_{i=1}^n \theta_a \mathbb{I} \left\{ z_i = a, \hat{z}_i^{(s)} = h \right\} + \frac{1}{n_h} \sum_{i \in T_h} w_i \\
\text{(A.9)} \quad &= \sum_{a=1}^k \frac{n_{ah}}{n_h} \theta_a + \frac{1}{n_h} W_{T_h}.
\end{aligned}$$

Then, the error of $\hat{\theta}_h^{(s)}$ can be upper bounded as

$$\left\| \hat{\theta}_h^{(s)} - \theta_h \right\| = \left\| \sum_{a=1}^k \frac{n_{ah}}{n_h} (\theta_a - \theta_h) + \frac{1}{n_h} W_{T_h} \right\| \leq \left\| \sum_{a \neq h} \frac{n_{ah}}{n_h} (\theta_a - \theta_h) \right\| + \left\| \frac{1}{n_h} W_{T_h} \right\|.$$

By triangle inequality,

$$\text{(A.10)} \quad \left\| \sum_{a \neq h} \frac{n_{ah}}{n_h} (\theta_a - \theta_h) \right\| \leq \sum_{a \neq h} \frac{n_{ah}}{n_h} \|\theta_a - \theta_h\| \leq \lambda \Delta \sum_{a \neq h} \frac{n_{ah}}{n_h} \leq \lambda \Delta G_s.$$

This, together with Lemma A.1 and (A.7), implies

$$\text{(A.11)} \quad \left\| \hat{\theta}_h^{(s)} - \theta_h \right\| \leq \lambda \Delta G_s + \sigma \sqrt{\frac{3(n+d)}{n_h}} \leq \left(\lambda G_s + \frac{3}{r} \right) \Delta$$

for all $h \in [k]$. The proof is complete. \square

PROOF OF LEMMA A.7. For any $g \neq h \in [k] \times [k]$,

$$\begin{aligned}
\mathbb{I} \left\{ z_i = g, \hat{z}_i^{(s+1)} = h \right\} &\leq \mathbb{I} \left\{ \|\theta_g + w_i - \hat{\theta}_h^{(s)}\|^2 \leq \|\theta_g + w_i - \hat{\theta}_g^{(s)}\|^2 \right\} \\
\text{(A.12)} \quad &= \mathbb{I} \left\{ \|\theta_g - \hat{\theta}_h^{(s)}\|^2 - \|\theta_g - \hat{\theta}_g^{(s)}\|^2 \leq 2 \left\langle w_i, \hat{\theta}_h^{(s)} - \hat{\theta}_g^{(s)} \right\rangle \right\}.
\end{aligned}$$

Triangle inequality implies

$$\|\theta_g - \hat{\theta}_h^{(s)}\|^2 \geq \left(\|\theta_g - \theta_h\| - \|\theta_h - \hat{\theta}_h^{(s)}\| \right)^2 \geq (1 - \Lambda_s)^2 \|\theta_g - \theta_h\|^2.$$

Using the fact that $(1-x)^2 - y^2 \geq (1-x-y)^2$ when $y(1-x-y) \geq 0$, we obtain

$$\text{(A.13)} \quad \|\theta_g - \hat{\theta}_h^{(s)}\|^2 - \|\theta_g - \hat{\theta}_g^{(s)}\|^2 = (1 - 2\Lambda_s)^2 \|\theta_g - \theta_h\|^2 \geq \epsilon^2 \|\theta_g - \theta_h\|^2.$$

Denote by $\Delta_h = \hat{\theta}_h^{(s)} - \theta_h$ for $h \in [k]$. Then,

$$\begin{aligned} & \mathbb{I} \left\{ z_i = g, \hat{z}_i^{(s+1)} = h \right\} \\ & \leq \mathbb{I} \left\{ \epsilon^2 \|\theta_g - \theta_h\|^2 \leq 2 \langle w_i, \theta_h - \theta_g + \Delta_h - \Delta_g \rangle \right\} \\ & \leq \mathbb{I} \left\{ \frac{\epsilon^2}{2} \|\theta_g - \theta_h\|^2 \leq 2 \langle w_i, \theta_h - \theta_g \rangle \right\} + \mathbb{I} \left\{ \frac{\epsilon^2}{2} \Delta^2 \leq 2 \langle w_i, \Delta_h - \Delta_g \rangle \right\}. \end{aligned}$$

Taking a sum over $i \in T_g^*$ and using Markov's inequality on the second term, we obtain

$$(A.14) \quad n_{gh}^{(s+1)} \leq \sum_{i \in T_g^*} \mathbb{I} \left\{ \frac{\epsilon^2}{4} \|\theta_g - \theta_h\|^2 \leq \langle w_i, \theta_h - \theta_g \rangle \right\} + \sum_{i \in T_g^*} \frac{16}{\epsilon^4 \Delta^4} (w_i'(\Delta_h - \Delta_g))^2$$

Note that $\mathbb{I} \left\{ \frac{\epsilon^2}{4} \|\theta_g - \theta_h\|^2 \leq \langle w_i, \theta_h - \theta_g \rangle \right\}$ are independent Bernoulli random variables. By Lemma A.5, the first term in RHS of (A.14) can be upper bounded by

$$(A.15) \quad n_g^* \exp \left(-\frac{\epsilon^4 \Delta^2}{32\sigma^2} \right) + \sqrt{5n_g^* \log n}.$$

By Lemma A.2, the second term in RHS of (A.14) can be upper bounded by

$$(A.16) \quad \sum_{i \in T_g^*} \frac{16}{\epsilon^4 \Delta^4} (w_i'(\Delta_h - \Delta_g))^2 \leq \frac{96(n_g^* + d)\sigma^2}{\epsilon^4 \Delta^4} \|\Delta_g - \Delta_h\|^2.$$

Combining (A.14), (A.15) and (A.16) and using the fact that $\|\Delta_g - \Delta_h\|^2 \leq 4\Lambda_s^2 \Delta^2$, we get

$$n_{gh}^{(s+1)} \leq n_g^* \exp \left(-\frac{\epsilon^4 \Delta^2}{32\sigma^2} \right) + \sqrt{5n_g^* \log n} + \frac{384(n_g^* + d)\sigma^2}{\epsilon^4 \Delta^2} \Lambda_s^2.$$

Consequently,

$$(A.17) \quad \max_{g \in [k]} \sum_{h \neq g} \frac{n_{gh}^{(s+1)}}{n_g^*} \leq k \exp \left(-\frac{\epsilon^4 \Delta^2}{32\sigma^2} \right) + k \sqrt{\frac{5 \log n}{\alpha n}} + \frac{384}{\epsilon^4 r^2} \Lambda_s^2.$$

Since $\Lambda_s \leq 1/2$ and $r \geq 20\epsilon^{-2}$, the RHS of (A.17) is smaller than 1/2 when $\alpha n \geq 32k^2 \log n$. Thus,

$$n_h^{(s+1)} \geq n_{hh}^{(s+1)} \geq \frac{1}{2} n_h^* \geq \frac{1}{2} \alpha n$$

for all $h \in [k]$ and we have

$$(A.18) \quad \max_{h \in [k]} \sum_{\substack{g \neq h \\ g \in [k]}} \frac{n_{gh}^{(s+1)}}{n_h^{(s+1)}} \leq \frac{2}{\alpha} \exp\left(-\frac{\epsilon^4 \Delta^2}{32\sigma^2}\right) + \sqrt{\frac{5k \log n}{\alpha^2 n}} + \frac{768}{\epsilon^4 r^2} \Lambda_s^2,$$

which, together with (A.17), implies

$$G_{s+1} \leq \exp\left(-\frac{\epsilon^4 \Delta^2}{32\sigma^2} + \log(2/\alpha)\right) + \sqrt{\frac{5k \log n}{\alpha^2 n}} + \frac{768}{\epsilon^4 r^2} \Lambda_s^2$$

Under the assumptions that $\epsilon^4 \alpha \Delta^2 / \sigma^2 \geq r^2 \epsilon^4 \geq 36$, we have the desired result (A.6). \square

A.3. Proof of Theorem 3.2.

PROOF OF (3.6) IN THEOREM 3.2. From Lemma A.6, a necessary condition for $\Lambda_0 \leq \frac{1}{2} - \frac{4}{\sqrt{r}}$ is $G_0 \leq (\frac{1}{2} - \frac{6}{\sqrt{r}}) \frac{1}{\lambda}$. Setting $\epsilon = \frac{7}{\sqrt{r}}$ in Lemma A.7, we have $G_1 \leq 0.35$. Plugging it into Lemma A.6 gives us $\Lambda_1 \leq 0.4$, under the assumption that $r \geq 16\sqrt{k}$. Then it can be easily proved by induction that $G_s \leq 0.35$ and $\Lambda_s \leq 0.4$ for all $s \geq 1$. Consequently, Lemma A.6 yields

$$\Lambda_s \leq \frac{3}{r} + \frac{3}{r} \sqrt{k G_s} + G_s \leq \frac{1}{2} + G_s$$

which, combined with (A.6), implies

$$G_{s+1} \leq \frac{C}{r^2} + \frac{C}{r^2} \left(\frac{1}{4} + 2G_s + G_s^2\right) + \sqrt{\frac{5k \log n}{\alpha^2 n}} \leq \frac{2C}{r^2} + \frac{3C}{r^2} G_s + \sqrt{\frac{5k \log n}{\alpha^2 n}}$$

for some constant C . Here we have chosen $\epsilon = 1/5$ in Lemma A.6 to get the first inequality. \square

PROOF OF (3.7) IN THEOREM 3.2. From the proof of Lemma A.6, the error of estimating θ_h at iteration s can be written as $\hat{\theta}_h^{(s)} - \theta_h = \frac{1}{n_h} W_{T_h^*} + u_h$, with

$$(A.19) \quad \|u_h\| \leq \left(\frac{3}{r} \sqrt{k G_s} + G_s\right) \Delta \leq \sqrt{G_s} \Delta$$

In addition, by Lemma A.6 and Lemma A.7, there is a constant C_1 such that

$$\Lambda_s \leq \frac{3}{r} + \sqrt{G_s} + 2G_s \Lambda_{s-1} \leq \frac{C_1}{r} + \frac{C_1}{r} \Lambda_{s-1} + 0.7 \Lambda_{s-1} + \left(\frac{C_1 k \log n}{\alpha^2 n}\right)^{1/4}$$

for all $s \geq 1$. Therefore, when r is large enough, we have

$$\Lambda \leq C_2 r^{-1} + C_2 \left(\frac{k \log n}{\alpha^2 n} \right)^{1/4}$$

for all $s \geq \log n$. Then by (A.13), we have

$$\mathbb{I} \left\{ z_i = g, \hat{z}_i^{(s+1)} = h \right\} \leq \mathbb{I} \left\{ \beta_1 \|\theta_g - \theta_h\|^2 \leq 2 \langle w_i, \theta_h - \theta_g + \Delta_h - \Delta_g \rangle \right\}$$

where $(1 - 2\Lambda_s)^2 \geq \beta_1 := 1 - 4C_2 r^{-1} - 4C_2 \left(\frac{k \log n}{\alpha^2 n} \right)^{1/4}$.

In order to prove that A_s attains convergence rates (3.7), we first upper bound the expectation of A_s and then derive the high probability bound using Markov's inequality. Similar to the two-mixture case, we need to upper bound the inner product $\langle w_i, \Delta_h - \Delta_g \rangle$ more carefully. Note that $\{T_h^*, h \in [k]\}$ are deterministic sets, we could use concentration equalities to upper bound $W_{T_h^*}$ and u_h parts separately.

Let $v_h = \frac{1}{n_h} W_{T_h^*}$ for $h \in [k]$ and we decompose $\mathbb{I} \left\{ z_i = g, \hat{z}_i^{(s+1)} = h \right\}$ into three terms.

$$\begin{aligned} \mathbb{I} \left\{ z_i = g, \hat{z}_i^{(s+1)} = h \right\} &\leq \mathbb{I} \left\{ \beta \|\theta_g - \theta_h\|^2 \leq 2 \langle w_i, \theta_h - \theta_g \rangle \right\} \\ &\quad + \mathbb{I} \left\{ \beta_2 \Delta^2 \leq 2 \langle w_i, u_h - u_g \rangle \right\} \\ &\quad + \mathbb{I} \left\{ \beta_4 \Delta^2 \leq 2 \langle w_i, v_h - v_g \rangle \right\}, \end{aligned}$$

where β_2 and β_4 will be specified later and $\beta = \beta_1 - \beta_2 - \beta_4$. Taking a sum over $h \in [k]$ and $i \in [n]$, we obtain

$$\mathbb{E}A_{s+1} \leq \mathbb{E}J_1 + \mathbb{E}J_2 + \mathbb{E}J_3$$

with

$$(A.20) \quad J_1 = \sum_{h \in [k]} \frac{1}{n} \sum_{i=1}^n \mathbb{I} \left\{ \beta \|\theta_{z_i} - \theta_h\|^2 \leq 2 \langle w_i, \theta_h - \theta_{z_i} \rangle \right\}$$

$$(A.21) \quad J_2 = \sum_{h \in [k]} \frac{1}{n} \sum_{i=1}^n \mathbb{I} \left\{ \beta_2 \Delta^2 \leq 2 \langle w_i, u_h - u_{z_i} \rangle \right\}.$$

$$(A.22) \quad J_3 = \sum_{h \in [k]} \frac{1}{n} \sum_{i=1}^n \mathbb{I} \left\{ \beta_4 \Delta^2 \leq 2 \langle w_i, v_{z_i} - v_h \rangle \right\}.$$

Let us first consider the expectation of J_1 . Using Chernoff's bound, we have

$$\mathbb{P} \left\{ \beta \|\theta_g - \theta_h\|^2 \leq 2 \langle w_i, \theta_h - \theta_g \rangle \right\} \leq \exp \left(-\frac{\beta^2 \|\theta_h - \theta_g\|^2}{8\sigma^2} \right) \leq \exp \left(-\frac{\beta^2 \Delta^2}{8\sigma^2} \right).$$

Thus,

$$\mathbb{E} J_1 \leq k \exp \left(-\frac{\beta^2 \Delta^2}{8\sigma^2} \right) = \exp \left(-\frac{\gamma \Delta^2}{8\sigma^2} \right),$$

with $\gamma = \beta^2 - \frac{8\sigma^2 \log k}{\Delta^2} \geq \beta^2 - 8/r^2$.

We use Markov Inequality to upper bound J_2 . Markov's inequality and Lemma A.2 give us

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \mathbb{I} \left\{ \beta_2 \Delta^2 \leq 2 \langle w_i, u_h - u_{z_i} \rangle \right\} &\leq \frac{4}{n\beta_2^2 \Delta^4} \sum_{g \in [k]} \sum_{i \in T_g^*} (w'_i (u_h - u_g))^2 \\ &\leq \frac{24\sigma^2}{n\beta_2^2 \Delta^4} \sum_{g \in [k]} (n_g^* + d) \|u_h - u_g\|^2. \end{aligned}$$

(A.19) implies

$$J_2 \leq \frac{96\sigma^2 G_s}{n\beta_2^2 \Delta^2} \sum_{h \in [k]} \sum_{g \in [k]} (n_g^* + d) \leq \frac{96\sigma^2 k(n + kd)}{\alpha n \beta_2^2 \Delta^2} A_s = \frac{12\sqrt{k}}{r} A_s.$$

Here the second inequality is due to the fact that $G_s \leq A_s/\alpha$. And we choose $\beta_2 = \sqrt{8k/r}$ in the last equality.

Finally, we upper bound the expectation of J_3 . Given $z_i = g$, we have

$$\begin{aligned} &\mathbb{P} \left\{ \beta_4 \Delta^2 \leq 2 \langle w_i, v_g - v_h \rangle \right\} \\ &\leq \mathbb{P} \left\{ \frac{\beta_4}{4} \Delta^2 \leq \langle w_i, v_g \rangle \right\} + \mathbb{P} \left\{ -\frac{\beta_4}{4} \Delta^2 \geq \langle w_i, v_h \rangle \right\} \\ &\leq \mathbb{P} \left\{ \frac{\beta_4}{8} \Delta^2 \leq \left\langle w_i, \frac{1}{n_g^*} W_{T_g^*} \right\rangle \right\} + \mathbb{P} \left\{ -\frac{\beta_4}{8} \Delta^2 \geq \left\langle w_i, \frac{1}{n_h^*} W_{T_h^*} \right\rangle \right\} \end{aligned}$$

Choosing $t = \max\{\frac{\sqrt{d}\Delta}{\sigma}, \frac{\Delta^2}{\sigma^2}\}$, $\delta = \exp\left(-\frac{\Delta^2}{4\sigma^2}\right)$ in Lemma A.3, and

$$\beta_4 = \frac{64}{r} \geq \frac{8}{\Delta^2} \left(\frac{3 \max\{\sqrt{d}\sigma\Delta, \Delta^2\}}{\sqrt{\alpha n}} + \frac{3\sigma^2 d + \Delta^2}{\alpha n} \right),$$

we obtain $\mathbb{P}\{\beta_4\Delta^2 \leq 2\langle w_i, v_g - v_h \rangle\} \leq 2\exp(-\Delta^2/(4\sigma^2))$, where we have used the assumption that $n_g^* \geq \alpha n$ and $\alpha n \geq 36r^2$. Thus,

$$\mathbb{E}J_3 \leq 2k \exp\left(-\frac{\Delta^2}{\sigma^2}\right),$$

Combining the pieces, we have

$$\begin{aligned} \mathbb{E}A_{s+1} &\leq \mathbb{E}[J_1] + \mathbb{E}[J_2\mathbb{I}\{\mathcal{E}\}] + \mathbb{E}[J_3] + \mathbb{P}\{\mathcal{E}^c\} \\ &\leq \exp\left(-\frac{\gamma\Delta^2}{8\sigma^2}\right) + \frac{12\sqrt{k}}{r}\mathbb{E}A_s + 2k \exp\left(-\frac{\Delta^2}{\sigma^2}\right), \end{aligned}$$

with $\gamma = (\beta_1 - \sqrt{8k/r} - 64/r)^2 - 8/r^2 = 1 - o(1)$. Here only prove the case that $r \rightarrow \infty$. For the finite case, all the $o(1)$ in the following proof can be substituted by a small constant.

$$\mathbb{E}A_s \leq \frac{1}{2^{s-\lceil \log r \rceil}} + 2 \exp\left(-\frac{(1-\eta)\Delta^2}{8\sigma^2}\right) + \frac{2}{n^3} \leq 2 \exp\left(-\frac{(1-\eta)\Delta^2}{8\sigma^2}\right) + \frac{3}{n^3}$$

when $s \geq 4 \log n$. By Markov's inequality, for any $t > 0$,

$$(A.23) \quad \mathbb{P}\{A_s \geq t\} \leq \frac{1}{t}\mathbb{E}A_s \leq \frac{2}{t} \exp\left(-\frac{(1-\eta)\Delta^2}{8\sigma^2}\right) + \frac{3}{n^3 t}.$$

If $(1-\eta)\frac{\Delta^2}{8\sigma^2} \leq 2 \log n$, choose $t = \exp\left(-\frac{(1-\eta-\frac{8\sigma}{\Delta})\Delta^2}{8\sigma^2}\right)$ and we have

$$\mathbb{P}\left\{A_s \geq \exp\left(-\frac{(1-\eta-\frac{8\sigma}{\Delta})\Delta^2}{8\sigma^2}\right)\right\} \leq \frac{4}{n} + 2 \exp\left(-\frac{\Delta}{\sigma}\right).$$

Otherwise, since A_s only takes discrete values of $\{0, \frac{1}{n}, \dots, 1\}$, choosing $t = \frac{1}{n}$ in (A.23) leads to

$$\mathbb{P}\{A_s > 0\} = \mathbb{P}\left\{A_s \geq \frac{1}{n}\right\} \leq 2n \exp(-2 \log n) + \frac{3}{n^2} \leq \frac{4}{n}.$$

The proof is complete. \square

A.4. Proof of Lower Bounds. The key difficulty in proving the lower bound is to deal with infinitum of all label permutations. Here we adapt the proof idea from [23]. We define a subset of the parameter space, in which a large portion of the labels in each cluster are fixed. Then any permutation other than identity gives us bigger mis-clustering rate.

PROOF OF THEOREM 3.3. For any $z \in [k]^n$, let us define $n_u(z) = \sum_{i=1}^n \mathbb{I}\{z_i = u\}$. Let $z^* \in [k]^n$ satisfying $n_1(z^*) \leq n_2(z^*) \leq \dots \leq n_k(z^*)$ with $n_1(z^*) = n_2(z^*) = \lfloor \alpha n \rfloor$. It is easy to check the existence of z^* . For each $u \in [k]$, we choose a subset of $\{i : z^*(i) = u\}$ with cardinality $\lceil n_u(z^*) - \frac{\alpha n}{4k} \rceil$, denoted by T_u . Define $T = \cup_{u=1}^k T_u$ and

$$(A.24) \quad \mathcal{Z}^* = \{z \in \mathcal{Z}_0, z_i = z_i^* \text{ for all } i \in T\}$$

A key observation is that for any $z \neq \tilde{z} \in \mathcal{Z}^*$, we have $\frac{1}{n} \sum_{i=1}^n \mathbb{I}\{z_i \neq \tilde{z}_i\} \leq \frac{k}{n} \frac{\alpha n}{4k} = \frac{\alpha}{4}$ and

$$(A.25) \quad \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{\pi(z_i) \neq \tilde{z}_i\} \geq \frac{1}{n} (\alpha n - \frac{\alpha n}{4k}) \geq \frac{\alpha}{2}$$

for all $\pi \in \mathcal{S}_k \neq \mathbb{I}_k$. Thus, $\ell(z, \tilde{z}) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{z_i \neq \tilde{z}_i\}$ for all $z, \tilde{z} \in \mathcal{Z}^*$. Then following the same arguments in the proof of Theorem 2 in [23], we can obtain

$$(A.26) \quad \inf_{\hat{z}} \sup_{z \in \mathcal{Z}} \mathbb{E} \ell(\hat{z}, z) \geq \frac{\alpha}{6} \frac{1}{|T^c|} \sum_{i \in T^c} \left[\frac{1}{2k^2} \inf_{\hat{z}_i} (\mathbb{P}_1\{\hat{z}_i = 2\} + \mathbb{P}_2\{\hat{z}_i = 1\}) \right]$$

Here $\mathbb{P}_t, t \in \{1, 2\}$ denote the probability distribution of our data given $z_i = t$. By Neyman-Pearson Lemma, the infimum of the right hand side of (A.26) is achieved by the likelihood ratio test

$$\hat{z}_i = \operatorname{argmin}_{g \in \{1, 2\}} \|y_i - \theta_g\|^2.$$

Thus,

$$\begin{aligned} \inf_{\hat{z}_i} \left(\frac{1}{2} \mathbb{P}_1\{\hat{z}_i = 2\} + \frac{1}{2} \mathbb{P}_2\{\hat{z}_i = 1\} \right) &= \mathbb{P}\{\|\theta_1 + w_i - \theta_2\|^2 \leq \|w_i\|^2\} \\ &= \mathbb{P}\{\|\theta_1 - \theta_2\|^2 \leq 2 \langle w_i, \theta_1 - \theta_2 \rangle\}. \end{aligned}$$

Since $w_{ij}, j \in [d]$ are independent $\mathcal{N}(0, \sigma^2)$, $\langle w_i, \theta_1 - \theta_2 \rangle \sim \mathcal{N}(0, \sigma^2 \|\theta_1 - \theta_2\|^2)$. Let $\Phi(t)$ be the cumulative function of $\mathcal{N}(0, 1)$ random variable. By calculating the derivatives, it can be easily proved that

$$1 - \Phi(t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2} dx \geq \frac{1}{\sqrt{2\pi}} \frac{t}{t^2 + 1} e^{-t^2/2}.$$

Then when $\|\theta_1 - \theta_2\| \geq \sigma$, we have

$$\mathbb{P}\left\{ \langle w_i, \theta_1 - \theta_2 \rangle \geq \frac{1}{2} \|\theta_1 - \theta_2\|^2 \right\} \geq \frac{\sigma}{\sqrt{2\pi} \|\theta_1 - \theta_2\|} \exp\left(-\frac{\|\theta_1 - \theta_2\|^2}{8\sigma^2}\right).$$

Consequently,

$$\inf_{\hat{z}} \sup_{z \in \mathcal{Z}} \mathbb{E} \ell(\hat{z}, z) \geq \exp \left(-\frac{\Delta^2}{8\sigma^2} - 2 \log \frac{2k\Delta}{\alpha\sigma} \right).$$

The proof is complete. \square

A.5. A counterexample. Now we give a counterexample to show that the initialization condition in Theorem 3.2 is almost necessary. Consider a noiseless case with 6 equal size clusters, as showed in Figure 3. Suppose $\|\theta_i - \theta_j\| = \Delta$ for $i \neq j \in [3]$ and $\|\theta_{i+3} - \theta_i\| = \lambda\Delta$ for $i \in [3]$. We are given an initializer that mixes cluster i and cluster $i + 3$, $i \in [3]$, with $m/(2\lambda)$ data points from cluster $i + 3$ and $m - m/(2\lambda)$ data points from cluster i for some integer m . Consequently, the estimated (initialized) center $\hat{\theta}_i^{(0)}$ lies in the middle of two true centers. For the next label update step, since there are ties, we may assign half points at θ_1 to cluster $\hat{\theta}_1^{(0)}$ and another half to cluster $\hat{\theta}_3^{(0)}$ and the estimated centers remain the same as before. Therefore, $\{\hat{\theta}_i^{(0)}, i \in [3]\}$ is a stationary point for Lloyd's algorithm and it may not convergence to the true centers. We would like to note that this counterexample is a worst case in theoretical analysis, which may not happen in practice.

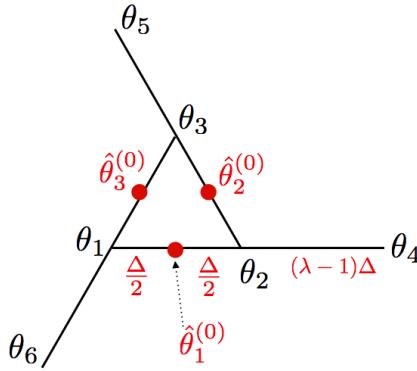


Fig 3: A counter example showing that Lloyd's algorithm may not converge when $G_0 = \frac{1}{2\lambda}$ or $\Lambda_0 = \frac{1}{2}$.

A.6. Proofs of Random Initialization.

PROOF OF THEOREM 6.1. For a data independent initializer, we have the following result.

LEMMA A.8. *For any $\delta < 1/4$, if we have a data independent initializer satisfying*

$$(A.27) \quad A_0 \leq \frac{1}{2} - \frac{\sigma}{\|\theta^*\|} \sqrt{\frac{2\log(1/\delta)}{n}} \left(1 + \frac{18\sqrt{d}}{r}\right)$$

the conclusion of Theorem 3.1 continues to hold with probability greater than $1 - \delta$.

Given Lemma A.8, let us study the condition under which completely random initialization falls into the basin of attraction (A.27). When we randomly assign initial labels, we have $|A_0 - 1/2| \geq n^{-1/2}$ with probability greater than 0.3. We can boost this probability to $1 - 1/\delta$ by independently drawing $3\log(1/\delta)$ different random initializers. Recall our loss function (2.4). We can assume $A_0 < 1/2 - n^{-1/2}$. Otherwise, we may flip all the labels in the initial step. Combining this with condition (A.27), we have the following corollary.

Now it remains to prove Lemma A.8. We focus on the first iteration. By (7.3), we have $\hat{\theta}^{(1)} - \theta^* = -2A_0\theta^* + R$ with $R = \frac{1}{n} \sum_{i=1}^n (1 - 2c_i)w_i$. Then

$$\begin{aligned} \mathbb{I}\{\hat{z}_i^{(1)} \neq z_i\} &= \mathbb{I}\{\langle \theta^* + w_i, (1 - 2A_0)\theta^* + R \rangle \leq 0\} \\ &\leq \mathbb{I}\left\{\langle w_i, \theta^* \rangle \leq -\frac{1}{2}\|\theta^*\|^2\right\} + \mathbb{I}\left\{\frac{\langle R, \theta^* + w_i \rangle}{1 - 2A_0} \leq -\frac{1}{2}\|\theta^*\|^2\right\}. \end{aligned}$$

Since the initializer is data independent, $\{c_i\} \subseteq \{-1, 1\}^n$ are independent of $\{w_i\}$. Chernoff's bound implies

$$\langle \theta^*, R \rangle = \frac{1}{n} \sum_{j=1}^n (1 - 2c_j)w_j \theta^* \geq -\frac{\sigma\|\theta^*\|\sqrt{2\log(1/\delta)}}{\sqrt{n}}$$

with probability greater than $1 - \delta$. Consequently, we obtain

$$(A.28) \quad A_1 \leq \frac{1}{n} \sum_{i=1}^n \mathbb{I}\left\{\langle w_i, \theta^* \rangle \leq -\frac{1}{2}\|\theta^*\|^2\right\} + \frac{1}{n} \sum_{i=1}^n \mathbb{I}\left\{\langle R, w_i \rangle \leq -\beta\|\theta^*\|^2\right\},$$

where $\beta = \frac{1}{2} - A_0 - \frac{\sigma\sqrt{2\log(1/\delta)}}{\sqrt{n}\|\theta^*\|}$. By Chernoff's bound and Hoeffding's inequality, the first term in the RHS of (A.28) can be upper bounded by

$$\exp\left(-\frac{\|\theta^*\|^2}{8\sigma^2}\right) + \sqrt{\frac{\log(1/\delta)}{n}}$$

with probability greater than $1 - \delta$. Markov's inequality and proof of Lemma 7.3 give us

$$\frac{1}{n} \sum_{i=1}^n \mathbb{I}\{\langle R, w_i \rangle \leq -\beta \|\theta^*\|^2\} \leq \frac{1}{n\beta^2 \|\theta^*\|^4} \sum_{i=1}^n (R' w_i)^2 \leq \frac{9\sigma^2(1+d/n)}{\beta^2 \|\theta^*\|^4} \|R\|^2.$$

with probability greater than $1 - \exp(-n/2)$. From (D.3), we have $\|R\|^2 \leq \frac{4\sigma^2 d \log(1/\delta)}{n}$ with probability greater than $1 - \delta$. Combining the pieces, we obtain

$$(A.29) \quad A_1 \leq \exp\left(-\frac{\|\theta^*\|^2}{8\sigma^2}\right) + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{36\sigma^4(1+d/n)}{\beta^2 \|\theta^*\|^4} \frac{d \log(1/\delta)}{n}$$

with probability greater than $1 - 3\delta$. Under the condition that

$$A_0 \leq \frac{1}{2} - \frac{\sigma}{\|\theta^*\|} \sqrt{\frac{2 \log(1/\delta)}{n}} \left(1 + \frac{18\sigma \sqrt{d(1+d/n)}}{\|\theta^*\|}\right),$$

we have $A_1 \leq \frac{1}{4}$ with probability greater than $1 - 3\delta$. Then using the result of Theorem 3.1, the proof is complete. \square

APPENDIX B: PROOFS OF COMMUNITY DETECTION

B.1. Proof of Theorem 4.1. We use the same notation as in the Gaussian case. Let T_g^* be the true cluster g and $T_g^{(s)}$ be the estimated cluster g at iteration s . And we will drop the dependency of s when there is no ambiguity in the context. A_s is the mis-clustering rate at iteration s and G_s is the group-wise mis-clustering rate at iteration s . Similar to the analysis of Gaussian mixture model, the proof consists of two steps. We first prove that given $G_0 \leq \frac{1}{2} - \epsilon_0$ with $\epsilon_0 \geq \frac{C_0 \sqrt{a\beta k} \log(\beta k)}{a-b}$ for a sufficiently large constant C_0 , we have $G_s \leq \frac{1}{3}$ for all $1 \leq s \leq \lfloor 3 \log n \rfloor$ with high probability. Then we prove the mis-clustering rate A_s geometrically decays to the rate (4.4).

Given $z_i = g$, we decompose (4.2) as follows.

$$\begin{aligned} \frac{1}{n_h} \sum_{j=1}^n A_{ij} \mathbb{I}\{\hat{z}_j^{(s)} = h\} &= \frac{1}{n_h} \sum_{j=1}^n w_{ij} \mathbb{I}\{\hat{z}_j^{(s)} = h\} + \frac{1}{n_h} \sum_{j=1}^n \sum_{l=1}^k B_{gl} \mathbb{I}\{\hat{z}_j^{(s)} = h, z_j = l\} \\ &= \frac{1}{n_h} \sum_{j \in T_h} w_{ij} + \frac{a}{n} \frac{n_{gh}}{n_h} + \frac{b}{n} \frac{\sum_{l \neq g} n_{lh}}{n_h} \\ &= \frac{1}{n_h} W_i(T_h) + \frac{a}{n} \frac{n_{gh}}{n_h} + \frac{b}{n} \left(1 - \frac{n_{gh}}{n_h}\right) \end{aligned}$$

Consequently, the error of estimating z_i at iteration $s + 1$ can be upper bounded as

$$\begin{aligned}
& \mathbb{I}\{\hat{z}_i^{(s+1)} \neq g, z_i = g\} \\
\leq & \mathbb{I}\left\{\max_{h \neq g} \sum_{j=1}^n A_{ij} \mathbb{I}\{\hat{z}_j^{(s)} = h\} \geq \sum_{j=1}^n A_{ij} \mathbb{I}\{\hat{z}_j^{(s)} = g\}, z_i = g\right\} \\
\leq & \mathbb{I}\left\{\max_{h \neq g} \frac{1}{n_h} W_i(T_h) - \frac{1}{n_g} W_i(T_g) \geq \frac{a-b}{n} \left(\frac{n_{gg}}{n_g} - \max_{h \neq g} \frac{n_{gh}}{n_h}\right)\right\} \\
\leq & \mathbb{I}\left\{\max_{h \neq g} \frac{1}{n_h} W_i(T_h) - \frac{1}{n_g} W_i(T_g) \geq \frac{2\epsilon(a-b)}{n}\right\} \\
\text{(B.1)} \leq & \mathbb{I}\left\{\max_{h \neq g} \frac{1}{n_h} W_i(T_h) \geq \frac{\epsilon(a-b)}{n}\right\} + \mathbb{I}\left\{\frac{1}{n_g} W_i(T_g) \leq -\frac{\epsilon(a-b)}{n}\right\}
\end{aligned}$$

where the second inequality is due to our induction assumption that $n_{gg} \geq (1 - 1/2 + \epsilon)n_g$ and $n_{gh} \leq (1/2 - \epsilon)n_h$. Union bound implies

$$\text{(B.2)} \quad \mathbb{I}\{\hat{z}_i^{(s+1)} \neq z_i\} \leq \sum_{h=1}^k \mathbb{I}\left\{W_i(T_h) \geq \frac{\epsilon(a-b)n_h}{n}\right\} + \mathbb{I}\left\{W_i(T_{z_i}) \leq -\frac{\epsilon(a-b)n_{z_i}^*}{n}\right\}.$$

Now we give two upper bounds on the sum of indicator variables uniformly over $\{T_h\}$. These proofs are deferred to Section B.2.

LEMMA B.1. *Given $T^* \subseteq [n]$ with cardinality greater than $n/(\beta k)$ and $\epsilon > 0$ such that $\epsilon^2(a-b)^2 \geq C_0 \beta a k \log(n/m)$ for a sufficiently large constant C_0 . There is an universal constant C such that with probability greater than $1 - n^{-2}$,*

$$\text{(B.3)} \quad \sum_{i=1}^m \mathbb{I}\left\{\sum_{j \in T} w_{ij} \geq \frac{\epsilon(a-b)|T|}{n}\right\} \leq m \exp\left(-\frac{\epsilon^2(a-b)^2}{C \beta a k}\right) + \frac{|T \Delta T^*|}{5}$$

holds for all $T \subseteq [n]$ with $|T \Delta T^*| \leq \frac{1}{2}|T^*|$.

LEMMA B.2. *Given $T^* \subseteq [n]$ with cardinality greater than $n/(\beta k)$ and $\epsilon > 0$ such that $\epsilon^2(a-b)^2 \geq C_0 \beta^2 a k^2 \log(n/m)$ for a sufficiently large constant C_0 . There is an universal constant C such that with probability greater than $1 - n^{-2}$,*

$$\text{(B.4)} \quad \sum_{i=1}^m \mathbb{I}\left\{\sum_{j \in T} w_{ij} \geq \frac{\epsilon(a-b)|T|}{n}\right\} \leq m \exp\left(-\frac{\epsilon^2(a-b)^2}{C \beta^2 a k^2}\right) + \frac{|T \Delta T^*|}{4 \beta k}$$

holds for all $T \subseteq [n]$ with $|T \Delta T^*| \leq \frac{1}{2}|T^*|$.

In Lemma B.1 and Lemma B.2, we present results for the upper tail $\sum_{i=1}^m \mathbb{I}\{W_i(T_h) \geq \epsilon(a-b)|T|/n\}$. By slightly modifying the proof, the same results hold for the lower tail $\sum_{i=1}^m \mathbb{I}\{W_i(T_h) \leq -\epsilon(a-b)|T|/n\}$. Taking an sum over $i \in T_g^*$ in (B.2) and using Lemma B.2, we obtain

$$\begin{aligned}
 & \sum_{i \in T_g^*} \mathbb{I}\{\hat{z}_i^{(s+1)} \neq z_i\} \\
 & \leq \sum_{h \neq g} \sum_{i \in T_g^*} \mathbb{I}\left\{W_i(T_h) \geq \frac{\epsilon(a-b)n_h}{k}\right\} + \sum_{i \in T_g^*} \mathbb{I}\left\{W_i(T_g) \leq -\frac{\epsilon(a-b)n_h}{k}\right\} \\
 & \leq kn_g^* \exp\left(-\frac{\epsilon^2(a-b)^2}{C\beta^2 ak^2}\right) + \frac{1}{4\beta k} \sum_{h=1}^k |T_h \Delta T_h^*| \\
 & \leq kn_g^* \exp\left(-\frac{\epsilon^2(a-b)^2}{C\beta^2 ak^2}\right) + \frac{n}{2\beta k} A_s.
 \end{aligned}$$

with probability greater than $1 - kn^{-2}$. Consequently,

$$\frac{1}{n_g^*} \sum_{h \neq g} n_{gh}^{(s+1)} \leq k \exp\left(-\frac{\epsilon^2(a-b)^2}{C\beta^2 ak^2}\right) + \frac{1}{2} G_s.$$

with probability greater than $1 - kn^{-2}$. Here we have used the fact that $A_s \leq G_s$. Using similar arguments, we obtain the same high probability upper bound for $\frac{1}{n_h} \sum_{g \neq h} n_{gh}^{(s+1)}$. Therefore, we have

$$G_{s+1} \leq k \exp\left(-\frac{\epsilon^2(a-b)^2}{C\beta^2 ak^2}\right) + \frac{1}{2} G_s$$

with probability greater than $1 - k^2 n^{-2}$. When $s = 0$, we choose $\epsilon = \epsilon_0$. Since $\epsilon_0 \geq \frac{C_0 \sqrt{a\beta k}}{a-b}$, we have $G_1 \leq 1/12 + G_0/2 \leq 1/3$ when $C_0 \geq (C+1) \log 12$. For $s \geq 1$, we choose $\epsilon = 1/3$ and it is straight forward to prove by induction that $G_{s+1} \leq 1/12 + 1/6 \leq 1/3$ for all $0 \leq s \leq 3 \log n$, with probability greater than $1 - n^{-1}$, provided $n \geq 3k^2 \log n$.

Now we are ready to give the convergence rate of A_s . Since $G_s \leq 1/3$ for all $s \in [1, 3 \log n]$, (B.2) holds for all $s \in [1, 3 \log n]$. Taking average over $i \in [n]$

in (B.2) and using Lemma B.1 with $\epsilon = 1/3$, we obtain

$$\begin{aligned}
A_{s+1} &\leq \frac{1}{n} \sum_{h=1}^k \sum_{i=1}^n \mathbb{I} \left\{ W_i(T_h) \geq \frac{a-b}{6k} \right\} + \frac{1}{n} \sum_{g=1}^k \sum_{i \in T_g^*} \mathbb{I} \left\{ W_i(T_g) \leq -\frac{a-b}{6k} \right\} \\
&\leq 2k \exp \left(-\frac{(a-b)^2}{C\beta ak^2} \right) + \frac{2}{5n} \sum_{h=1}^k |T_h \Delta T_h^*| \\
&\leq \exp \left(-\frac{(a-b)^2}{2C\beta ak^2} \right) + \frac{4}{5} A_s
\end{aligned}$$

with probability greater than $1 - kn^{-2}$, where C is some universal constant.

B.2. Proof of Lemma B.1 and Lemma B.2. For any fixed T with $|T \Delta T^*| \leq \gamma n/k$, let

$$a_i(T) = \mathbb{I} \left\{ \sum_{j \in T} w_{ij} \geq \frac{\epsilon(a-b)|T|}{n} \right\}.$$

By Bernstein's inequality, the success probability of $a_i(T)$ is upper bounded by

$$\begin{aligned}
\mathbb{P} \left\{ \sum_{j \in T} w_{ij} \geq \frac{\epsilon(a-b)|T|}{n} \right\} &\leq \exp \left(-C_1 \min \left\{ \frac{\epsilon(a-b)|T|}{n}, \frac{\epsilon^2(a-b)^2|T|}{an} \right\} \right) \\
&\leq \exp \left(-\frac{C_2 \epsilon^2 (a-b)^2}{\beta ak} \right),
\end{aligned}$$

for some universal constant C_1 and C_2 . Here the last inequality is due to the fact that $|T| \geq |T^*| - |T \Delta T^*| \geq n/(2\beta k)$. Note that $\{a_i(T), i \in [n]\}$ are independent Bernoulli random variables. By Bennett's inequality (see Lemma 9 in [22]), for any $t > 0$,

$$\begin{aligned}
\mathbb{P} \left\{ \sum_{i=1}^m a_i(T) \geq mp + t \right\} &\leq \exp \left(t - (mp + t) \log \left(1 + \frac{t}{mp} \right) \right) \\
\text{(B.5)} \qquad \qquad \qquad &\leq \exp \left(-t \log \left(\frac{t}{emp} \right) \right),
\end{aligned}$$

where

$$p = \frac{1}{n} \sum_{i=1}^n \mathbb{E} a_i(T) \leq \exp \left(-\frac{C_2 \epsilon^2 (a-b)^2}{\beta ak} \right) \triangleq \exp(-R).$$

To prove (B.3) holds for all $T \subseteq [n]$ with $|T\Delta T^*| \leq \frac{\gamma n}{k}$, we need a chaining argument. For $s = m \exp(-R/30)$, define $\mathcal{D}_0 = \{T \subseteq [n], |T\Delta T^*| \leq s\}$ and

$$\mathcal{D}_r = \{T \subseteq [n], 2^{r-1}s \leq |T\Delta T^*| \leq 2^r s\}.$$

Then $\mathcal{D} \subseteq \cup_{r=0}^u \mathcal{D}_r$ with $u = \lceil \log n \rceil$. Union bound implies

$$\begin{aligned} \mathcal{A} &\triangleq \mathbb{P} \left\{ \exists T \in \mathcal{D}, \text{ s.t. } \sum_{i=1}^m a_i(T) \geq mp + \frac{1}{5}|T\Delta T^*| + s \right\} \\ &\leq \sum_{r=0}^u \mathbb{P} \left\{ \exists T \in \mathcal{D}_r, \text{ s.t. } \sum_{i=1}^m a_i(T) \geq mp + \frac{1}{5}|T\Delta T^*| + s \right\} \\ &\leq \sum_{r=0}^u |\mathcal{D}_r| \max_{T \in \mathcal{D}_r} \mathbb{P} \left\{ \sum_{i=1}^m a_i(T) \geq mp + \frac{1}{5}|T\Delta T^*| + s \right\} \end{aligned}$$

Note that $T = T^* \setminus (T^* \cap T^c) \cup (T \cap (T^*)^c)$. Given T^* and $|T\Delta T^*| = u$, there are at most $\binom{n}{u} 2^u$ possible choices of T . Thus, the cardinality of \mathcal{D}_r is upper bounded by

$$\sum_{u=0}^{2^r s} \binom{n}{u} 2^u \leq \sum_{u=0}^{2^r s} \left(\frac{2en}{u} \right)^u \leq 2^r s \exp \left(2^r s \log \left(\frac{2en}{2^r s} \right) \right) \leq \exp \left(2^r s \log \left(\frac{en}{2^{r-2}s} \right) \right).$$

Combining this with (B.5), we obtain

$$\begin{aligned} &|\mathcal{D}_0| \max_{T \in \mathcal{D}_0} \mathbb{P} \left\{ \sum_{i=1}^m a_i(T) \geq mp + \frac{1}{5}|T\Delta T^*| + s \right\} \\ &\leq \exp \left(-s \log \left(\frac{s}{emp} \right) + s \log \left(\frac{2en}{s} \right) \right) \\ &\leq \exp \left(-\frac{Rs}{2} \right) \end{aligned}$$

and

$$\begin{aligned} &|\mathcal{D}_r| \max_{T \in \mathcal{D}_r} \mathbb{P} \left\{ \sum_{i=1}^m a_i(T) \geq mp + \frac{1}{5}|T\Delta T^*| + s \right\} \\ &\leq \exp \left(-\frac{2^r}{10} s \log \left(\frac{2^r s}{10emp} \right) + 2^r s \log \left(\frac{en}{2^{r-2}s} \right) \right) \\ &\leq \exp \left(-\frac{2^r Rs}{20} \right) \end{aligned}$$

for all $r \geq 1$, where we have used the assumption that $\frac{(a-b)^2}{\beta ak} \geq C \log \frac{n}{m}$ for a sufficiently large constant C . Thus,

$$\mathcal{A} \leq \exp\left(-\frac{Rs}{2}\right) + \sum_{r=1}^u \exp\left(-\frac{2^r Rs}{20}\right) \leq 3 \exp\left(-\frac{Rs}{10}\right).$$

By our assumption, we have $R = \frac{C_2 \epsilon^2 (a-b)^2}{\beta ak} \geq \frac{3}{4}R + 30 \log \frac{n}{m}$. Consequently, $\mathcal{A} \leq 3 \exp\left(-3s \log \frac{n}{s}\right) \leq n^{-2}$. Therefore, the proof Lemma B.1 is complete. Lemma B.2 follows from almost identical arguments except that we choose $s = m \exp\left(-\frac{R}{8\beta k}\right)$.

APPENDIX C: PROOFS OF CROWDSOURCING

C.1. Proof of Theorem 4.2. Suppose $z_j = g$, then Chernoff bound gives us

$$\begin{aligned} \mathbb{P}\left\{\hat{z}_j^{(0)} \neq z_j\right\} &= \mathbb{P}\left\{\exists h \neq g, \sum_{i=1}^m \mathbb{I}\{X_{ij} = h\} \geq \sum_{i=1}^m \mathbb{I}\{X_{ij} = g\}\right\} \\ &\leq \sum_{h \neq g} \mathbb{P}\left\{\sum_{i=1}^m (\mathbb{I}\{X_{ij} = h\} - \mathbb{I}\{X_{ij} = g\}) > 0\right\} \\ &\leq \sum_{h \neq g} \prod_{i=1}^m \mathbb{E} \exp(\lambda \mathbb{I}\{X_{ij} = h\} - \lambda \mathbb{I}\{X_{ij} = g\}). \end{aligned}$$

for all $\lambda > 0$. Since X_{ij} is from a multinomial distribution $(\pi_{ig1}, \pi_{ig2}, \dots, \pi_{igk})$, we have

$$\begin{aligned} &\prod_{i=1}^m \mathbb{E} \exp(\lambda \mathbb{I}\{X_{ij} = h\} - \lambda \mathbb{I}\{X_{ij} = g\}) \\ &= \prod_{i=1}^m \left(\pi_{igh} e^\lambda + \pi_{igg} e^{-\lambda} + (1 - \pi_{igh} - \pi_{igg})\right) \\ &\stackrel{(a)}{\leq} \prod_{i=1}^m \exp\left(\pi_{igh} e^\lambda + \pi_{igg} e^{-\lambda} - \pi_{igh} - \pi_{igg}\right) \\ &= \exp\left(Ae^\lambda + Be^{-\lambda} - A - B\right), \end{aligned}$$

where $A = \sum_{i=1}^m \pi_{igh}$ and $B = \sum_{i=1}^m \pi_{igg}$. Here inequality (a) is due to the fact that $1 + x \leq e^x$ for all x . Choosing $\lambda = 0.5 \log(B/A)$ yields

$$\begin{aligned} \mathbb{P} \left\{ \hat{z}_j^{(0)} \neq z_j \right\} &\leq \sum_{h \neq g} \exp \left(- \left(\sqrt{\sum_{i=1}^m \pi_{igg}} - \sqrt{\sum_{i=1}^m \pi_{igh}} \right)^2 \right) \\ &\leq k \exp(-V(\pi)) \end{aligned}$$

where $V(\pi) = \min_{h \neq g} \left(\sqrt{\sum_{i=1}^m \pi_{igg}} - \sqrt{\sum_{i=1}^m \pi_{igh}} \right)^2$. Then using Markov's inequality, we have

$$\mathbb{P} \left\{ \frac{1}{n} \sum_{j=1}^n \mathbb{I} \{ \hat{z}_j^{(0)} \neq z_j \} \geq t \right\} \leq \exp(-V(\pi) + \log k - \log t)$$

for all $t > 0$. A choice of $t = 1/(4k)$ completes the proof.

C.2. Proof of Corollary 4.2. Let us first calculate the sub-Gaussian parameter of w_i . For any $\lambda > 0$ and $a = (a_{11}, a_{12}, \dots, a_{mk}) \in \mathbb{R}^{mk}$ with $\|a\| = 1$,

$$\mathbb{E} \exp(\lambda \langle a, w_i \rangle) = \prod_{u=1}^m \mathbb{E} \exp \left(\lambda \sum_{h=1}^k w_{uih} a_{uh} \right).$$

For any fixed u , define a Multinoulli random variable X such that

$$\mathbb{P}\{X = a_{uh}\} = \pi_{uzih}, h \in [k].$$

Recall the definition of w_{uih} that $w_{uih} = \mathbb{I}\{X_{ui=h}\} - \pi_{uzih}$. Then,

$$\mathbb{E} \exp \left(\lambda \sum_{h=1}^k w_{uih} a_{uh} \right) = \mathbb{E} e^{\lambda(X - \mathbb{E}X)} \leq e^{2\lambda^2 \max_{h \in [k]} a_{uh}^2},$$

where the last inequality is because $|X| \leq \max_{h \in [k]} |a_{uh}|$ is a bounded random variable. Consequently, we have

$$\mathbb{E} \exp(\lambda \langle a, w_i \rangle) \leq \mathbb{E} \exp \left(2\lambda^2 \sum_{u=1}^m \max_{h \in [k]} a_{uh}^2 \right) \leq \mathbb{E} \exp(2\lambda^2).$$

Therefore, the sub-Gaussian parameter of w_i is upper bounded by 2.

APPENDIX D: PROOFS OF TECHNICAL LEMMAS

To help readers better understand the results, we first give proofs for spherical Gaussians and then extend it to the general sub-Gaussians. Without loss of generality, we assume $\sigma = 1$ in this section.

D.1. Proofs for spherical Gaussians. We denote by $\mathcal{N}(0, \Sigma)$ the normal random vectors with mean zero and covariance matrix Σ . For $Z \sim \mathcal{N}(0, 1)$, we have the following two facts on its moment generating functions.

$$(D.1) \quad \mathbb{E}e^{\lambda Z_1} = e^{\frac{\lambda^2}{2}}.$$

$$(D.2) \quad \mathbb{E}e^{\lambda Z_1^2} = \frac{1}{\sqrt{1-2\lambda}} \quad \text{for all } \lambda < 1/2.$$

PROOF OF LEMMA 7.1. For any fixed $S \subseteq [n]$, $W_S = \sum_{i \in S} w_i \sim \mathcal{N}(0, |S|I_d)$. By (D.2), we have

$$\mathbb{E}e^{\lambda \|W_S\|^2} = (1 - 2\lambda|S|)^{-d/2}$$

for $\lambda < 1/(2|S|)$. Then Chernoff's bound yields

$$(D.3) \quad \mathbb{P} \{ \|W_S\|^2 \geq t \} \leq \exp \left(-\lambda t - \frac{d}{2} \log(1 - 2\lambda|S|) \right).$$

Choosing $\lambda = 0.49/|S|$ and $t = 1.62(n + 4d)|S|$,

$$\mathbb{P} \{ \|W_S\|^2 \geq 1.62(n + 4d)|S| \} \leq \exp \left(-0.7938n - 3.17d + \frac{d}{2} \log 50 \right).$$

Since there are 2^n subsets of $[n]$, an union bound argument gives us

$$\mathbb{P} \left\{ \max_S \left(\|W_S\| - \sqrt{1.62(n + 4d)|S|} \right) \geq 0 \right\} \leq 2^n \exp(-0.7938n) \leq \exp(-0.1n).$$

□

PROOF OF LEMMA 7.2. Since $w_i' \theta^* \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \|\theta^*\|^2)$, we have $\bar{w}' \theta^* \sim \mathcal{N}(0, \|\theta^*\|^2/n)$. Consequently,

$$\mathbb{P} \left\{ -\bar{w}' \theta^* \geq \frac{\|\theta^*\|^2}{\sqrt{n}} \right\} \leq \exp \left(-\frac{\|\theta^*\|^2}{2} \right).$$

Choosing $\lambda = 0.4/n$ in (D.3), we have

$$\mathbb{P} \{ \|\bar{w}\|^2 \geq t \} \leq \exp \left(-0.4nt + \frac{d}{2} \log 5 \right).$$

A choice of $t = \frac{1}{n} (3d\sigma^2 + \|\theta^*\|^2)$ yields the desired result. □

PROOF OF LEMMA 7.3. Let $A = [w_1, w_2, \dots, w_n]$. Then

$$\sup_{\|a\|=1} \sum_{i=1}^n (a'w_i)^2 = \sup_{\|a\|=1} \|Aa\|_2^2.$$

Let \mathcal{B}_1 be the unit ball in \mathbb{R}^d and \mathcal{C} be a ϵ -net of \mathcal{B}_1 such that for any $a \in \mathcal{B}_1$, there is a $b \in \mathcal{C}$ satisfying $\|a - b\| \leq \epsilon$. Then we have $|\mathcal{C}| \leq (1 + 2/\epsilon)^d$ [50, Lemma 4.1] and

$$\begin{aligned} \|Aa\| &\leq \|Ab\| + \|A(a - b)\| \\ &\leq \max_{b \in \mathcal{C}} \|Ab\| + \|a - b\| \sup_{a \in \mathcal{B}_1} \|Aa\| \\ &\leq \max_{b \in \mathcal{C}} \|Ab\| + \epsilon \sup_{a \in \mathcal{B}_1} \|Aa\|. \end{aligned}$$

Taking a supreme over $a \in \mathcal{B}_1$ on both sides and rearranging, we get

$$(D.4) \quad \sup_{a \in \mathcal{B}_1} \|Aa\| \leq \frac{1}{1 - \epsilon} \max_{b \in \mathcal{C}} \|Ab\|.$$

For any fixed $b \in \mathcal{B}_1$, $b'w_i \stackrel{i.i.d}{\sim} \mathcal{N}(0, 1)$. Then by (D.2),

$$\mathbb{P} \left\{ \sum_{i=1}^n (b'w_i)^2 \geq t \right\} \leq e^{-0.25t} \prod_{i=1}^n \mathbb{E} e^{0.25(b'w_i)^2} \leq \exp(-0.25t + 0.35n),$$

where we have set $\lambda = 0.25$. Then (D.4) and union bound give us

$$\begin{aligned} \mathbb{P} \left\{ \sup_{a \in \mathcal{B}_1} \|Aa\|^2 \geq t \right\} &\leq \sum_{b \in \mathcal{C}} \mathbb{P} \{ \|Ab\|^2 \geq (1 - \epsilon)^2 t \} \\ &\leq \exp(-0.25(1 - \epsilon)^2 t + 0.35n + d \log(1 + 2/\epsilon)). \end{aligned}$$

A choice of $\epsilon = 0.05$ and $t = 2(n + 9d)$ yields the desired result. \square

PROOF OF LEMMA 7.4. For each $i \in [n]$, define $G_i = S \cap \{i\}^c$. Then,

$$(D.5) \quad \left\langle w_i, \sum_{l \in S} w_l \right\rangle \leq \left\langle w_i, \sum_{l \in G_i} w_l \right\rangle + \|w_i\|^2$$

$W_{G_i} = \sum_{l \in G_i} w_l$ is an isotopic Gaussian random vector with variance smaller than $|S|$ on every direction. By (D.1) and (D.2), for any independent $Z_1 \sim N(0, 1)$ and $Z_2 \sim N(0, 1)$,

$$\mathbb{E} e^{\lambda Z_1 Z_2} = \mathbb{E} e^{\frac{\lambda^2 Z_1^2}{2}} = \frac{1}{\sqrt{1 - \lambda^2}}$$

for all $\lambda \in [0, 1)$. Since w_i and W_{G_i} are independent, Chernoff bound implies

$$\mathbb{P} \left\{ \langle w_i, W_{G_i} \rangle \geq \sqrt{|S|t} \right\} \leq \exp \left(-\lambda t - \frac{d}{2} \log(1 - \lambda^2) \right) \leq \exp(-\lambda t + \lambda^2 d).$$

for $\lambda \leq 1/2$. Choosing $\lambda = \min\{\frac{t}{2d}, \frac{1}{2}\}$ yields

$$(D.6) \quad \mathbb{P} \left\{ \langle w_i, W_{G_i} \rangle \geq \sqrt{|S|t} \right\} \leq \exp \left(-\min \left\{ \frac{t^2}{4d}, \frac{t}{4} \right\} \right).$$

From (D.3), a choice of $\lambda = 1/3$ gives us

$$\mathbb{P} \left\{ \|w_i\|_2^2 \geq t \right\} \leq \exp \left(-\frac{t}{3} + 0.55d \right).$$

Setting $t = 3(d + 4 \log n)$, we obtain

$$(D.7) \quad \mathbb{P} \left\{ \|w_i\|_2^2 \geq 3\sigma^2(d + \log(1/\delta)) \right\} \leq \delta$$

with probability greater than $1 - \delta$. Combining (D.1), (D.6) and (D.7) completes the proof. \square

D.2. Proofs of sub-Gaussians.

PROOF OF LEMMA A.1. For any fixed $S \subseteq [n]$, W_S is a d dimensional random vector satisfying

$$\mathbb{E} e^{\langle a, W_S \rangle} = \prod_{i \in S} \mathbb{E} e^{\langle a, w_i \rangle} \leq e^{\frac{|S| \|a\|^2}{2}}$$

for all $a \in \mathbb{R}^d$. By [29, Theorem 2.1], we have

$$(D.8) \quad \mathbb{P} \left\{ \|W_S\|^2 \leq |S|(d + 2\sqrt{dt} + 2t) \right\} \leq \exp(-t).$$

Note that there are at most 2^n possible choices of S . Using the union bound completes the proof. \square

PROOF OF LEMMA A.2. For any fixed $b \in \mathbb{R}^d$ with $\|b\| = 1$, $b'w'_i$ s are independent sub-gaussian random variables with parameter 1. Using [29, Theorem 2.1] again,

$$(D.9) \quad \mathbb{P} \left\{ \sum_{i=1}^n (b'w_i)^2 \leq n + 2\sqrt{nt} + 2t \right\} \leq \exp(-t).$$

Following the same argument as in the proof of Lemma 7.3 and choosing $t = 2d + n$, we obtain the desired result. \square

PROOF OF LEMMA A.3. Similarly to the proof of Lemma 7.4, we upper bound $\langle w_i, W_S \rangle$ as

$$\langle w_i, W_S \rangle \leq \langle w_i, W_{G_i} \rangle + \|w_i\|^2.$$

Since w_i and W_{G_i} are independent sub-Gaussian random vectors, we have

$$\mathbb{E}e^{\lambda \langle w_i, W_{G_i} \rangle} \leq \mathbb{E}e^{\frac{\lambda^2 \|W_{G_i}\|^2}{2}} \leq \mathbb{E}e^{\frac{\lambda^2 |S| \|z\|^2}{2}}$$

for $\lambda^2 \leq 1/|S|$, where $z \sim \mathcal{N}(0, I_d)$. Here the last inequality is due to [29, Remark 2.3]. Then D.6) also holds for sub-Gaussian random vectors. Applying [29, Theorem 2.1] on w_i , we have

$$\mathbb{P} \left\{ \|w_i\|^2 \geq d + 2\sqrt{d \log(1/\delta)} + 2 \log(1/\delta) \right\} \leq \delta$$

for $\delta > 0$. Following the same arguments as in the proof of Lemma 7.4, the proof is complete. \square

PROOF OF LEMMA A.4. For any $h \in [k]$, (D.8) implies

$$\mathbb{P} \left\{ \|W_{T_h^*}\|^2 \leq |T_h^*| (d + 4\sqrt{d \log n} + 8 \log n) \right\} \leq n^{-4}.$$

Since $n \geq k$, a union bound argument completes the proof. \square

PROOF OF LEMMA A.5. Let $u_i = \mathbb{I} \{ a \|\theta_h - \theta_g\|^2 \leq \langle w_i, \|\theta_h - \theta_g\| \} \}$. Then $u_i, i \in T_g^*$ are independent Bernoulli random variables. Hoeffding's inequality implies

$$\sum_{i \in T_g^*} u_i \leq \sum_{i \in T_g^*} \mathbb{E}u_i + 2\sqrt{n_g^* \log n}$$

with probability greater than $1 - n^{-4}$. Using Chernoff bound and the sub-Gaussian property of w_i , we have

$$\begin{aligned} \mathbb{E}u_i &\leq \mathbb{E} \exp(-\lambda a \|\theta_h - \theta_g\|^2 + \lambda \langle w_i, \|\theta_h - \theta_g\| \rangle) \\ &\leq \exp(-\lambda a \|\theta_h - \theta_g\|^2 + \lambda^2 \|\theta_h - \theta_g\|^2 / 2) \\ &= \exp(-a^2 \|\theta_h - \theta_g\|^2 / 2), \end{aligned}$$

where we set $\lambda = a$ in the last equality. By union bound and the fact that $n \geq k$, we get the desired result. \square

DEPARTMENT OF STATISTICS
 YALE UNIVERSITY
 NEW HAVEN, CT 06511
 E-MAIL: yu.lu@yale.edu
 E-MAIL: huibin.zhou@yale.edu
 URL: <http://www.stat.yale.edu/~hz68/>