A provable initialization and robust clustering method for general mixture models

Soham Jana, Jianqing Fan, and Sanjeev Kulkarni

Abstract

Clustering is a fundamental tool in statistical machine learning in the presence of heterogeneous data. Most recent results focus primarily on optimal mislabeling guarantees when data are distributed around centroids with sub-Gaussian errors. Yet, the restrictive sub-Gaussian model is often invalid in practice, since various real-world applications exhibit heavy-tail distributions around the centroids or suffer from possible adversarial attacks that call for robust clustering with a robust data-driven initialization. In this paper, we present initialization and subsequent clustering methods that provably guarantee near-optimal mislabeling for general mixture models when the number of clusters and data dimensions are finite. We first introduce a hybrid clustering technique with a novel multivariate trimmed mean type centroid estimate to produce mislabeling guarantees under a weak initialization condition for general error distributions around the centroids. A matching lower bound is derived, up to factors depending on the number of clusters. In addition, our approach also produces similar mislabeling guarantees even in the presence of adversarial outliers. Our results reduce to the sub-Gaussian case in finite dimensions when errors follow sub-Gaussian distributions. To solve the problem thoroughly, we also present novel data-driven robust initialization techniques and show that, with probabilities approaching one, these initial centroid estimates are sufficiently good for the subsequent clustering algorithm to achieve the optimal mislabeling rates. Furthermore, we demonstrate that Lloyd's algorithm is suboptimal for more than two clusters even when errors are Gaussian and for two clusters when error distributions have heavy tails. Both simulated data and real data examples further support our robust initialization procedure and clustering algorithm.

Index Terms

Heavy tail, adversarial outliers, initialization, mislabeling, adaptive methods

I. Introduction

A. Problem

CLUSTERING is an essential task in statistics and machine learning [1], [2] that has diverse practical applications (e.g., wireless networks [3], [4], studying geographical location data [5], grouping biological species [6], [7], medical imaging [8], [9] and defining peer firms in finance [10], [11]). One of the simplest and most studied clustering models is the additive k-centroid setup, where data points are distributed around one of the centroids according to some unknown additive error distribution. Mathematically, this model can be described as

$$Y_i = \theta_{z_i} + w_i, \quad z_1, \ldots, z_n \in [k], \quad \theta_1, \ldots, \theta_k \in \mathbb{R}^d, \tag{1}$$

for a given k, where $[k]$ denotes the set $\{1, \ldots, k\}$. Here $Y_1, \ldots, Y_n$ are the data, θ_g is the centroid corresponding to the g-th cluster, $z = (z_1, \ldots, z_n)$ is the unknown label vector of the data describing which clusters they belong to, and $w_1, \ldots, w_n$ are unknown independent errors. Recent advances in the literature have focused on recovering the labels z. Given any estimate $\widehat{z} = (\widehat{z}_1, \ldots, \widehat{z}_n)$ of z, define the mislabeling error as the proportion of label estimates that do not match the correct ones (up to a permutation of labels):

$$\ell(\widehat{z}, z) = \inf_{\pi \in \mathcal{S}_k} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\pi(z_i) \neq \widehat{z}_i\}} \right], \tag{2}$$

where $\mathcal{S}_k$ is the collection of all mappings from $[k]$ to $[k]$. Then, clustering algorithms are constructed to partition the data into k groups such that the data labels are recovered with a small mislabeling error.

A surge of recent work focuses on constructing provable algorithms to guarantee smaller mislabeling errors as the minimum separation of centroids increases. Let $\|\cdot\|$ denote the Euclidean norm for the rest of the paper, unless mentioned otherwise. The majority of this work studies specific light-tailed models. For example, in the Gaussian mixture model, [12] established that Lloyd's algorithm with spectral initialization can achieve the mislabeling rate $\exp\left(-(1+o(1))\frac{\Delta^2}{8\sigma^2}\right)$, where $\Delta = \min_{g \neq h \in [k]} \|\theta_g - \theta_h\|$ and σ is the common standard deviation of the error coordinates. They also show that the above rate is minimax optimal in finite dimensions. Extending on the above work, [13], [14] show that spectral clustering algorithms can achieve the above rate of error in high dimensions. In the very specific case of $k = 2$, when d is much larger

S. J. is with the Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN, USA, email:soham.jana@nd.edu. J. F. and S. K. are with the Department of Operations Research and Financial Engineering and Department of Electric and Computer Engineering, Princeton University, Princeton, NJ, USA email:jqfan@princeton.edu, kulkarni@princeton.edu. J.F. is partially supported by NSF grants DMS-2210833, DMS-2053832, DMS-2052926 and ONR grant N00014-22-1-2340.

than the sample size n, [15] used a variant of Lloyd's algorithm with spectral initialization to achieve a more precise error rate $\exp\left(-\Theta\left(\frac{\Delta^4/\sigma^4}{\Delta^2/\sigma^2+d/n}\right)\right)$. Recently [16] extended these results to Laplace distributions for the w_i. They showed that a variant of the spectral initialization combined with a maximum likelihood-based estimation strategy can achieve the minimax mislabeling error $\exp\left(-(1+o(1))\left(\frac{\min_{g\neq h\in[k]}\|\theta_g-\theta_h\|_1}{\sigma}\right)\right)$, where $\|\cdot\|_1$ denotes the L_1 norm of vectors and σ is the common standard deviation in each coordinate of the error vector. Their work also addresses the clustering problem for specific types of sub-exponential mixture distributions, for which they used variants of Lloyd's algorithms based on the Bregman divergence. Notably, the mislabeling in the sub-exponential case is significantly different from that in the Gaussian case.

The above exposition illustrates that given any fixed constellation of centroids, the statistical guarantees for mislabeling depend significantly on the properties of the tail of the distributions of the w_i. In particular, given two data-generating setups with the same centroids, the heavy-tailed setup is expected to have a higher mislabeling error. The clustering and initialization methods in the existing literature, such as spectral techniques, are mostly geared to tackle specific light-tailed distributional setups, and it is unknown whether their guarantees extend to general heavy-tailed errors where outliers are prevalent. We address the above issues in the current work. Our proposed algorithm is a significant departure from the existing literature as it is designed to adapt to heavy-tailed error distributions, a scenario that has yet to be extensively explored. In summary, our paper deviates from most of the existing literature in trying to answer the following.

> *Can we construct an initialization method and subsequent clustering algorithm that provably adapt to the decay conditions for general heavy-tailed error distributions and produce (nearly) optimal mislabeling under generic assumptions?*

To address the above question, we primarily focus on generalizing the existing results concerning the tails of the error distributions. As a first work in this direction, we restrict ourselves to the scenario where the actual clusters are of similar sizes (see, e.g., [14] for similar assumptions), and d, k are finite and known (the problem of estimating k is another significant direction in the clustering literature [17], which is beyond the scope of our current work). In particular, we assume the following:

(C1) There is a constant C such that $k, d \leq C$.

(C2) $\alpha = \min_{g\in[k]} \sum_{i=1}^n \mathbf{1}_{\{z_i=g\}}/n \geq c/k$ for some $c > 0$

(C3) (G-decay condition) The error distributions in (1) satisfy
- $\mathbb{P}\left[\|w_i\| > x\right] \leq G(\frac{x}{\sigma}), \sigma > 0, x \geq 0, i \in \{1, \ldots, n\}$,
- $G(\cdot)$ is decreasing on $\mathbb{R}_+ \cup \{0\}$.

(C4) (Parameter space) The parameters $(z, \{\theta_g\}_{g\in[k]})$ is generated from the parameter space $\mathcal{P}_\Delta$

$$\mathcal{P}_\Delta = \left\{(z, \{\theta_g\}_{g\in[k]}) : z \in [k]^n, \min_{g\neq h} \|\theta_g - \theta_h\| \geq \Delta\right\}.$$

Discussion of the assumptions: The constructions of our algorithms do not require the above conditions to be satisfied, and these assumptions are made purely to establish the theoretical guarantees. The condition $d \leq C$ in (C1) is imposed to ensure that the data is concentrated, which is essential to output a consistent clustering. The explicit dependency of k in our algorithms' mislabeling guarantee is provided below in Theorem 1, and it shows that our method are minimax rate optimal when the data has a finite number of clusters. The conditions (C2) and (C4) are standard assumptions in the literature [12], [14]. (C2) ensures that the clusters are balanced, simplifying the analysis, and (C4) provides that the clusters are separable. Finally, we discuss the condition (C3). In some robust problems, such as robust linear regression and robust location estimation, it is traditional to assume constraints on moments of the coordinates of the error variable w_i. In finite-dimensional setups, which is the primary focus of our paper, moment constraints are equivalent to tail constraints. For example, if we consider the case where the coordinates of w_i-s have p-th moment bounded from above by M for some constant $p \geq 1$, then in view of Markov's inequality, it is equivalent to having the tail constraint $G(x) = \frac{c_p d^{p/2} M}{x^p}$, for a constant $c_p > 0$. In particular, having a rapidly decreasing error tail G implies that the data points corresponding to any cluster are less likely to appear closer to the centroids of the other clusters. In any such situation, one should be able to produce a consistent clustering guarantee, and this intuitive argument serves as the inspiration behind our work.

Note that we do not assume much knowledge about the decay function G, making our result general in the true sense. We propose a novel initialization algorithm IOD (Initialization via Ordered Distances, see Algorithm 4 for $k = 2$ and Algorithm 5 for $k \geq 3$) and a novel subsequent iterative clustering technique COD (Clustering via Ordered Distances, provided in Algorithm 2) that are also oblivious to the decay condition G. In addition to the observations $\{Y_1, \ldots, Y_n\}$, the only extra information our algorithms use is knowledge about a lower bound on α. This is often assumed to be known in practice, e.g., the smallest cluster contains at least 5-10% of the data points. Our main result is the following.

Theorem 1. *Suppose that we have data $\{Y_i\}_{i=1}^n$ generated by the additive noise model (1), where the noise w_i-s independently follow the G-decay condition in (C3) for some unknown but fixed G, and d, k are known. Also assume that $\alpha > c/k$ for some constant $c > 0$ and let $\widehat{z}$ be the output label vector when we run the* COD *clustering method with* IOD *initialization scheme*

that knows this c and k. There exist constants $c_k, c_{G,\alpha}, C_{G,\alpha}$ such that the following are satisfied. Whenever $\Delta \geq \sigma C_{G,\alpha}$, we have

$$\sup_{\mathcal{P}_\Delta} \mathbb{E}\left[\ell(\widehat{z}, z)\right] \leq k^2 G\left(\mathsf{SNR} - c_{G,\alpha}\right) + 8ke^{-\frac{n\alpha}{4}}.$$

The entire algorithm runs in $c_k n^2(d + \log n)$ time. In addition, under minor smoothness condition on G, whenever $\alpha \in \left(\frac{c}{k}, \frac{\bar{c}}{k}\right)$ for some constants $0 < c < \bar{c} < 1$, there exists $c_G > 0$ such that

$$\inf_{\widehat{z}} \sup_{\mathcal{P}_\Delta} \mathbb{E}\left[\ell(\widehat{z}, z)\right] \geq \frac{1 - k\alpha - k/n}{12} \cdot G\left(\frac{\Delta}{2\sigma} + c_G\right).$$

Remark 1. The closest work to ours that we could find is [16], which attempts to provide algorithms to achieve similar generalized guarantees. However, the construction of their algorithm hinges on knowing the data-generating model (Algorithm 1 in their paper), and even then, their method works only in the presence of a good initialization. Their work can achieve such initialization via spectral methods for mixture models based on specific parametric families of sub-exponential distributions. Generalizing such results to heavy-tailed models is unknown in the literature.

Remark 2. Our result aims to provide a provable adaptive clustering technique that guarantees consistent clustering (i.e., with vanishing mislabeling) under general decay conditions. The vanishing mislabeling is only achieved in the regime where the signal-to-noise ratio $\mathsf{SNR} = \Delta/2\sigma$ is significant, and we do not explore the territory of small SNR. In particular, our result indicates that when $k \leq C$ for some constant $C > 0$, our algorithm achieves the minimax mislabeling error rate (the additive term $8ke^{-n\alpha/4}$ can be ignored as, for large n such that $8kne^{-n\alpha/4}$ is much smaller than 1, this corresponds to no extra mislabeled points with a high probability). The decay condition (C3) on the errors w_i is presented based on the Euclidean norm to simplify the theoretical results. The above decay condition translates to decay conditions on the distribution of each coordinate of the w_i when the data dimension d is at most a constant. Hence, our results easily extend to the sub-Gaussian and sub-exponential mixture models in finite dimensions. The generalization of such results to high-dimensional setups is left to future work. Our results also translate to obtaining the regime of exact recovery, i.e. $\frac{\Delta}{2\sigma} > c_{G,\alpha} + G^{-1}\left(\frac{1}{k^2 n}\right)$ which corresponds to expected mislabeling dropping below $\frac{1}{n}$. However, this phenomenon is traditionally studied in high-dimensional regimes [15], [18], which is beyond the scope of the current work.

Remark 3 (Dependence on k)**.** Our aim in this work is to generalize existing clustering results in terms of the decay condition of the error distributions and provide provable initialization and clustering algorithms that run in polynomial time for any finite number of clusters. We leave it for later work to optimize the dependence of our results in terms of the number of clusters k. In particular, the current result implies that as long as $k = O(1)$ and n is large, our lower and upper bound differs by a multiple of $k^2 = O(1)$. In terms of the runtime of the algorithm we have $c_k = \left(\frac{O(1)k^2}{\alpha^2}\right)^{k-1}$ which is at most a constant if $k = O(1)$. Notably, this dependency stems from the use of a recursive framework in our initialization algorithms (Algorithm 4 and Algorithm 5), as explained in the proof of Theorem 10, and is unavoidable for recursive techniques. See, e.g., the classical $(1+\epsilon)$-approximated k-means method of [19] which has a runtime of $2^{k^{O(1)}} dn$ for a general k.

Remark 4 (Comparison with existing initialization algorithms)**.** We point out that robust initialization techniques are lacking in the literature that can balance both statistical guarantees and fast runtime, and our paper prioritizes the statistical guarantee part. In particular, initialization schemes that can provably guarantee good outputs for data generated by a general mixture model are almost nonexistent, even for finite k, d. For comparing with existing initialization schemes in the literature, consider the classical work of [19], one of the top choices for initialization algorithms for clustering sub-Gaussian mixtures [13], [14], [20]. Their algorithm uses a similar recursive scheme to ours, although it runs in linear time in n. A significant improvement in our results compared to the above work is that by optimizing our iterative schemes, we can guarantee a good initialization with a probability tending to one as n increases, as mentioned in Theorem 6 and Theorem 8. In contrast, the final theoretical guarantees of the work mentioned above hold with a probability γ^k for a constant γ much smaller than one [19, Theorem 4.1]. This is because, to guarantee a linear runtime, the author first sampled a constant number of data points in their algorithm and then used this sample to find the initialization. Another relevant fast initialization method is the classical k-means++ [21], [22]. The runtime of the above method has a weaker dependence on k. However, k-means++ is known to be highly sensitive to outliers and heavy-tailed errors [23]. The last paper aims to provide a robust modification of k-means++. However, their algorithm's runtime has a similar dependence as ours, also when they try to output exactly k clusters. Their paper also mentions that recently [24] aimed to produce a robust initialization method with much faster improved runtime dependence on k. However, their method requires an initial knowledge of the cost of clustering. It is also popular to use spectral methods for centroid initialization; however, theoretical guarantees for mislabeling minimization tend to exist only in the sub-Gaussian setup. For example, [12] uses the spectral method presented in [25, Claim 3.4], which bounds the centroid estimation error using the Frobenius norm of the error matrix, and then uses a concentration of the sub-Gaussian errors to bound it. Unfortunately, such concentration results fail to work for heavier tails, such as with a polynomial tail. We also provide a general solution to this initialization problem in our work. On a side note, the vanilla version of spectral methods is also vulnerable to noisy setups [26], [27]. This creates issues in the modern setting of adversarial data contamination, which has become prevalent recently.

In addition to the above, we also touch upon the robustness properties of our algorithms to adversarial data contamination. Adversarial data contamination is one of the important challenges of modern machine learning literature, and it often arises due to security concerns. Clustering techniques that are robust to such noises are advantageous in practice and have garnered significant importance in recent trends (see, e.g., [28], [29] for related works). As a counterpart to the above result in Theorem 1, we show that when k is bounded from above with a constant, our initialization algorithm IOD can tolerate a constant fraction of adversarially contaminated data and still produce the desired initialization guarantees as in Theorem 12. In addition, the subsequent iterative algorithm COD can also be improved so that the mislabeling error for the uncorrupted data points can be retained at the minimax optimal level. The result is presented in Theorem 11.

B. Our contributions

Our paper serves as a valuable proof of concept for the universality of nearest-neighbor-based approaches for robust initialization and clustering.

(i) **Initialization:** In the core of our proposed initialization method is a robust improvement on the classical concept of within-cluster sum of squares (WCSS) used in clustering. Given the data set $\{Y_i\}_{i=1}^n$ and centroid estimates $\{\widehat{\theta}_j\}_{j=1}^k$, the WCSS is defined as

$$\text{WCSS}(\{\widehat{\theta}_j\}_{j=1}^k; \{Y_i\}_{i=1}^n) = \sum_{i=1}^n \min_{j\in[k]} \|Y_i - \widehat{\theta}_j\|^2. \tag{3}$$

The primary purpose of the WCSS metric is to assess the quality of the centroid estimates with respect to the data. The classical use of WCSS can be traced to the k-means problem [30], and its other extensions [31] with different metrics. Most recently, the linear time approximate k-means clustering technique of [19] for initialization has come to prominence, as the typical runtime of the exact k-means minimization schedule is $n^{k^2}+1$ [13]. This algorithm is also based on the WCSS metric.

Although promising in the sub-Gaussian noise model, the WCSS is rendered useless in the heavy-tailed setup. This is essentially due to the fact that the WCSS considers the sum based on all the points, and with a high probability, a fraction of points in heavy-tailed setups behave as outliers and significantly inflate the total sum. To deal with the issue, we propose an adaptive quantile-based version of WCSS (see the quantity totdist in Algorithm 4 and Algorithm 5), which computes the sum of distances after removing certain outliers. The strategy ensures that the relevant guarantees adhere to different decay conditions to help produce the aforementioned minimax guarantees. As a consequence of this quantile-based trimming strategy, our method is able to tolerate adversarial points to some extent. As a result of this adaptive strategy, our analysis is significantly involved.

(ii) **Iterative clustering:** Iterative clustering methods broadly follow two major steps:

- *Labeling step:* Given an estimate of the centroids $\widehat{\theta}_h^{(s)}$, construct cluster estimates using the Euclidean distance
- *Estimation step:* For each of the estimated clusters, compute the new centroid estimates $\widehat{\theta}_h^{(s+1)}$ using a suitable estimator.

The classical Lloyd's algorithm [12] uses the sample means of the clusters to update the centroids. However, the sample mean lacks a robustness property. To induce robustness properties in the iterative setup provided above, [29] uses the coordinatewise median to update centroids. However, the coordinatewise median is often too conservative for outlier removal purposes.

The novel centroid updating method we propose in Algorithm 2 centers around the idea of distance-based trimming. To update the centroid of any particular cluster, we first compute the mutual distances of the points and figure out which data point (say P) in the cluster is the most central, relative to a quantile value of the distances. We then compute a trimmed mean of the points by removing a fraction of the cluster's points farthest from the point P. Our analysis points out that even without knowledge of the decay condition (C3), this strategy of trimmed mean adaptively achieves the required mislabeling guarantees in Theorem 1. The algorithm also runs in quadratic time in the sample size n, with most time spent computing the mutual distance.

We present a comparison with relevant robust centroid updating strategies. In the presence of outliers, the centroid estimation guarantees for the coordinatewise median have a dependency on the dimension, which prevents us from attaining the dimension-free guarantees in Theorem 1. Another useful contender might be the geometric median [32], which is also relevant for dealing with data models based on the Euclidean distance. However, the concentration properties of this estimator are lacking in the literature. This makes the analyses challenging, and we leave it for future work. In contrast, the trimmed mean is known to produce sub-Gaussian concentration guarantees, which is the central part of the analysis in this paper. Another well-known robust estimator is Tukey's median, which is established to produce exponential concentration around the true centroid as well [33]. Unfortunately, the estimator requires exponential computation time for multiple dimensions. This defeats our purpose of obtaining a polynomial-time algorithm.

C. Related work

There are limited results in the literature that consider the heavy-tail regime, which is the main focus of our work. Notably, [34] studied the mislabeling minimization problem with different moment constraints. However, their work only aims to produce a mislabeling rate that is a vanishing proportion of the minimum cluster size and does not guarantee optimality or quantify the mislabeling. Robust centroid estimation techniques are also central to our approach. It would be interesting if relevant robust methods in the modern literature, such as mean estimation [35]–[37], vector mean estimation [38]–[40], regression [41], [42] can contribute to improving our results.

A long list of work uses a robust centroid estimation technique in clustering. The classical partitioning around the medoid (PAM) algorithm [43], [44] updates the centroid estimates using a point in the data set (these centroid estimates are referred to as the medoids of the clusters) based on some dissimilarity metric. For example, [44] used the ℓ_1 distance and argued the robustness of the corresponding ℓ_1 based PAM algorithm.

In our paper, we also use the adversarial contamination model. In this model, upon observing the actual data points, powerful adversaries can add new points of their choosing, and our theoretical results depend on the number of outliers added. This contamination model is arguably stronger than the traditional Huber contamination model [45], [46], which assumes that the outliers originate from a fixed distribution via an iid mechanism. Our model is similar to the adversarial contamination model studied in [39], [47] for robust mean estimation. For robust clustering of Gaussian mixtures, [48] examines a similar contamination model. However, these works do not study adversarial outliers in the presence of general heavy-tail error distributions, as is done in this paper.

Another critical related direction is clustering anisotropic mixture models, where the error probabilities decay in particular directions more than others. This differs from our setup, as our decay condition is independent of any direction. Clustering anisotropic mixtures has been studied previously in the sub-Gaussian setup, e.g., in [20] using a variant of Lloyd's algorithm, in [49] for high-dimensional clustering in the specific case of $k = 2$, and [50], [51] with the target of approximating the mixture distribution in Total Variation distance. Specific heavy-tail regimes with non-spherical mixtures are also discussed in [52]; however, they do not characterize the mislabeling in terms of the minimum separation distance. It would be interesting to study whether modifications of our clustering methodology can also work in such setups.

D. Organization

The rest of the paper is organized as follows. In Section II, we re-introduce our mathematical model and present the clustering algorithm we use, given a good initialization. The theoretical results, i.e., the mislabeling rate upper bound under good initialization conditions and the worst case mislabeling lower bound, are presented in Section II-B and Section II-C respectively. These two results jointly characterize the expected mislabeling as a function of the minimum centroid separation. As an application of our results, we study the mislabeling errors for the sub-Gaussian distributions and distributions with moment constraints in Section III. We present our initialization algorithm and their theoretical guarantees in Section IV. The results involving the robustness of our algorithms to adversarial outliers are presented in Section V. In Section VI, we show that Lloyd's algorithm might produce non-converging mislabeling errors even with good initialization. We demonstrate the effectiveness of our algorithms with application on actual and simulated data sets in Section VII. All the proofs and technical details have been provided in the appendix.

II. Robust clustering under mislabeling guarantees under good initialization

A. Algorithm

In this section, we present our iterative clustering algorithm with the assumption that an initial estimate of either the centroids or the label vector is available. As mentioned before in Section I-B, our iterative clustering technique involves a centroid estimation step. For the above purpose, we use a novel multivariate trimmed mean algorithm (TM_δ) based on the ordered distances between all points in the estimated clusters. Algorithm 1 presents the TM_δ estimator for finding the trimmed mean of a dataset $S = \{X_1, \ldots, X_n\}$.

Here is an intuitive explanation of the TM_δ estimator. We first aim to find out a point X_{i^*} from the data set $X = \{X_1, \ldots, X_m\}$ such that the radius of the ball around X_{i^*} that contains $\lceil (1-\delta)m \rceil$ many points in X is the smallest. In other words, X_{i^*} is the point among the data that has the tightest neighborhood of size $(1-\delta)m$ points within the set X. Then, the algorithm computes an average of data points from the tightest neighborhood. Notably, if all the points in X were independent and were generated via a distribution from the class G_σ around a centroid θ, then our analysis based on an argument about the quantiles of G, shows that with a high probability the tightest neighborhood will be contained in a ball around θ, where, the radius of the ball depends on G, σ, δ. Hence, the estimator $\mathsf{TM}_\delta(\{X_1, \ldots, X_m\})$ will be close to θ. When δ is very small, the estimator is approximately the sample mean, which is unbiased for θ.

In the main clustering algorithm, we apply the above estimator on each of the approximated clusters. Consider one of the approximated clusters. In the approximated cluster the data points are not necessarily independent, and there are misclustered points. In such scenarios, we will require the approximated cluster to contain at least half of the points from the corresponding

Algorithm 1 The Trimmed Mean (TM_δ) estimator

Input: Set of points $S = \{X_1, \ldots, X_m\}$, truncation parameter δ

Compute $D = \left\{\{D_{ij}\}_{i,j\in[m]} : D_{i,j} = \|X_i - X_j\|\right\}$
for Each $i \in [m]$ **do**
    Compute R_i as $\lceil(1-\delta)m\rceil$-th smallest number in $\{D_{ij}, j \in [m]\}$
end for
Find $i^* = \operatorname{argmin}_{i\in[m]} R_i$.
Compute an $\lceil(1-\delta)m\rceil$-sized set $V \subseteq [m]$, with a priority to the points closer to X_{i^*}, ties broken arbitrarily

$$V = \{j \in [m] : \|X_j - X_{i^*}\| \le R_{i^*}\}$$

Output: $\mathsf{TM}_\delta(\{X_1, \ldots, X_m\}) = \frac{\sum_{j\in V} X_j}{\lceil(1-\delta)m\rceil}$

true cluster to make the estimation meaningful. Let us assume that the true cluster contains m points and $\lceil m(\frac{1}{2}+c)\rceil$ many points from it belong to the approximated cluster. Then our analysis, using a union bound to deal with the possible dependency issue, shows that the TM_δ algorithm for any $\frac{1}{2} - c < \delta < \frac{1}{2}$ can estimate the centroid of the true cluster, although with a bias that depends on G, σ, δ. Notably, with a large enough Δ, the mislabeling error will be asymptotically unaffected by this bias.

In view of the above centroid estimation algorithm, we present our primary clustering technique, the Clustering via Ordered Distances (COD_δ), below in Algorithm 2. Our algorithm requires that an initial estimate of the centroids or label vector to kick off the clustering process. With a lousy initialization, the method can converge to local optima. This is similar to most off-the-shelf methods like the k-means algorithm, Fuzzy C-means algorithm, EM algorithms, etc. [53, Section 3]. We will provide a novel robust centroid initialization technique later in Section IV that will guarantee a global convergence for our COD_δ algorithm.

Algorithm 2 The Clustering via Ordered Distances (COD_δ) - algorithm

Input: Data $\{Y_1, \ldots, Y_n\}$. Initial centroid estimates $(\widehat{\theta}_1^{(0)}, \ldots, \widehat{\theta}_k^{(0)})$ (or initial label estimates $\left\{\widehat{z}_i^{(0)}\right\}_{i=1}^n$). Maximum number of iterations M. Error threshold ϵ. Truncation level $\delta \in (0, \frac{1}{2})$.

Set s = 1
for $h \in \{1, 2, \ldots, k\}$ **do**
    Labeling step:
    if $s = 1$ and initial estimate of the label vector is available **then**
        Compute $T_h^{(s)} = \{i \in \{1, \ldots, n\} : \widehat{z}_i^{(0)} = h\}$
    else
        Compute the clusters, with ties broken arbitrarily,

$$T_h^{(s)} = \Bigg\{ i \in \{1, \ldots, n\} : \|Y_i - \widehat{\theta}_h^{(s-1)}\| \le \|Y_i - \widehat{\theta}_a^{(s-1)}\|, \; a \in [k], a \neq h \Bigg\},$$

    end if
    Estimation step: Update the new estimate of θ_h as $\widehat{\theta}_h^{(s)} = \mathsf{TM}_\delta(\{Y_j : j \in T_h^{(s)}\})$.
end for
if $s = 1$ or $\{2 \le s < \mathsf{M}$ and $\frac{1}{k}\sum_{h=1}^k \|\widehat{\theta}_h^{(s)} - \widehat{\theta}_h^{(s-1)}\|^2 > \epsilon\}$ **then**
    Update $s \leftarrow s + 1$ and go back to the **Labeling step** and repeat
end if

Output: $(\widehat{\theta}_1^{(s)}, \ldots, \widehat{\theta}_k^{(s)})$ and $\widehat{z}_i^{(s)} = \operatorname{argmin}_{h\in[k]} \|Y_i - \widehat{\theta}_h^{(s)}\|$.

B. Mixture model and mislabeling guarantees

In this section we present the mislabeling upper bound achieved by the COD_δ algorithm when a reasonable initialization is present. Our result is presented in terms of the decay condition G. To this end, we restate our full data generating model.

Fix a monotonically decreasing function G with $\lim_{x\to\infty} G(x) = 0$. We say that a random variable w is distributed according to a G-decay condition with a scale parameter σ, denoted by $w \in G_\sigma$, if w satisfies the condition (C3). We observe independent samples $Y_1, \ldots, Y_n \in \mathbb{R}^d$ from a mixture of k many G_σ distributions as follows:

$$Y_i = \theta_{z_i} + w_i, i \in [n], w_i \in G_\sigma, z_i \in [k], \theta_h \in \mathbb{R}^d, h \in [k], \tag{4}$$

where $z = \{z_i\}_{i=1}^n \in [k]^n$ denotes the underlying true label vector, $\theta_1, \ldots, \theta_k$ are the unknown centroids. We study the mislabeling loss function between the estimated label vector $\widehat{z} = \{\widehat{z}_i\}_{i=1}^n$ and true label vector $z = \{z_i\}_{i=1}^n$ given by (2). To better present our results, we first introduce some notations. For all $h, g \in [k]$, define

$$\begin{aligned} T_h^* = \{i \in [n] : z_i = h\}, T_h^{(s)} = \left\{i \in [n] : z_i^{(s-1)} = h\right\} \\ n_h^* = |T_h^*|, n_h^{(s)} = \left|T_h^{(s)}\right|, n_{hg}^{(s)} = \left|T_h^* \cap T_g^{(s)}\right| \end{aligned} \tag{5}$$

Note that for $s \geq 1$ this implies

$$T_h^{(s)} = \left\{i \in [n] : \|Y_i - \widehat{\theta}_h^{(s-1)}\| \leq \|Y_i - \widehat{\theta}_a^{(s-1)}\|, a \in [k]\right\}. \tag{6}$$

with ties broken arbitrarily. Recall the minimum fraction of points in the data set that come from a single component defined previously in (C2)

$$\alpha = \min_{g\in[k]} \frac{n_g^*}{n}. \tag{7}$$

Define the cluster-wise correct labeling proportion at step s as

$$H_s = \min_{g\in[k]} \left\{ \min \left\{ \frac{n_{gg}^{(s)}}{n_g^*}, \frac{n_{gg}^{(s)}}{n_g^{(s)}} \right\} \right\}. \tag{8}$$

We denote by $\Delta = \min_{g\neq h\in[k]} \|\theta_g - \theta_h\|$ the minimum separation between the centroids. Let

$$\Lambda_s = \max_{h\in[k]} \frac{1}{\Delta} \|\widehat{\theta}_h^{(s)} - \theta_h\|. \tag{9}$$

be the error rate of estimating the centroids at iteration s. Our results are presented based on the signal-to-noise ratio in the model, defined as

$$\mathsf{SNR} = \frac{\Delta}{2\sigma}.$$

We have the following result.

Theorem 2. *There exists a constant $c_0 > 0$ such that the following holds.*

- *If we have an initial estimate of the label vector satisfying $H_0 \geq \frac{1}{2} + \gamma$ for a $\gamma \in (\frac{10}{n\alpha}, \frac{1}{2})$, then whenever $\mathsf{SNR} \geq G^{-1}\left(\exp\left\{-\frac{c_0}{\alpha\gamma}\right\}\right)$, the output $\widehat{z}^{(s)}$ of the COD_δ algorithm at iteration s with $\delta = \frac{1}{2} - \frac{\gamma}{4}$ achieves the following expected mislabeling rate for all $s \geq 2$*

$$\mathbb{E}\left[\ell(\widehat{z}^{(s)}, z)\right] \leq k^2 G\left(\mathsf{SNR} - \exp\left\{-\frac{c_0}{\alpha}\right\}\right) + 8ke^{-\frac{n\alpha}{4}}.$$

- *Instead of initial labels, if we have initial centroid estimates that satisfy*

$$\Lambda_0 \leq \frac{1}{2} - \frac{G^{-1}\left(\exp\left\{-\frac{c_0}{\alpha}\right\}\right)}{\mathsf{SNR}},$$

then the last conclusion holds with $\gamma = 0.3$.

Remark 5. The above result shows that our algorithm reaches the desired level of mislabeling accuracy after only a couple of steps. This is essentially due to the robustness of the underlying centroid estimation algorithm, which is observed similarly in [29]. As we increase the number of iterations, we expect to improve the mislabeling error. However, the corresponding theoretical analysis is beyond the scope of the current paper. For all practical purposes, we iterate the clustering process a pre-specified large number of times to obtain the final label estimates.

C. Optimality of mislabeling: lower bound

In this section, we show that when SNR is significantly large, even if we have a good centroid initialization, the mislabeling error can be as high as $\Theta(G(\mathsf{SNR}))$, up to factors depending on k. Suppose that $\theta_1^*, \ldots, \theta_k^*$ are the true centroids, that are known to us, with $\min_{h \neq g} \|\theta_h^* - \theta_g^*\| = \|\theta_1^* - \theta_2^*\| = \Delta$. Consider the following set of parameters and label vectors

$$\mathcal{P}_0 = \Bigg\{ \big(z, \{\theta_i\}_{i=1}^k\big) : \theta_i = \theta_i^*, i \in [k], \\ |\{i \in [n] : z_i = g\}| \geq n\alpha, g \in [k] \Bigg\}$$

In addition, we assume that the decay function G satisfies the following smoothness condition:

(Q) There exists $c_G > 0$ such that $G(\cdot)$ is differentiable in the interval (c_G, ∞) and $|G'(y)|\,|_{y \geq c_G}$ is monotonically decreasing.

Then we have the following guarantee. The proof is provided in Appendix C.

Theorem 3. *Suppose that $\alpha \in (c/k, \bar{c}/k)$ for some constants $0 < c < \bar{c} < 1$. Then given any decay function G satisfying (Q), there exists $C_G > 0$ such that*

$$\inf_{\widehat{z}} \sup_{\mathcal{P}_0} \mathbb{E}\left[\ell(\widehat{z}, z)\right] \geq \frac{1 - k\alpha - k/n}{12} \cdot G(\mathsf{SNR} + C_G), \;\; \Delta \geq \sigma C_G.$$

III. Applications to specific distributions

In this section, we showcase our general results for two specific mixture models with error distributions having sub-Gaussian tails and polynomial tails.

A. Sub-Gaussian mixture model

In this model, the observed data $Y_1, \ldots, Y_n \in \mathbb{R}^d$ are distributed as

$$Y_i = \theta_{z_i} + w_i, \;\; i = 1, \ldots, n, \tag{10}$$

where $\{z_i\}_{i=1}^n \in [k]^n$ denotes the underlying unknown label vector of the points, and $\{w_i\}_{i=1}^n$ denote the error variables distributed independently as zero mean sub-Gaussian vectors with parameter $\sigma > 0$ (denoted by $w_i \in \mathsf{SubG}(\sigma)$), i.e.,

$$\mathbb{E}\left[e^{\langle a, w_i \rangle}\right] \leq e^{\frac{\sigma^2 \|a\|^2}{2}}, \;\; \text{for all } i \in \{1, \ldots, n\} \text{ and } a \in \mathbb{R}^d. \tag{11}$$

In order to apply our main results to the sub-Gaussian clustering problem, we need to derive a decay condition similar to G_σ. To this end, we note the next result from Remark 2.2 of [54]: given any $t > 0$ and $w \in \mathsf{SubG}(\sigma)$ we have

$$\mathbb{P}\left[\|w\|^2 > \sigma^2 \cdot (d + 2\sqrt{dt} + 2t)\right] \leq e^{-t}. \tag{12}$$

Simplifying the above, we get for all $x \geq \sqrt{d}$

$$\mathbb{P}\left[\|w\| > \sigma \cdot x\right] \leq \exp\left(-(\sqrt{x^2 - d/2} - \sqrt{d/2})^2/2\right). \tag{13}$$

Hence we apply Theorem 2 with

$$G(x) = \exp\left(-(\sqrt{x^2 - d/2} - \sqrt{d/2})^2/2\right), x \geq \sqrt{d}, \\ G^{-1}(y) = \left(2\log(1/y) + 2\sqrt{d\log(1/y)} + d\right)^{1/2} \\ \leq \sqrt{2\log(1/y)} + \sqrt{d}.$$

In view of the above, the next result directly follows.

Corollary 4. *There are absolute constants c_0 and c_1 such that the following holds true. Fix $\gamma \in (\frac{10}{n\alpha}, \frac{1}{2})$ and suppose that the clustering initialization satisfies either one of the following conditions*

$$H_0 \geq \frac{1}{2} + \gamma \quad \text{or} \quad \Lambda_0 \leq \frac{1}{2} - \frac{c_1(d + 1/\alpha)^{1/2}}{\mathsf{SNR}}.$$

Then, for sub-Gaussian w_i, whenever $\mathsf{SNR} \geq c_0(d + 1/(\alpha\gamma))^{1/2}$, the COD_δ algorithm with $\delta = \frac{1}{2} - \frac{\gamma}{4}$ achieves the mislabeling rate for all $s \geq 2$

$$\mathbb{E}\left[\ell(\widehat{z}^{(s)}, z)\right] \leq \exp\left\{-\frac{1}{2}\left(\mathsf{SNR} - c_1^2(d + 1/\alpha)^{1/2}\right)^2 - 2\log k\right\}.$$

Remark 6. The implications of the above result are the following: whenever SNR is significantly larger than $(d+1/(\alpha\gamma))^{1/2}$ and $\log k$, the mislabeling rate in the sub-Gausssian mixture model is approximately $\exp\left(-\Delta^2/(8\sigma^2)\right)$. This matches the theoretical limit for mislabeling proportion in the sub-Gaussian mixture model with a constant d; see [12] for an example which demonstrates that Lloyd's algorithm achieves a similar error rate. When d is fixed, the initialization conditions stated above in Corollary 4 are weaker than the conditions required for Lloyd's algorithm. In particular, the initialization condition on H_0 for Lloyd's algorithm depends on the relative distance between the closest cluster centroid and the farthest cluster centroids, given by $\lambda = \max_{h\neq g\in[k]} \|\theta_g - \theta_h\|/\Delta$. As the value of λ increases, Lloyd's algorithm requires a stronger initialization condition to guarantee the optimal mislabeling. Notably, this dependency of initialization condition on λ is necessary for Lloyd's algorithm to converge as the mean based centroid estimate for any cluster can be destabilized via contamination from the farthest away clusters. We believe that the dependency on d in the condition involving SNR can be further improved by first running a spectral method on the dataset and then applying the COD_δ algorithm. However, the analysis is beyond the scope of the current paper.

B. Mixture models with moment constraints on the norm

In this section, we explore the clustering guarantees of COD_δ when the data generating distributions have moment constraints. We say that a random variable w is distributed according to a p-th moment constraint on the norm with a scale parameter σ, denoted by $w \in \mathcal{R}_p(\sigma)$ for a given $p > 0$, if it satisfies the following condition:

(P) There exists $x_0 > 0$ such that $\mathbb{P}\left[\|w\| > x\right] < \frac{\sigma^p}{x^p}$ for all $x \geq x_0$. Without a loss of generality we will assume $x_0 \geq \sigma$ as otherwise the bound is trivial.

We observe independent samples $Y_1, \ldots, Y_n \in \mathbb{R}^d$ from a mixture of k many $\mathcal{R}_p(\sigma)$ distributions

$$Y_i = \theta_{z_i} + w_i,\ i \in [n], \quad w_i \in \mathcal{R}_p(\sigma),$$
$$z_i \in \{1, 2, \ldots, k\}, \theta_h \in \mathbb{R}^d, h \in [k],$$

where $z = \{z_i\}_{i=1}^n \in [k]^n$ denote the underlying label vector. The mislabeling proportion for the estimated label vector $\widehat{z}$ produced by the COD_δ algorithm is summarized as follows.

Theorem 5. *Suppose that $\gamma \in (\frac{10}{n\alpha}, \frac{1}{2})$. Then there exist absolute constants $c_1, c_2 > 0$ such that the following holds. If the clustering initialization satisfies*

$$H_0 \geq \frac{1}{2} + \gamma \quad \textit{or} \quad \Lambda_0 \leq \frac{1}{2} - \frac{e^{c_1/p\alpha}}{\mathsf{SNR}},$$

then whenever $\mathsf{SNR} \geq e^{c_2/p\alpha\gamma}$ we have that the COD_δ algorithm with $\delta = \frac{1}{2} - \frac{\gamma}{4}$ achieves the expected mislabeling rate

$$\mathbb{E}\left[\ell(\widehat{z}^{(s)}, z)\right] \leq k^2(\mathsf{SNR} - e^{2c_1/p\alpha})^{-p} + 8ke^{-\frac{n\alpha}{4}}, \quad s \geq 2.$$

In addition, this rate is optimal, up to a factor depending on k, α.

Notably, in the above result, we never assume that the error distributions are centered around zero. As long as there is sufficient decay around the location parameter θ_h, our result states that we should be able to produce good clustering guarantees. Note that the second term is usually negligible.

IV. Provable initialization methods

In this section, we propose centroid initialization algorithms which guarantee that the conditions on Λ_0 required in Theorem 2 are met with a high probability. We deal with the cases $k = 2$ and $k \geq 3$ separately. The algorithm in the case of $k \geq 3$ follows from a recursive structure which calls the algorithm for $k = 2$ at the end.

A. Two centroids

We first present our initialization algorithm for the setup with $k = 2$. Our algorithm revolves around searching for data points with dense neighborhoods. With a high signal-to-noise ratio, such dense neighborhoods are expected to be close to the centroids. Hence, the data points with a high density neighborhoods can be chosen as good estimates of the true centroids. Our algorithm for finding such data points is presented in Algorithm 3: Given a data set with size n and neighborhood size parameter q, the algorithm outputs a data point with the tightest neighborhood in the data set with at least nq points from the set.

Algorithm 3 The High Density Point (HDP_q) - algorithm

Input: Set of points $S = \{Y_1, \ldots, Y_n\}$, neighborhood size parameter q

Create distance matrix $D = \{D_{ij} : i, j \in [n], D_{i,j} = \|Y_i - Y_j\|\}$
for Each $i \in [n]$ **do**
    Compute the R_i as the $\lceil nq \rceil$-th smallest number in $\{D_{ij}, j \in [n]\}$
end for
Find $i^* = \operatorname{argmin}_{i \in [n]} R_i$.

Output: Y_i^*

In view of the above, we present the initialization algorithm for $k = 2$ below.

Algorithm 4 The Initialization via Ordered Distances ($\mathsf{IOD}_{2,m_1,m,\beta}$)- algorithm with 2 centroids

Input: Data $Y_1, \ldots, Y_n$, truncation parameter β, batch size m, and initial cluster size m_1

Compute $\mu_1^{(1)} = \mathsf{HDP}_{\frac{m_1}{n}}(\{Y_1, \ldots, Y_n\})$.
Order the rest of the points in increasing Euclidean distance from $\mu_1^{(1)}$.
Denote the first m_1 points in the list as $\mathcal{P}_1^{(1)}$ and the rest of the points list as $\overline{\mathcal{P}_1^{(1)}}$ in increasing order of distance form $\mu_1^{(1)}$.
Compute $\mathsf{dist}_1^{(1)}$ as the $(1-\beta)m_1$-th smallest value among the distances from $\mu_1^{(1)}$ to $\mathcal{P}_1^{(1)}$.
for $\ell = 1, \ldots, \left\lceil \frac{n-m_1}{m} \right\rceil$ **do**
    Assign $\mu_1^{(\ell)} = \mu_1^{(1)}$. Compute $\mathsf{dist}_1^{(\ell)}$ as the $(1-\beta)\left|\mathcal{P}_1^{(\ell)}\right|$-th smallest value among the distances from $\mu_1^{(1)}$ to $\mathcal{P}_1^{(\ell)}$
    Compute $\mu_2^{(\ell)} = \mathsf{HDP}_{1-\beta}(\overline{\mathcal{P}_1^{(\ell)}})$.
    Compute $\mathsf{dist}_2^{(\ell)}$ as the $(1-\beta)\left|\overline{\mathcal{P}_1^{(\ell)}}\right|$-th smallest value among the distances from $\mu_2^{(\ell)}$ in the set $\overline{\mathcal{P}_1^{(\ell)}}$.
    Store $\mathsf{totdist}^{(\ell)} = \mathsf{dist}_1^{(\ell)} + \mathsf{dist}_2^{(\ell)}$.
    Move the first m points in the list $\overline{\mathcal{P}_1^{(\ell)}}$ to $\mathcal{P}_1^{(\ell)}$ to construct $\overline{\mathcal{P}_1^{(\ell+1)}}, \mathcal{P}_1^{(\ell+1)}$
end for
Find $(\mu_1^*, \mu_2^*) = (\mu_1^{(\ell^*)}, \mu_2^{(\ell^*)})$ and $\mathsf{totdist}^* = \mathsf{totdist}^{(\ell^*)}$ corresponding to

$$\ell^* = \operatorname*{argmin}_{\ell \in \{1, \ldots, \left\lceil \frac{n-m_1}{m} \right\rceil - 1\}} \mathsf{totdist}^{(\ell)}.$$

Output: (μ_1^*, μ_2^*) and $\mathsf{totdist}^*$.

Remark 7 (Explanation of Algorithm 4)**.** We apply Algorithm 4 to the full data set $\mathcal{S} = \{Y_1, \ldots, Y_n\}$ with the choice $m_1 = \frac{n\alpha}{4}$. The first centroid estimate $\widehat{\theta}_1$ is chosen by picking the index $i^* \in \{1, \ldots, n\}$ such that the tightest neighborhood in $\mathcal{S}$ of size m_1 around Y_{i^*} has the smallest radius compared to any other such neighborhood around any other point Y_i. Using results on concentration of the quantiles of G Lemma 22, it is not too difficult to show that, for some constant $C = \sigma \widetilde{C}_{G,\alpha}$ depending on the decay function G and minimum cluster proportion α, with a high probability

$$\widehat{\theta}_1 = Y_{i^*} \in \cup_{i=1,2} \mathcal{B}(\theta_i, C),$$

where $\mathcal{B}(x, R)$ denotes the Euclidean ball of radius R around the point x. Without a loss of generality, suppose that $\mathcal{B}(\theta_1, C)$ is the set containing $\widehat{\theta}_1$. Denote the first m_1 points in the data set closest to $\widehat{\theta}_1$ as $\mathcal{P}_1$ and denote the complement set of points as $\overline{\mathcal{P}_1}$.

In view of the above, it is clear that the inherent challenge in finding a good initialization lies in obtaining a good estimate of θ_2. At this stage, it might seem reasonable to apply the HDP algorithm again on the remaining set of points $\overline{\mathcal{P}_1}$ to estimate θ_2. Unfortunately, a direct application of the HDP on the set $\overline{\mathcal{P}_1}$ need not guarantee a good estimate of θ_2. This is because as there are significantly many points in $\overline{\mathcal{P}_1} \cap \{Y_i : i \in T_1^*\}$ (at least $n_1^* - \frac{n\alpha}{4} \geq \frac{3n\alpha}{4}$ points) and the data point chosen via HDP can indeed belong to $\{Y_i : i \in T_1^*\}$, which will be closer to θ_1 than θ_2 with a high probability. To remedy this issue, we gradually move m points from $\overline{\mathcal{P}_1}$ to $\mathcal{P}_1$, prioritizing the points in $\overline{\mathcal{P}_1}$ that are closer to $\widehat{\theta}_1$. At each transfer step, we can compute the corresponding centroid estimate $\widehat{\theta}_2$, using HDP estimator, while keeping $\widehat{\theta}_1$ as it is. To control the stopping point at which we terminate the transfer of points from $\overline{\mathcal{P}_1}$ to $\mathcal{P}_1$ we use the quantile-based measure totdist. The reason behind using the quantiles of the intra-cluster distances rather than their sum, which is often used in the WCSS metric defined in (3), is that the quantiles are more robust to outlying observations. Notably, once we transfer a significant number of points

from $\overline{\mathcal{P}_1}$ that belong to T_1^* and keep a substantial number of points in $\overline{\mathcal{P}_1}$ that belong to T_2^*, a second application of the HDP algorithm will guarantee a good estimate $\widehat{\theta}_2$ of θ_2

The following result describes our choice for the parameters m_1, m, β in the above algorithm and the corresponding centroid estimation guarantees.

Theorem 6. *Suppose that out of the n many observed data points, n_i^* many are from cluster $T_i^*, i = 1, 2$ and assume that for some constant $\alpha > 0$ the counts satisfy $n_1^*, n_2^* > n\alpha$. Then there are constants $c_1, c_2 > 0$ such that if $\Delta \geq c_1 \sigma G^{-1}\left(e^{-\frac{c_2}{\alpha^2}}\right)$ then the* IOD $_{2,m_1,m,\beta}$ *algorithm with $m_1 = \lceil \frac{n\alpha}{4} \rceil, m = \max\{1, \left\lfloor \frac{n\alpha^2}{16} \right\rfloor\}, \beta = \frac{\alpha}{4}$ guarantees, for a permutation π on $\{1, 2\}$*

$$\max_{i=1,2} \|\theta_{\pi(i)} - \mu_i^*\| \leq \Delta/3$$

with probability at least $1 - 4e^{-n\alpha/4}$.

Remark 8. Our main result Theorem 2 states that for a large enough SNR, the mislabeling guarantee for the COD algorithm holds for any initial centroid estimates that satisfy $\Lambda_0 \leq \frac{1}{2+c}$ for some constant $c > 0$. In other words, given centroid estimates μ_1^*, μ_2^* of θ_1, θ_2, it is sufficient to satisfy

$$\max_{i=1,2} \|\theta_{\pi(i)} - \mu_i^*\| = \Delta\Lambda_0 \leq \frac{\Delta}{2+c}, \tag{14}$$

for some $c > 0$. In view of Theorem 6, our proposed initialization paired with the proposed COD algorithm leads to the desired mislabeling.

The following result resolves the time complexity to run Algorithm 4.

Theorem 7. *The runtime of* IOD $_{2,m_1,m,\beta}$ *is at most $O\left(\frac{1}{\beta^2}\left(n^2 d + n^2 \log n\right)\right)$.*

Proof. We first find the point in the data set with the tightest neighborhood of m_1 other points and the corresponding $(1-\beta)$-quantile of the distances from it. We have the following observations.

- Computing the tightest neighborhood of m_1 involves computing all the pairwise distances, which has a time complexity of $n^2 d$.
- Computing the m_1/n-quantile of the distances for all the points has a time complexity $O(n^2 \log n)$, and then subsequently computing the minimum of those values takes at most $O(\log n)$ time.

Once we have found the first centroid, for each $1 \leq \ell \leq 2/\beta^2$ we construct $\mathcal{P}_1^{(\ell)}, \overline{\mathcal{P}_1^{(\ell)}}$ according to distances from the first centroid. For each values of ℓ, we repeat the following.

- We find the $(1-\beta)$ quantiles of the distances from $\mu_1^{(1)}$ in $\mathcal{P}_1^{(\ell)}$, which takes $n \log n$ time.
- We then find the point in the data set with the tightest neighborhood of $(1-\beta)\left|\overline{\mathcal{P}_1^{(\ell)}}\right|$ other points in $\overline{\mathcal{P}_1^{(\ell)}}$ and the corresponding $(1-\beta)$-quantile of the distances from it. This will again take at most $O(n^2 d + n^2 \log n)$ time similar to above considerations.

Combining the above, we get that the total runtime is $\frac{O(1)}{\beta^2}\left(n^2 d + n^2 \log n\right)$. □

B. Algorithm with a general k

To extend the above algorithm for a general cluster number k we use a recursive framework that utilizes the structure of Algorithm 4. We first locate a point from the data set that has the tightest neighborhood of size m_1 (denote it by $\mathcal{P}$). This will serve as the first centroid estimate. Then for the remaining point set (call it $\overline{\mathcal{P}}$) we recursively apply the initialization algorithm to find the best $k-1$ cluster centers. We repeat the process of finding the best $k-1$ cluster centroids from $\overline{\mathcal{P}}$ after successively removing m points from $\overline{\mathcal{P}}$ and adding it to $\mathcal{P}$. In each step, say ℓ, we compute an appropriate distance measure similar to $\mathsf{totdist}^{(\ell)} = \mathsf{dist}_1^{(\ell)} + \mathsf{dist}_2^{(\ell)}$ in Algorithm 1, that quantifies the goodness of the clustering at that step. Finally, the centroids generated in the step with the lowest distance measure chosen to be the final output. Whenever we are left with the task of finding two centroids from $\overline{\mathcal{P}}$, we resort to $\mathsf{IOD}_{2,m_1,m,\beta}$. The details are provided in Algorithm 5.

The following result describes a choice of the parameters m_1, m, β that guarantees a good initialization, sufficient to meet the requirements related to Λ_0 in Theorem 2. Hence, our initialization algorithm, paired with the clustering technique COD, produces the desired mislabeling starting from scratch.

Theorem 8. *Suppose that out of the n many observed data points, n_i^* many are from cluster $T_i^*, i = 1, \ldots, k$ and for some constant $\alpha > 0$ the counts satisfy $n_i^* > n\alpha, i = 1, \ldots, k$. Then there are constants c_1, c_2 such that the following is satisfied.*

Algorithm 5 The Initialization via Ordered Distances ($\mathsf{IOD}_{k,m_1,m,\beta}$)- algorithm

Input: Data $\{Y_1,\dots,Y_n\}$, k clusters to be found, truncation parameter β, batch size m, initial cluster size m_1.

Output: Centroid estimates $\{\mu_i^*\}_{i=1}^k$ and Error measure $\mathsf{totdist}_k^*$.

if $k \geq 3$ **then**
  Compute $\mu_k^{(k,1)} = \mathsf{HDP}_{\frac{m_1}{n}}(\{Y_1,\dots,Y_n\})$
  Denote the first m_1 points closest to $\mu_k^{(k,1)}$ as $\mathcal{P}_k^{(1)}$ and the rest of the points as $\overline{\mathcal{P}_k^{(1)}}$.
  for $\ell_k = 1,\dots,\lfloor \frac{n-m_1}{m} \rfloor$ **do**
    Set $\mu_k^{(k,\ell_k)} = \mu_k^{(k,1)}$ and compute $\mathsf{dist}_k^{(\ell_k)}$ as the the distance to the $(1-\beta)|\mathcal{P}_k^{(\ell_k)}|$-th closest point from $\mu_k^{(k,\ell_k)}$ in $\mathcal{P}_k^{(\ell_k)}$.
    Run the $\mathsf{IOD}_{k-1,m_1,m,\beta}$ algorithm on the set $\overline{\mathcal{P}_k^{(\ell_k)}}$ and note the outputs:

$$\text{centroid set } \{\mu_i^{(k,\ell_k)}\}_{i=1}^{k-1} \text{ and error measure as } \mathsf{totdist}_{k-1}^{(\ell_k)}.$$

    Store $\mathsf{totdist}_k^{(\ell_k)} = \mathsf{dist}_k^{(\ell_k)} + \mathsf{totdist}_{k-1}^{(\ell_k)}$.
    Move the first m points in $\overline{\mathcal{P}_k^{(\ell_k)}}$, that are closer to $\mu_k^{(k,1)}$, to $\mathcal{P}_k^{(\ell_k)}$ to construct $\overline{\mathcal{P}_k^{(\ell_k+1)}}, \mathcal{P}_k^{(\ell_k+1)}$.
  end for
  $\ell_k^* = \mathrm{argmin}_{\ell_k} \mathsf{totdist}_k^{(\ell_k)}$, $\{\mu_i^*\}_{i=1}^k = \{\mu_i^{(k,\ell_k^*)}\}_{i=1}^k$, $\mathsf{totdist}_k = \mathsf{totdist}_k^{(\ell_k^*)}$.
else if k=2 **then**
  Run the $\mathsf{IOD}_{2,m_1,m,\beta}$ algorithm and note the output as $\{\mu_1^*,\mu_2^*\}$ and $\mathsf{totdist}_k^*$.
end if

Whenever $\Delta > c_1 k\sigma G^{-1}\left(e^{-c_2/\beta^2}\right)$, *there is a permutation* π *of the set* $[k]$ *that satisfies* $\max_{i\in[k]} \|\theta_{\pi(i)} - \mu_i^*\| \leq \Delta/3$ *with probability at least* $1 - 2ke^{-n\alpha/4}$, *where the* $\{\mu_i^*\}_{i=1}^k$ *are centroids generated via the* $\mathsf{IOD}_{k,m_1,m,\beta}$ *algorithm with*

$$m_1 = \left\lceil \frac{n\alpha}{4} \right\rceil, m = \max\left\{1, \left\lfloor \frac{n\beta^2}{2} \right\rfloor\right\}, \beta = \frac{\alpha}{4k}.$$

Remark 9. The task of accurately estimating a lower bound to the minimum cluster proportion α is a challenging problem and this is not the primary focus of our work. In addition, apart from improving the runtime of our algorithm, obtaining a good lower bound on α is mainly required for technical reasons, such as proving the theoretical guarantees of our algorithm. In practice, our algorithm usually produces correct clustering output even when we replace α by a slightly larger value. See, for example, supporting discussion and simulation results in Remark 14 and Fig. 1 where our algorithm works when replace alpha by larger quantities, and our method works even in presence of outliers. Nonetheless, below we propose a modification to our algorithm, to tackle the scenarios of unknown α, and achieve good clustering while circumventing the issue of estimating α. We recommend that one initially clusters the data with a guess of α given by $\widehat{\alpha}^{(0)} = \frac{1}{k}$, which is the largest possible value for α for k clusters. Next we check if updating $\widehat{\alpha}^{(t+1)} = \widehat{\alpha}^{(t)}/2$ significantly changes the clustering output for $t = 0, 1, \dots$. To compute the similarity of clustering outputs $\{\widehat{z}_i^{(t)}\}_{i=1}^n$ and $\{\widehat{z}_i^{(t+1)}\}_{i=1}^n$ at subsequent iterations $t, t+1$, we can use the minimum label alignment error among all possible permutations of labels, given by $\xi_t = \ell(\widehat{z}^{(t+1)}, \widehat{z}^{(t)})$, where ℓ is as defined in (2). We keep updating $\widehat{\alpha}^{(t)}$ until ξ_t falls below some prespecified threshold $\xi > 0$ and use $\widehat{z}^{(t)}$ as our final output. The intuition behind the above modification is that as soon as $\widehat{\alpha}^{(t)}$ becomes smaller than α, our method should be able to produce a vanishing mislabeling. On the other hand, if $\widehat{\alpha}^{(t)}$ leads to an incorrect clustering, it is unlikely that it would exactly match the output of clustering algorithm which relies on $\widehat{\alpha}^{(t+1)}$. We leave it for future work to explore the theoretical guarantees of the above heuristics.e above heuristics.

In view of Theorem 8, our initialization paired with the COD algorithm leads to the desired mislabeling. Notably, Lloyd's algorithm [12] and the hybrid k-median algorithm in [29] also required the initialization condition $\Lambda_0 < 1/(2+c)$, for a constant $c > 0$, to produce the optimal mislabeling rate in the sub-Gaussian clustering problem. In view of Section III, and the proof of Theorem 8 in Appendix F, we note that Theorem 8 will require $\Delta \geq \sigma c_\alpha \sqrt{d}$ for some constant c_α depending on α. This implies the following.

Corollary 9. *There is a constant* c_α *depending on* α *such that the following holds true. When initialized with our initialization scheme* IOD *, the hybrid* k*-median algorithm in [29] and Lloyd's algorithm [12] produce the optimal mislabeling rate in the sub-Gaussian error setup, provided* $\Delta > \sigma c_\alpha \sqrt{d}$ *and* d *is bounded by a constant.*

The following result resolves the time complexity to run Algorithm 5.

Theorem 10. *The runtime of* $\mathsf{IOD}_{k,m_1,m,\beta}$ *is at most* $(O(1)/\beta^2)^{k-1}(n^2 d + n^2 \log n)$.

Proof. As our method is a recursive process, we construct a recursion that relates the computation time of finding the best k centroids to that of finding $k-1$ best centroids. In the recursion process, when we want to find out the best k centroids from the data, we first find the point in the data set with the tightest neighborhood of m_1 other points and the corresponding $\frac{m_1}{n}$-quantile of the distances from it. This involves the computation of all the pairwise distances, which has a time complexity of $O(n^2 d)$, computing the $(1-\beta)$-quantile of the distances for all the points, which has a time complexity $O(n^2 \log n)$, and finally computing the minimum which takes $\log n$ at most. Once we have found the first centroid, for each $1 \leq \ell_k \leq 2/\beta^2$ we construct $\mathcal{P}_k^{(\ell_k)}, \overline{\mathcal{P}_k^{(\ell_k)}}$ according to distances from the first centroid, which takes another n unit time and perform the $k-1$ centroid finding algorithm on $\overline{\mathcal{P}_k^{(\ell_k)}}$ which has at most n points. Let U_k be the time complexity of finding the best k-centroids given n data points. Then, given the above reasoning, we have

$$U_k \leq (O(1)/\beta^2)\,[U_{k-1} + n] + O(n^2 d + n^2 \log n).$$

Solving the above recursion we get

$$U_k \leq (O(1)/\beta^2)^{k-2} U_2 + (O(1)/\beta^2)^{k-2} \Big[n^2 d + n^2 \log n + 2n/\beta^2 \Big].$$

note that via a similar argumentthe 2-centroid finding problem takes $\frac{O(1)}{\beta^2}\left(n^2 d + n^2 \log n\right)$ time. Hence, we simplify the above to get the desired result. □

V. Clustering in the presence of adversarial outliers

In this section we show that the inherent robustness of our algorithms extends to the scenario when a constant fraction of data (for a finite k) might be adversarial outliers. We study the setup where an adversary, after accessing the original data set, adds n^{out} many new points of its choice. Our results in this section primarily address how the previous theoretical guarantees change when outliers might be present and how much adversarial contamination our initialization methods can tolerate without misbehaving significantly. We do not aim to optimize the outlier levels that our algorithms can tolerate; it is left for future works.

We first discuss the extension of Theorem 2. To retain the above mislabeling guarantees, we need to apply a higher value of the truncation parameter δ for running TM_δ. We have the following guarantees.

Theorem 11. *Suppose that an adversary, after analyzing the data $Y_1, \ldots, Y_n$ coming from the general mixture model (4), adds $n^{\mathsf{out}} = n\alpha(1-\psi)$ many outliers of its choice for some $\psi \in (0,1]$. Then there exists a constant $c_0 > 0$ such that the following holds.*

- *If we have an initial estimate of the label vector satisfying $H_0 \geq \frac{1}{2} + \gamma$ with $\gamma \in (\frac{10}{n\alpha}, \frac{1}{2})$, then whenever $\mathsf{SNR} \geq G^{-1}\left(\exp\left\{-\frac{c_0}{\alpha \cdot \min\{\gamma,\psi\}}\right\}\right)$ we have that the label vector output $\widehat{z}^{(s)}$ obtained from the COD_δ algorithm after s iterations, with $\delta = \frac{1}{2} - \frac{1}{4}\min\left\{\gamma, \frac{\psi}{6}\right\}$, achieves*

$$\mathbb{E}\left[\ell(\widehat{z}^{(s)}, z)\right] \leq k^2 G\left(\mathsf{SNR} - G^{-1}\left(\exp\left\{-\frac{c_0}{\alpha\psi}\right\}\right)\right) + 8ke^{-\frac{n\alpha}{4}}, \quad s \geq 2.$$

- *Instead of initial cluster labels, if we have initial centroid estimates satisfying $\Lambda_0 \leq \frac{1}{2} - \frac{G^{-1}\left(\exp\left\{-\frac{c_0}{\alpha\psi}\right\}\right)}{\mathsf{SNR}}$, then the above conclusion holds with $\gamma = 0.3$.*

Remark 10. Since the adversarial data are arbitrary, the result also applies to the case where the number of clusters is undetermined. In that case, we can simply regard data beyond the first k class as adversarial attacks, so long as the total number of such data points is not too large. Then, Theorem 11 gives an accurate bound on the first k clusters.

Remark 11. The recent work of [29] studies a similar setup of adversarial outliers, although in the specific case of sub-Gaussian mixture models. The previous work uses the coordinate-wise median for centroid estimation. We expect that the similar theoretical guarantees as in Theorem 11 for the above coordinate-wise median based clustering algorithm will require $\mathsf{SNR} \geq \sqrt{d} G^{-1} \exp\left\{-\frac{c_0}{\alpha \cdot \min\gamma,\psi}\right\}$, which is worse than the results in this paper by a factor of $\sqrt{d}$. This is because in the worst case, the outliers can equally impact the estimation guarantees of the coordinate-wise median in each coordinate, which produces this $\sqrt{d}$ factor. Our algorithm can avoid this extra factor of $\sqrt{d}$ as it utilizes the underlying Euclidean distance-based structure of the error distributions, in contrast to the algorithm based on the coordinate-wise median, which originates from an ℓ_1-norm based minimization procedure.

Remark 12. Our method requires that the proportion of outliers is, at most, the proportion of points in the smallest cluster. It is a requirement to produce a small mislabeling as adversarial setups that allow larger outliers than the above can allow

the adversary to substitute the smallest cluster with a different cluster of their choice and mislead any reasonable clustering algorithm.

Next, we discuss the results regarding our initialization algorithms in the presence of adversarial outliers.

Theorem 12. *Suppose that out of the n many observed data points, n_i^* many are from cluster $T_i^*, i = 1, \ldots, k$ and n^{out} many are adversarial outliers (i.e., $\sum_{i=1}^k n_i^* + n^{\mathsf{out}} = n$). Also, assume that for some constant $\alpha > 0$ the counts satisfy $n_i^* > n\alpha, i = 1, \ldots, k$. We apply Algorithm 4 for $k = 2$ and Algorithm 5 for $k \geq 3$. Then the consequences of Theorem 6 (i.e., for $k = 2$) hold if $n^{\mathsf{out}} \leq \frac{n\alpha^2}{32}$ and the consequences of Theorem 8 (i.e., for a general $k \geq 3$) hold if $n^{\mathsf{out}} \leq \frac{n\alpha^2}{64k}$.*

Remark 13. The above result guarantees that for $k = O(1), \alpha > c/k > 0$, our clustering algorithm can tolerate $O(n)$ outliers in the data and still retain the initialization guarantees similar to the case without outliers. It is reasonable to comment that with an increase in the number of clusters, i.e., comparatively fewer points in each cluster, the tolerable amount of adversarial outliers should decrease for any algorithm. However, optimizing the number of adversarial outliers our algorithm can tolerate for a general k is beyond the scope of the current work and is left for future directions.

Remark 14 (Flexibility of constraints on α, n^{out})**.** The theoretical bound on the allowable amount of outliers in our initialization results has not been optimized, and we leave the work for future. In practice, our algorithm can perform excellent clustering even when the number of outliers is significantly larger compared to the theoretical requirement of $\frac{n\alpha^2}{64k}$. As an example, we considered a three cluster setups with centroids $[0, 0], [20, 0], [10, 20]$, with error distributions $N(0, 4I_{2\times 2})$. For clustering purposes, we generated 50 data points from cluster 1, 100 data points from cluster 2, and 150 data points from cluster 3, hence the minimum cluster proportion $\alpha = 0.125$. Additionally, we introduced 40 more outlier points to the data, where each coordinate of the outliers points were generated uniformly from $[0, 20]$. Note that this is significantly more than the threshold $n\alpha^2/64k \approx 0$ of outliers that is required for our theoretical results to hold. To implement our initialization algorithms, we varied the guess of α in the set $[0.333, 0.167, 0.083, 0.042]$ and picked the parameter δ for computing trimmed mean as $\delta = 0.1$. We also considered $m = 5$ in our initialization algorithm to speed up the iterative process. We found that in all the above instances of α-guesses our algorithms were able to provide excellent clustering outputs. One of the simulations is given below in Fig. 1.

VI. Suboptimality of Lloyd's algorithm

In this section, we establish that Lloyd's algorithm might produce a suboptimal mislabeling even when the initial labels are reasonably good. In the case of at least three centroids, even when error distributions have bounded support, if one of the centroids is far away from the rest, then the mislabeled points originating from that centroid can destabilize the cluster means and hence lead to the poorly estimated centroid. In the two-centroid setup, the suboptimality occurs when error distributions exhibit heavy tails.

A. The case of at least three centroids

This section assumes that whenever Lloyd's algorithm produces an empty cluster, it randomly picks one of the data points as the missing centroid for the next iteration step. Then, we have the following result.

Lemma 13. *Given any $\beta \in (0, 1)$, there exists a system of three centroids and an initialization with mislabeling proportion β such that Lloyd's algorithm does not produce better than a constant proportion of mislabeling.*

Proof. We consider the one dimensional setup with three centroids, located at $-\frac{\Delta}{2}, \frac{\Delta}{2}$ and $\frac{c\Delta}{2\beta}$ for some constant $c > 2$ and sufficiently large Δ. Consider the data generating model

$$\begin{aligned} Y_i &= \theta_{z_i} + w_i, \ \ i \in [n], \\ w_i &\overset{\text{iid}}{\sim} \text{Uniform}(-1, 1), \ \ z_i \in \{1, 2, 3\}, \\ \theta_1 &= -\frac{\Delta}{2}, \theta_2 = \frac{\Delta}{2}, \theta_3 = \frac{c\Delta}{2\beta}. \end{aligned} \tag{15}$$

Let $T_h^* = \{i \in [n] : z_i = h\}, h \in \{1, 2, 3\}$ be as before. We assume equal number of points in all three clusters, i.e., $|n_h^*| = n/3$. To define the initial label estimates, choose any $\lceil n\beta/3 \rceil$ points from T_3^*, say S and consider the initialization

$$\widehat{z}_i^{(0)} = \begin{cases} 2 & \text{if } i \in S, \\ z_i & \text{otherwise.} \end{cases} \tag{16}$$

This is a good initialization, except for a fraction of β mislabels in class 2.

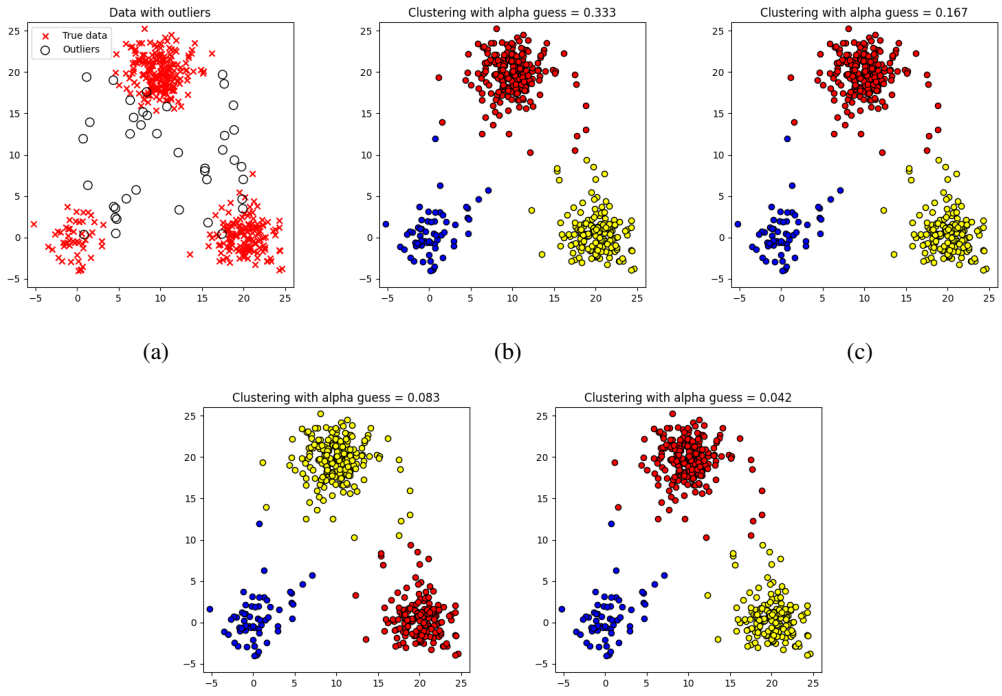

Fig. 1: Clustering with different guesses of α. In plot (a) we have the data, including outliers. The information on which data points are outliers are unknown to the algorithm. Plots (b)–(e) presents the clustering outputs for different guesses of α and the estimated cluster labels are shown using the different colors $yellow, red, blue$. The above plots show that our algorithm performs excellent clustering even in presence of significant number of outliers and with incorrect guesses of lower bounds on $\alpha = 0.125$. Plots (d) and (e) have different cluster labels depending on the ordering of how the labeling was performed during the execution of the algorithms, however the clustering is similar up to label permutations.

We now study the iteration steps for Lloyd's algorithm. After the first iteration, assuming Δ is sufficiently large, the centroid estimates satisfy

$$\begin{aligned}
\theta_1^{(0)} &\in \left(-\frac{\Delta}{2} - 1, \frac{\Delta}{2} + 1\right), \\
\widehat{\theta}_2^{(0)} &\in \left(\frac{(c+1)\Delta}{2(1+\beta)} - 1, \frac{(c+1)\Delta}{2(1+\beta)} + 1\right), \\
\theta_3^{(0)} &\in \left(\frac{c\Delta}{2\beta} - 1, \frac{c\Delta}{2\beta} + 1\right).
\end{aligned}$$

Note that the above implies that given any data point, it is either closer to $\widehat{\theta}_1^{(0)}$ or to $\widehat{\theta}_3^{(0)}$, depending on whether the data is from clusters 1 and 2 or from cluster 3. As a result, $T_2^{(1)}$ is empty, and we randomly pick one of the data points as $\widehat{\theta}_2^{(1)}$. With a constant probability, the choice is given by one of the points in $\{Y_i : i \in T_3^*\}$. In that scenario, in all subsequent stages $\widehat{\theta}_2^{(s)}, \widehat{\theta}_3^{(s)}$ will continue to be inside the interval $(\frac{c\Delta}{2\beta} - 1, \frac{c\Delta}{2\beta} + 1)$. As a result, all the points from T_2^* are mislabeled. This shows that with constant probability we will have a constant proportion of mislabeling even if all possible label permutations are considered. □

B. The case of two centroids

We produce a counter example where Lloyd's algorithm fails even with a good initialization. Fix $\epsilon \in (0, 1)$. Given any $\Delta > 0$ we choose a sample size so big that $n^\epsilon > 4\Delta$. Next consider the decay function

$$G(x) = \frac{1}{1 + x^{1-\epsilon}}, \quad x > 0. \tag{17}$$

The model we use is

$$Y_i = \theta_{z_i} + w_i,\ i \in [n],$$
$$w_i \overset{\text{iid}}{\sim} W, \quad \mathbb{P}[W > x] = G(x), x > 0,$$
$$z_i \in \{1,2\}, \theta_1 = 0, \theta_2 = \Delta$$

with equal cluster sizes. Then, given n samples from the above mixture model, we have

$$\begin{aligned}
&\mathbb{P}\left[\cup_{i=1}^n \left\{w_i > n^{1+\epsilon}\right\}\right] \\
&= 1 - \mathbb{P}\left[\cap_{i=1}^n \left\{w_i \le n^{1+\epsilon}\right\}\right] \\
&= 1 - \Pi_{i=1}^n \mathbb{P}\left[w_i \le n^{1+\epsilon}\right] \\
&= 1 - \left(1 - \frac{1}{1+n^{1-\epsilon^2}}\right)^n \ge 1 - e^{-n^{\epsilon^2}}.
\end{aligned}$$

This implies that with probability at least 1/2 there is at least one index i^* such that $w_{i^*} > n^{1+\epsilon}$. Then, whichever cluster contains Y_{i^*}, its corresponding centroid estimate will be bigger than n^ϵ. Notably, in the next step, when we use the Euclidean distance to cluster estimate, the best estimated clusters will be of the form

$$\begin{aligned}
T_1^{(s+1)} &= \{i \in [n] : Y_i \in [0, x]\}, \\
T_2^{(s+1)} &= \{i \in [n] : Y_i \in (x, \infty)\}, \\
x &= (\widehat{\theta}_1^{(s)} + \widehat{\theta}_2^{(s)})/2.
\end{aligned}$$

As one of the centroid estimates is bigger than n^ϵ we get that $x \ge n^\epsilon/2 \ge 2\Delta$. Next, we present the following concentration result.

Lemma 14. *Fix $\epsilon_0 > 0$. Then there is an event $\mathcal{E}_{\epsilon_0}^{\text{con}}$ with probability at least $1 - k \cdot e^{-\frac{\min_{g\in[k]} n_g^*}{4}}$ on which*

$$\sum_{i \in T_g^*} \mathbf{1}_{\{\epsilon\Delta \le \|w_i\|\}} \le \frac{5n_g^*}{4\log(1/G(\epsilon_0\Delta/\sigma))}, \quad \epsilon \ge \epsilon_0, \forall g \in [k].$$

A proof of the above result is presented at the end of this section. Note that in view of Lemma 14, for all large enough Δ and n we have

$$\mathbb{P}\left[\sum_{i \in T_h^*} \mathbf{1}_{\{w_i < \Delta\}} > \frac{3n}{8},\ h \in \{1,2\}\right] \ge \frac{3}{4}.$$

In view of $x \ge 2\Delta$, using the above inequality conditioned on the event $\cup_{i=1}^n \{w_i > n^{1+\epsilon}\}$ we have that

$$\mathbb{P}\left[|\widehat{T}_1^{(s+1)} \cap T_h^*| \ge \frac{3n}{8},\ h \in \{1,2\}\right] \ge \frac{3}{4}.$$

Hence, on the event $\cup_{i=1}^n \{w_i > n^{1+\epsilon}\}$, that has a probability at least 1/2, there will be at least $\frac{3n}{8}$ points that are mislabeled.

Proof of Lemma 14. We define $B_i = \mathbf{1}_{\{\epsilon_0\Delta \le \|w_i\|\}}$. As $\epsilon \ge \epsilon_0$, it is enough to find an event $\mathcal{E}_1$ with the said probability on which

$$\mathbb{P}\left[\sum_{i \in T_g^*} B_i \ge \frac{5n_g^*}{4\log(1/G(\epsilon_0\Delta/\sigma))}\right] \le e^{-\frac{n_g^*}{4}} \text{ for each } g \in [k]. \tag{18}$$

Note that

$$\mathbb{P}[\epsilon_0\Delta \le \|w_i\|] \le G(\epsilon_0\Delta/\sigma). \tag{19}$$

This implies $\sum_{i \in T_g^*} B_i$ is stochastically smaller than a random variable distributed as $\text{Binom}(n_g^*, G(\epsilon_0\Delta/\sigma))$. We continue to analyze (18) via Chernoff's inequality in Lemma 21 for a random variable with the $\text{Binom}(n_g^*, G(\epsilon_0\Delta/\sigma))$ distribution. Denote

$$a = \frac{5}{4\log(1/G(\epsilon_0\Delta/\sigma))}, \quad m = n_g^*, \quad q = e^{-5/(4a)} = G(\epsilon_0\Delta/\sigma).$$

Then we have $a = \frac{5}{4\log(1/q)} > \frac{1}{\log(1/q)} \geq q$. Using $a\log a \geq -0.5$ for $a \in (0,1)$ we get for $h_q(a) = a\log\frac{a}{q} + (1-a)\log\frac{1-a}{1-q}$

$$\begin{aligned}
&\mathbb{P}\left[\sum_{i\in T_g^*} B_i \geq \frac{5n_g^*}{4\log(1/G(\epsilon_0\Delta/\sigma))}\right] \\
&\leq \exp\left(-mh_q(a)\right) \\
&\leq \exp\left(-m\left(a\log\frac{a}{q} + (1-a)\log\frac{1-a}{1-q}\right)\right) \\
&\leq \exp\left(-m\left\{a\log\frac{a}{e^{-5/(4a)}} + (1-a)\log(1-a)\right\}\right) \\
&= \exp\left(-m\left\{a\log a + (1-a)\log(1-a) + \frac{5}{4}\right\}\right) \leq e^{-n_g^*/4}.
\end{aligned}$$

□

VII. Experiments

A. Synthetic datasets

In this section, we evaluate our proposed algorithm (IOD for initialization and COD for clustering) on synthetic datasets and compare its performance in terms of the mislabeling proportion with the classical Lloyd's algorithm (e.g., the Lloyd–Forgy algorithm [30]). For initializing Lloyd's algorithm, we consider three methods:

- the proposed IOD algorithm
- the k-means++ algorithm [21]
- randomly chosen initial centroid estimates from the dataset.

We simulate the data points with the errors $\{w_i\}$ independently from the multivariate t_ν-distribution with a scale parameter σ, i.e., the w_i random variable has a density

$$f(x) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)\nu^{d/2}\pi^{d/2}\sigma}\left[1 + \frac{\|x\|^2}{\sigma\nu}\right]^{-(\nu+d)/2}. \tag{20}$$

We study the effect of different dimensions d, degrees of freedom ν for the t distribution, and the scale parameter σ. We consider the number of centroids $k = 2, 3$ for our experiments. The centroids of the cluster components are generated randomly and then scaled to make sure that they are at least 25 units apart. For each of the clusters, we generate 200 data points. When running the IOD initialization method in Algorithm 4, Algorithm 5 and the COD clustering method in Algorithm 2, we use the parameters

$$m_1 = 20, m = 10, \beta = 0.05, \delta = 0.3.$$

Our experiments are divided into the following regimes.

- *Different degrees of freedom.* We fix the data dimension $d = 5$ and $\sigma = 5$. We vary the degrees of freedom ν in the set $\{1, 1.5, 10\}$ to cover the cases of a very heavy tail where the mean does not exist, a moderately heavy tail where the mean exists but variance does not, and finally a very light tail where other higher moments exist.
- *Different scale parameters.* We fix the data dimension $d = 10$ and $\nu = 1.5$. We vary the scale parameter σ in the set $\{1, 5, 10\}$ to cover the cases of large, moderate, and low signal-to-noise ratios, respectively.
- *Different dimensions.* The true points are generated with $\nu = 1.5, \sigma = 5$. We vary the data dimension d in the set $\{2, 10, 30\}$.

We repeat all the experiment setups 150 times to estimate the mislabeling proportion and its standard error. The average mislabeling errors are presented in Table I, Table II, Table III (along with the standard errors within the parenthesis).

Results: We first present the numerical study describing the effect of ν Table I. For the large value of $\nu = 10$ the data are supposed to be highly concentrated around the centroids, which should guarantee a low mislabeling error. Lloyd's algorithm should work well in such a light tail setup, even though its mislabeling optimality is unknown. Nonetheless, our simulations demonstrate a low mislabeling error for all the algorithms for both $k = 2, 3$. As we consider heavier tails by decreasing ν to 1.5, we observe a steep increase in the mislabeling error for all the methods, although our algorithm produces the best performance. Notably, Lloyd's algorithm, when paired with our proposed IOD initialization method, improves on the performance of the classical k-means++ initialization technique. However, further decreasing ν to 1, a setup where even the population mean does not exist, all instances of Lloyd's-type methods perform equally poorly, while our algorithm produces significantly lower mislabeling.

Next, we demonstrate the effect of the scale parameter σ in Table II. For a fixed ν, Δ this amounts to studying the effect of $\mathsf{SNR} = \frac{\Delta}{2\sigma}$ on the mislabeling error. The proportion of mislabeling should decay with large SNR, or equivalently with low σ

TABLE I: Effect of degrees of freedom: $n = 200k, \sigma = 5, d = 5, \Delta = 25$

k	ν	COD + IOD	Lloyd + IOD	Lloyd + k-means++
2	1	0.322 (0.011)	0.495 (0.002)	0.498 (0.000)
	1.5	0.128 (0.001)	0.322 (0.014)	0.48 (0.006)
	10	0.014 (0.000)	0.013 (0.000)	0.014 (0.000)
3	1	0.422 (0.005)	0.652 (0.004)	0.664 (0.000)
	1.5	0.364 (0.007)	0.411 (0.009)	0.576 (0.011)
	10	0.043 (0.008)	0.034 (0.007)	0.014 (0.000)

k	ν	Lloyd + random init
2	1	0.497 (0.000)
	1.5	0.366 (0.014)
	10	0.014 (0.000)
3	1	0.65 (0.005)
	1.5	0.403 (0.013)
	10	0.081 (0.013)

TABLE II: Effect of scale: $n = 200k, \nu = 1.5, d = 10, \Delta = 25$

k	σ	COD + IOD	Lloyd + IOD	Lloyd + k-means++
2	1	0.014 (0.000)	0.029 (0.006)	0.274 (0.018)
	5	0.173 (0.003)	0.424 (0.006)	0.496 (0.001)
	10	0.352 (0.005)	0.492 (0.001)	0.497 (0.000)
3	1	0.161 (0.012)	0.169 (0.012)	0.27 (0.017)
	5	0.412 (0.001)	0.485 (0.006)	0.654 (0.003)
	10	0.509 (0.003)	0.628 (0.005)	0.664 (0.000)

k	σ	Lloyd + random init
2	1	0.1 (0.014)
	5	0.451 (0.005)
	10	0.495 (0.001)
3	1	0.169 (0.013)
	5	0.53 (0.008)
	10	0.647 (0.004)

values, and this is supported by our demonstrations. Additionally, in all the setups, our algorithm performs significantly better than its competitors.

In Table III, we demonstrate how the data dimensions affect the performance of our algorithm. As the data dimension increases while keeping the centroid separation fixed, the performance of the clustering algorithm deteriorates. This is because the norm of the error random variables increases proportionally to the square root of the dimension, multiplied with variability in each coordinate. Nonetheless, our proposed clustering algorithm performs more robustly compared to the other methods in the simulation studies. It might be possible to improve all the clustering techniques by applying some dimension reduction, for example, feature screening approaches [55] and the spectral methods in [13], to the data set before applying the clustering methods. However, such analysis is beyond the scope of the current work.

B. Real data experiments

Furthermore, we evaluated our proposed algorithm on the publicly available Letter Recognition dataset [56]. The data set contains 16 primitive numerical attributes (statistical moments and edge counts) of black-and-white rectangular pixel displays of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts, and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. We apply our proposed algorithm to this data with the aim of clustering data points corresponding to the same letters together. Additionally, we explore the robustness

TABLE III: Effect of dimension: $n = 200k, \nu = 1.5, \sigma = 5, \Delta = 25$

k	d	COD + IOD	Lloyd + IOD	Lloyd + k-means++
2	2	0.099 (0.001)	0.154 (0.007)	0.398 (0.01)
	10	0.174 (0.004)	0.414 (0.008)	0.495 (0.001)
	30	0.309 (0.01)	0.492 (0.002)	0.497 (0.000)
3	2	0.156 (0.008)	0.2 (0.009)	0.38 (0.009)
	10	0.41 (0.002)	0.479 (0.009)	0.655 (0.004)
	30	0.467 (0.002)	0.64 (0.002)	0.664 (0.000)

k	d	Lloyd + random init
2	2	0.231 (0.01)
	10	0.445 (0.006)
	30	0.494 (0.002)
3	2	0.236 (0.009)
	10	0.528 (0.011)
	30	0.653 (0.004)

guarantees of the algorithms when some small number of data points corresponding to other contaminating letter classes are also present. In that setup, the goal is to minimize the mislabeling error corresponding to the letter classes with larger sample sizes.

Experiment setup: For our experiment, we consider two and three-cluster setups. In the two-cluster setup, we pick the data points corresponding to the letters "W" and "V" as the clusters, and in the three-cluster setup, we pick the data points corresponding to the letters "X", "M", and "A". We randomly sample 100 points from each cluster in both setups to simulate a contamination-free setup. To introduce outliers, in each scenario, we add 20 randomly chosen data points corresponding to the letter "R". Once the data set is prepared, we apply the following clustering algorithms

- the proposed IOD initialization algorithm and COD clustering algorithm
- the hybrid k-median algorithm in [29] for clustering in the presence of adversarial outliers, initialized with the IOD algorithm
- Lloyd's algorithm initialized with the IOD algorithm
- Lloyd's algorithm initialized with the k-means++ algorithm
- Lloyd's algorithm initializations from the dataset.

The relevant parameters for the clustering are the same as those in the simulation studies section, with the only modification being for the value of δ. This is in accordance with Theorem 11, which proposes that in the presence of outliers, it is meaningful to choose a more robust clustering algorithm, which corresponds to a higher value of δ. For our studies, we fix $\delta = 0.48$. The entire process, starting from data generation to applying the algorithms, is independently repeated 150 times to measure the average mislabeling proportion and the corresponding standard errors. The results are presented in Table IV (the standard errors are presented within the parentheses beside the average mislabeling values).

Results: All the results show that our method consistently yields the lowest proportion of mislabeling, outperforming the other algorithms. Remarkably, our method produces a better mislabeling rate even in the absence of outliers. This indicates a heavy tail structure in the data set. Interestingly, in the two cluster setup, the mislabeling proportion reduces in the presence of data points from the letter class "R". This is possible, as we did not aim to pick the outlier class that distorts the clustering process. We instead study the effect of a particular outlier class. This indicates a similarity of the data points from the outlier class with one of the clusters, resulting in more points being observed in the neighborhood of the corresponding cluster. When we observe more points in the clusters, separating the clusters becomes much more accessible, resulting in lower mislabeling.

TABLE IV: Results for clustering letters: $n = 100k, \delta = 0.48$, outlier proportion = 20%, outlier class = R

Classes	Outliers	COD + IOD	k-median + IOD	Lloyd + IOD
W, V	without	0.276 (0.008)	0.32 (0.005)	0.391 (0.005)
	with	0.269 (0.008)	0.317 (0.004)	0.381 (0.004)
X, M, A	without	0.194 (0.010)	0.245 (0.007)	0.374 (0.004)
	with	0.264 (0.009)	0.275 (0.006)	0.388 (0.003)

Classes	Outliers	Lloyd + k-means++	Lloyd + random
W, V	without	0.355 (0.004)	0.402 (0.004)
	with	0.352 (0.004)	0.398 (0.004)
X, M, A	without	0.342 (0.006)	0.357 (0.007)
	with	0.354 (0.004)	0.379 (0.005)

VIII. Conclusion:

Our paper investigates the problem of initialization and clustering in a general heavy-tailed setup, allowing for the possibility of adversarial contamination. We propose novel algorithms that can provide consistent clustering as the cluster separation and sample size increase. Our algorithm runs in $O(n^2(d+\log n))$ when the number of clusters is finite. We establish the minimax rate optimality of our algorithm by obtaining thresholds on the mislabeling proportions for the above clustering problem in finite dimensions. Potential future directions of our work include developing robust algorithms for high-dimensional data that can perform efficient and robust clustering with an increasing number of clusters.

Appendix A
Mislabeling upper bound in Theorem 2

A. Preparation

The proof of Theorem 2 is primarily based on the following two stages:

(a) analyzing accuracy of the clustering method based on current centroid estimates,
(b) analyzing the next centroid updates based on current labels.

We obtain results on these steps separately and then combine them to prove Theorem 2. Our analysis depends on high-probability events $\mathcal{E}_{\tau}^{\mathsf{norm}}, \mathcal{E}_{\gamma_0,\epsilon_0}$ given in the following lemmas.

Lemma 15. *Suppose that w_i-s are independent random variables satisfying the G_σ-decay condition (C3) and $\beta \in (0,1)$ is fixed. Then given any $\tau > 0$ there is an event $\mathcal{E}_\tau^{\mathsf{norm}}$ with probability at least $1 - e^{-0.3n}$ on which the following holds. For any $S \subseteq [n]$ with $|S| \geq n\beta$, the cardinality of the set*

$$\left\{ i \in S : \|w_i\| \leq \sigma G^{-1}\left(\exp\left\{-\frac{1+1/\beta}{\tau}\right\}\right)\right\}$$

is at least $(1-\tau)|S|$.

The following lemma provides a lower bound on H_{s+1} based on Λ_s and establishes an upper bound on Λ_s in terms of H_s.

Lemma 16. *Fix any $\epsilon_0 \in (0, \frac{1}{2}), \gamma_0 \in (\frac{10}{n\alpha}, \frac{1}{2})$. Then whenever $\Delta, \sigma > 0$ satisfies $\frac{5}{2\alpha \log(1/G(\epsilon_0\Delta/(2\sigma)))} < \frac{1}{2}$ the following holds true. There is an event $\mathcal{E}_{\gamma_0,\epsilon_0}$ with a probability of at least $1 - 2ke^{-n\alpha/4} - e^{-0.3n}$ on which for all $s \geq 0$, the COD_δ algorithm with $\delta \in (\frac{1}{2} - \frac{\gamma_0}{4}, \frac{1}{2})$ ensures:*

(i) if $\Lambda_s \leq \frac{1}{2} - \epsilon_0$ then $H_{s+1} \geq 1 - \frac{5}{2\alpha \log(1/G(\epsilon_0\Delta/(2\sigma)))}$,

(ii) if $H_{s+1} \geq \frac{1}{2} + \gamma_0$ then

$$\Lambda_{s+1} \leq \frac{8\sigma}{\Delta} G^{-1}\left(\exp\left\{-\frac{1+2/\alpha}{\tau}\right\}\right),$$

where $\tau = \frac{\gamma_0}{1+2\gamma_0}$.

Proof of Lemma 16. We first prove part (i). For any $g \neq h \in [k] \times [k]$,

$$\begin{aligned}
&\mathbf{1}_{\left\{z_i = g, \widehat{z}_i^{(s+1)} = h\right\}} \\
&\leq \mathbf{1}_{\left\{\|Y_i - \widehat{\theta}_h^{(s)}\|^2 \leq \|Y_i - \widehat{\theta}_g^{(s)}\|^2,\ i \in T_g^*\right\}} \\
&= \mathbf{1}_{\left\{\|\theta_g + w_i - \widehat{\theta}_h^{(s)}\|^2 \leq \|\theta_g + w_i - \widehat{\theta}_g^{(s)}\|^2\right\}} \\
&= \mathbf{1}_{\left\{\|\theta_g - \widehat{\theta}_h^{(s)}\|^2 - \|\theta_g - \widehat{\theta}_g^{(s)}\|^2 \leq 2\langle w_i, \widehat{\theta}_h^{(s)} - \widehat{\theta}_g^{(s)}\rangle\right\}}.
\end{aligned} \tag{21}$$

The triangle inequality and the fact

$$\|\theta_h - \widehat{\theta}_h^{(s)}\| \leq \Lambda_s \Delta \leq \Lambda_s \|\theta_g - \theta_h\|$$

we get

$$\begin{aligned}
\|\theta_g - \widehat{\theta}_h^{(s)}\|^2 &\geq \left(\|\theta_g - \theta_h\| - \|\theta_h - \widehat{\theta}_h^{(s)}\|\right)^2 \\
&\geq (1 - \Lambda_s)^2 \|\theta_g - \theta_h\|^2.
\end{aligned}$$

This implies

$$\begin{aligned}
&\|\theta_g - \widehat{\theta}_h^{(s)}\|^2 - \|\theta_g - \widehat{\theta}_g^{(s)}\|^2 \\
&\geq (1 - \Lambda_s)^2 \|\theta_g - \theta_h\|^2 - \Lambda_s^2 \|\theta_g - \theta_h\|^2 \\
&= (1 - 2\Lambda_s)\|\theta_g - \theta_h\|^2.
\end{aligned} \tag{22}$$

In view of the last inequality, using the fact

$$\begin{aligned}
&|\langle w_i, \widehat{\theta}_h^{(s)} - \widehat{\theta}_g^{(s)}\rangle| \\
&\leq \|w_i\| \cdot \|\widehat{\theta}_h^{(s)} - \widehat{\theta}_g^{(s)}\| \\
&\leq \|w_i\| \cdot (\|\widehat{\theta}_g^{(s)} - \theta_g\| + \|\widehat{\theta}_h^{(s)} - \theta_h\| + \|\theta_g - \theta_h\|) \\
&\leq \|w_i\|(2\Lambda_s + 1)\|\theta_g - \theta_h\|,
\end{aligned}$$

we continue (21) to get

$$\mathbf{1}_{\left\{z_i = g, \widehat{z}_i^{(s+1)} = h\right\}} \leq \mathbf{1}_{\left\{\frac{1-2\Lambda_s}{2(1+2\Lambda_s)}\|\theta_g - \theta_h\| \leq \|w_i\|\right\}}. \tag{23}$$

Simplifying above with $\Lambda_s \leq \frac{1}{2} - \epsilon_0$

$$\mathbf{1}_{\left\{z_i = g, \widehat{z}_i^{(s+1)} = h\right\}} \leq \mathbf{1}_{\left\{\frac{\epsilon_0}{2}\|\theta_g - \theta_h\| \leq \|w_i\|\right\}}$$

Summing $\mathbf{1}_{\left\{z_i=g,\widehat{z}_i^{(s+1)}=h\right\}}$ over $\left\{i \in T_g^*\right\}$, in view of Lemma 14, we get on the event $\mathcal{E}_{\epsilon_0}^{\mathsf{con}}$

$$\begin{aligned} n_{gh}^{(s+1)} &\leq \sum_{i \in T_g^*} \mathbf{1}_{\left\{\frac{\epsilon_0}{2}\Delta \leq \|w_i\|\right\}} \\ &\leq \frac{5n_g^*}{4\log(1/G(\epsilon_0\Delta/(2\sigma)))}, \ \forall h \in [k], h \neq g. \end{aligned} \tag{24}$$

Using the last display and noting that $k \leq \frac{1}{\alpha}$ (as the proportion of points in the smallest cluster is at most $\frac{1}{k}$ by the Pigeon-Hole Principle) and $n_g^* \geq n\alpha$ we get

$$\begin{aligned} \frac{\sum_{\substack{h \in [k] \\ h \neq g}} n_{gh}^{(s+1)}}{n_g^*} &\leq \frac{5k}{4\log(1/G(\epsilon_0\Delta/(2\sigma)))} \\ &\leq \frac{5}{4\alpha\log(1/G(\epsilon_0\Delta/(2\sigma)))}. \end{aligned} \tag{25}$$

Next, we switch g, h in (24) and sum over $h \in [k], h \neq g$. We get on the event $\mathcal{E}_{\epsilon_0}^{\mathsf{con}}$

$$\sum_{\substack{h \in [k] \\ h \neq g}} n_{hg}^{(s+1)} \leq \frac{5\sum_{h \in [k], h \neq g} n_h^*}{4\log(1/G(\epsilon_0\Delta/(2\sigma)))} \leq \frac{5n}{4\log(1/G(\epsilon_0\Delta/(2\sigma)))}.$$

Using the relation between $\epsilon_0, \Delta, \sigma, \alpha$ in the lemma statement we get

$$\frac{\sum_{\substack{h \in [k] \\ h \neq g}} n_{gh}^{(s+1)}}{n_g^*} \leq \frac{1}{2},$$

which implies

$$n_g^{(s+1)} \geq n_{gg}^{(s+1)} = n_g^* - \sum_{\substack{h \in [k] \\ h \neq g}} n_{gh}^{(s+1)} \geq \frac{1}{2} n_g^* \geq \frac{1}{2} n\alpha.$$

This gives us

$$\frac{\sum_{h \neq g} n_{hg}^{(s+1)}}{n_g^{(s+1)}} \leq \frac{5}{2\alpha\log(1/G(\epsilon_0\Delta/(2\sigma)))}.$$

Using the above with (25) we get with probability $1 - 2kn^{-c/4}$

$$\begin{aligned} H_{s+1} &= 1 - \max\left\{\frac{\sum_{h \neq g} n_{gh}^{(s+1)}}{n_g^*}, \frac{\sum_{h \neq g} n_{hg}^{(s+1)}}{n_g^{(s+1)}}\right\} \\ &\geq 1 - \frac{5}{2\alpha\log(1/G(\epsilon_0\Delta/(2\sigma)))}. \end{aligned}$$

Next we present below the proof of Lemma 16(ii). We use Proposition 17 provided below.

Proposition 17. *Suppose that given any $\tau \in (0,1)$, there is an event $\mathcal{F}_\tau$ and a number $D_\tau > 0$ such that, on $\mathcal{F}_\tau$, for any $S \subseteq [n]$ with $|S| > \frac{n\alpha}{2}$, the cardinality of the set $\{i \in S : \|w_i\| \leq D_\tau\}$ is at least $(1-\tau)|S|$. Then for any $\gamma \in (10/n\alpha, 1/2)$, using $\tau = \frac{\gamma}{1+2\gamma}$, we get that on the event $\mathcal{F}_\tau$, if $H_s \geq \frac{1}{2} + \gamma$, then the COD_δ algorithm with any $\delta \in (\frac{1}{2} - \frac{\gamma}{4}, \frac{1}{2})$ returns $\Lambda_s \leq 8D_\tau/\Delta$.*

A proof of the above result is provided at the end of this section. We choose $\mathcal{F}_\tau = \mathcal{E}_\tau^{\mathsf{norm}}$ and

$$\gamma = \gamma_0, \quad \tau = \frac{\gamma_0}{1+2\gamma_0}, \quad D_\tau = \sigma G^{-1}\left(\exp\left\{-\frac{1+2/\alpha}{\tau}\right\}\right).$$

In view of Lemma 15 note that the event $\mathcal{F}_\tau$ has probability at least $1 - e^{-0.3n}$ and satisfies the requirement in Proposition 17. This implies that we get the required bound on Λ_{s+1}.

Combining the proof of part (i) we conclude that both the claims hold with probability at least $1 - 8ke^{-\frac{n\alpha}{4}}$. □

Proof of Proposition 17. We prove the above result using a contradiction. Fix $\tau = \frac{\gamma}{1+2\gamma}$ as specified. Our entire analysis will be on the event $\mathcal{F}_\tau$. Let us assume $\Lambda_s > 8D_\tau/\Delta$. This implies that there exists a cluster h such that the centroid estimation error satisfies

$$\|\widehat{\theta}_h^{(s)} - \theta_h\| > 8D_\tau.$$

As $H_s \geq \frac{1}{2}+\gamma$, we know that $n_{hh}^{(s)} \geq \left(\frac{1}{2}+\gamma\right) n_h^*$. As we are on the set $\mathcal{F}_\tau$ and

$$S = T_h^{(s)} \cap T_h^*, \quad |S| = n_{hh}^{(s)} \geq \left(\frac{1}{2}+\gamma\right) n_h^* \geq \left(\frac{1}{2}+\gamma\right) n\alpha$$

we get $|\{j \in S : \|w_j\| \leq D_\tau\}| \geq (1-\tau)|S|$. In view of $n_{hh}^{(s)} \geq \left(\frac{1}{2}+\gamma\right) n_h^{(s)}$ from $H_s \geq \frac{1}{2}+\gamma$ the above implies

$$\begin{aligned}
&|\{j \in S : \|w_j\| \leq D_\tau\}| \geq \frac{1+\gamma}{1+2\gamma}|S| \\
&\geq \left(\frac{1+\gamma}{1+2\gamma}\right)\left(\frac{1}{2}+\gamma\right) n_h^{(s)} \geq \left(\frac{1}{2}+\frac{\gamma}{2}\right) n_h^{(s)}.
\end{aligned}$$

This gives us

$$\left|\left\{j \in T_h^{(s)} : \|Y_j - \theta_h\| \leq D_\tau\right\}\right| \geq \left(\frac{1}{2}+\frac{\gamma}{2}\right) n_h^{(s)} \tag{26}$$

Next we will show

$$\begin{aligned}
&\left|\left\{j \in T_h^{(s)} : \|Y_j - \widehat{\theta}_h^{(s)}\| \leq 4D_\tau\right\}\right| \\
&\geq (1-\delta) n_h^{(s)} \geq \left(\frac{1}{2}+\frac{\gamma}{4}\right) n_h^{(s)}.
\end{aligned} \tag{27}$$

To prove the above, we first set

$$W = \left\{j \in T_h^{(s)} : \|Y_j - \theta_h\| \leq D_\tau\right\}.$$

Then given any $j_0 \in W$, all the points in $\{Y_j : j \in W\}$ are within $2D_\tau$ distance of Y_{j_0}. This implies

$$\left|\left\{j \in T_h^{(s)} : \|Y_j - Y_{j_0}\| \leq 2D_\tau\right\}\right| \geq \left(\frac{1}{2}+\frac{\gamma}{2}\right) n_h^{(s)}. \tag{28}$$

Now, remember the computation of $\mathsf{TM}_\delta(\{Y_j : j \in T_h^{(s)}\})$ in Algorithm 2 according to Algorithm 1. In view of (28) we then have $R_{i^*} \leq 2D_\tau$. Hence, for $\delta = \frac{1}{2} - \frac{\gamma}{4}$

$$\begin{aligned}
&\left|\left\{j \in T_h^{(s)} : \|Y_j - Y_{i^*}\| \leq 2D_\tau\right\}\right| \\
&\geq \left|\left\{j \in T_h^{(s)} : \|Y_j - Y_{i^*}\| \leq R_{i^*}\right\}\right| \geq \left(\frac{1}{2}+\frac{\gamma}{4}\right) n_h^{(s)}.
\end{aligned} \tag{29}$$

Then, the steps in Algorithm 1 imply for some $V \subset T_h^{(s)}$ with $|V| = \left(\frac{1}{2}+\frac{\gamma}{4}\right) n_h^{(s)}$

$$\begin{aligned}
&\|\widehat{\theta}_h^{(s)} - Y_j\| \leq \|\widehat{\theta}_h^{(s)} - Y_{i^*}\| + \|Y_j - Y_{i^*}\| \\
&\leq \frac{\sum_{j \in V} \|Y_j - Y_{i^*}\|}{\left\lfloor (1-\delta) n_h^{(s)} \right\rfloor + 1} + R_{i^*} \leq 2R_{i^*} \leq 4D_\tau, \quad j \in V.
\end{aligned}$$

This completes the proof of (27).

Finally, combining (26) and (27) we get a contradiction as

- $$\left|\left\{j \in T_h^{(s)} : \|Y_j - \widehat{\theta}_h^{(s)}\| \leq 4D_\tau\right\} \cup \left\{j \in T_h^{(s)} : \|Y_j - \theta_h\| \leq D_\tau\right\}\right| \leq \left|\{j \in T_h^{(s)}\}\right| = n_h^{(s)}$$
- $$\left|\left\{j \in T_h^{(s)} : \|Y_j - \widehat{\theta}_h^{(s)}\| \leq 4D_\tau\right\} \cap \left\{j \in T_h^{(s)} : \|Y_j - \theta_h\| \leq D_\tau\right\}\right| = 0$$

- $$\left|\left\{j \in T_h^{(s)} : \|Y_j - \widehat{\theta}_h^{(s)}\| \le 4D_\tau\right\}\right| + \left|\left\{j \in T_h^{(s)} : \|Y_j - \theta_h\| \le D_\tau\right\}\right| \ge \left(1 + \frac{3\gamma}{4}\right) n_h^{(s)}.$$

□

B. Proof of Theorem 2

In view of the above results we provide the proof of Theorem 2 below. Note that it suffices to bound the larger quantity $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{z_i \neq \widehat{z}_i\}}\right]$ to obtain a bound on $\mathbb{E}\left[\ell(\widehat{z}, z)\right]$, as the later contains a minimum over all possible label permutations. For an ease of notations, denote

$$A_s = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\left\{\widehat{z}_i^{(s)} \neq z_i\right\}} = \frac{1}{n}\sum_{h \neq g \in [k]} n_{hg}^{(s)}. \tag{30}$$

For $c_1 > 0$ to be chosen later, we define

$$\epsilon_0 = \frac{G^{-1}(e^{-c_1/\alpha})}{\Delta/(2\sigma)}, \quad \gamma_0 = \gamma.$$

Then from Lemma 16 it follows that we can choose $c_1, c_2 > 0$ such that if

$$\Delta \ge c_2 \sigma G^{-1}\left(\exp\left\{-\frac{1+2/\alpha}{\tau}\right\}\right), \tau = \frac{\gamma_0}{1+2\gamma_0},$$

then on the set $\mathcal{E}_{\epsilon_0,\gamma_0}$, for a large enough c_1, we have

- if $\Lambda_0 \le \frac{1}{2} - \epsilon_0$ then $H_1 \ge 0.8$,
- if $H_0 \ge \frac{1}{2} + \gamma_0$ then $\Lambda_0 \le 0.2$

A second application of Lemma 16, with $\epsilon_1 = 0.3, \gamma_1 = 0.3$, guarantees that if $\Delta \ge G^{-1}\left(\exp\left\{-\frac{c_3}{\alpha}\right\}\right)$ for a large enough c_3, then on the set $\mathcal{E}_{\epsilon_1,\gamma_1}$ we have for all $s \ge 1$,

(P1) If $\Lambda_s \le \frac{1}{2} - \epsilon_1$ then $H_{s+1} \ge 1 - \frac{5}{2\alpha \log(1/G(\epsilon_0\Delta/(2\sigma)))} \ge 0.8$,

(P2) If $H_s \ge \frac{1}{2} + \gamma_1$ then

$$\begin{aligned}\Lambda_s &\le \frac{8\sigma}{\Delta} G^{-1}\left(\exp\left\{-\frac{(1+2/\alpha)(1+2\gamma_1)}{\gamma_1}\right\}\right) \\ &\le \frac{G^{-1}(e^{-c_4/\alpha})}{\mathsf{SNR}} \le 0.2,\end{aligned}$$

where c_4 is an absolute constant. Note that from Lemma 16 the probabilities of each of the sets $\mathcal{E}_{\epsilon_0,\gamma_0}, \mathcal{E}_{\epsilon_1,\gamma_1}$ are at least $1 - 2k^{\frac{n\alpha}{4}} - e^{-0.3n}$, and hence

$$\mathcal{E} = \mathcal{E}_{\epsilon_0,\gamma_0} \cap \mathcal{E}_{\epsilon_1,\gamma_1} \text{ with } \mathbb{P}\left[\mathcal{E}\right] \ge 1 - 8ke^{-\frac{n\alpha}{4}}. \tag{31}$$

In view of the above arguments, on the event $\mathcal{E}$ we have

$$\Lambda_s \le \frac{G^{-1}(e^{-c_4/\alpha})}{\mathsf{SNR}} \le 0.2, \ H_s \ge 0.8 \text{ for all } s \ge 1. \tag{32}$$

Next, we will show that $\mathbb{P}\left[z_i \neq \widehat{z}_i^{(s+1)} \middle| \mathcal{E}\right]$ is small for each $i \in [n]$, and then sum over i to achieve the required result. Fix a choice for z_i, say equal to $g \in [k]$. Remember (23)

$$\mathbf{1}_{\left\{z_i = g, \widehat{z}_i^{(s+1)} = h\right\}} \le \mathbf{1}_{\left\{\frac{1-2\Lambda_s}{2(1+2\Lambda_s)}\|\theta_g - \theta_h\| \le \|w_i\|\right\}}. \tag{33}$$

Then in view of the inequalities

- $(1+x)^{-1} \ge 1 - x$ with $x = 2\Lambda_s < 1$
- $(1-x)^2 \ge 1 - 2x$ with the above choices of x,

and $\|\theta_g - \theta_h\| \ge \Delta$ we continue the last display to get

$$\begin{aligned}\mathbf{1}_{\left\{z_i = g, \widehat{z}_i^{(s+1)} = h\right\}} &\le \mathbf{1}_{\left\{\frac{1}{2}(1-4\Lambda_s)\Delta \le \|w_i\|\right\}} \\ &\le \mathbf{1}_{\left\{\sigma(\mathsf{SNR} - 4G^{-1}(e^{-c_4/\alpha})) \le \|w_i\|\right\}},\end{aligned}$$

where the last inequality followed using the bound on $\Lambda_s \le \frac{G^{-1}(e^{-c_4/\alpha})}{\mathsf{SNR}}$ in (32). Taking expectation conditioned on the event $\mathcal{E}$ in (31) and using the inequality

$$\mathbf{1}_{\left\{z_i \neq \widehat{z}_i^{(s+1)}\right\}} \le \sum_{\substack{g,h\in[k]\\ g\neq h}} \mathbf{1}_{\left\{\widehat{z}_i^{(s+1)}=h, z_i=g\right\}}$$

we get

$$\begin{aligned}\mathbb{P}\left[z_i \neq \widehat{z}_i^{(s+1)} \middle| \mathcal{E}\right] &\le k^2 \max_{\substack{g,h\in[k]\\ g\neq h}} \mathbb{P}\left[z_i = g, \widehat{z}_i^{(s+1)} = h \middle| \mathcal{E}\right] \\ &\le k^2 G\left(\mathsf{SNR} - 4G^{-1}(e^{-c_4/\alpha})\right)\end{aligned}$$

This implies

$$\begin{aligned}\mathbb{E}\left[A_{s+1}|\mathcal{E}\right] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left[z_i \neq \widehat{z}_i^{(s+1)} \middle| \mathcal{E}\right] \\ &\le k^2 G\left(\mathsf{SNR} - 4G^{-1}(e^{-c_4/\alpha})\right).\end{aligned}$$

Combining the above with (31) we get

$$\mathbb{E}\left[A_{s+1}\right] \le 8ke^{-\frac{n\alpha}{4}} + k^2 G\left(\mathsf{SNR} - 4G^{-1}(e^{-c_4/\alpha})\right).$$

APPENDIX B
PROOF OF RESULTS WITH OUTLIER (THEOREM 11)

A. Preparation

The following lemma provides results that show a lower bound on H_{s+1} based on Λ_s and establish upper bound on Λ_s in terms of H_s, when $n\alpha(1-\psi)$ outliers are present.

Lemma 18. *Fix any $\epsilon_0 \in (0, \frac{1}{2}), \gamma_0 \in (\frac{10}{n\alpha}, \frac{1}{2})$. Then, whenever $\Delta, \sigma > 0$ satisfies $\frac{5}{2\alpha\log(1/G(\epsilon_0\Delta/(2\sigma)))} < \frac{1}{2}$ the following holds true. There is an event $\mathcal{E}_{\gamma_0,\epsilon_0}$, which has a probability at least $1 - 4ke^{-\frac{n\alpha}{4}}$, on which we have for all $s \ge 0$, the COD_δ algorithm with $\delta \in (\frac{1}{2} - \frac{\gamma_0}{4}, \frac{1}{2})$ ensures:*

(i) if $\Lambda_s \le \frac{1}{2} - \epsilon_0$ then

$$H_{s+1} \ge \frac{1}{2} + \frac{\psi - 2\xi}{2(2-\psi)}, \quad \xi = \frac{5}{4\alpha\log(1/G(\epsilon_0\Delta/(2\sigma)))},$$

(ii) if $H_{s+1} \ge \frac{1}{2} + \gamma_0$ then

$$\Lambda_{s+1} \le \frac{8\sigma}{\Delta} G^{-1}\left(\exp\left\{-\frac{1+2/\alpha}{\tau}\right\}\right),$$

where $\tau = \frac{\gamma_0}{1+2\gamma_0}$.

Proof of Lemma 18. Repeating the argument in (23) in the proof of Lemma 16 we have

$$\mathbf{1}_{\left\{z_i=g, \widehat{z}_i^{(s+1)}=h\right\}} \le \mathbf{1}_{\{\epsilon_0\Delta/(2\sigma) \le \|w_i\|\}}.$$

Summing $\mathbf{1}_{\left\{z_i=g, \widehat{z}_i^{(s+1)}=h\right\}}$ over $\left\{i \in T_g^*\right\}$, in view of Lemma 14, we get on the event $\mathcal{E}_{\epsilon_0}^{\mathsf{con}}$, for all $h \in [k], h \neq g$

$$n_{gh}^{(s+1)} \le \sum_{i\in T_g^*} \mathbf{1}_{\left\{\frac{\epsilon_0}{2}\Delta \le \|w_i\|\right\}} \le \frac{5n_g^*}{4\log(1/G(\epsilon_0\Delta/\sigma))}. \tag{34}$$

Using the last display and noting that $k \le \frac{1}{\alpha}$ and $n_g^* \ge n\alpha$ we get

$$\begin{aligned}\frac{\sum_{\substack{h\in[k]\\ h\neq g}} n_{gh}^{(s+1)}}{n_g^*} &\le \frac{5k}{4\log(1/G(\epsilon_0\Delta/(2\sigma)))} \\ &\le \frac{5}{4\alpha\log(1/G(\epsilon_0\Delta/(2\sigma)))}.\end{aligned} \tag{35}$$

Next we switch g, h in (34) and sum over $h \in [k], h \neq g$. We get with probability similar to $\mathbb{P}\left[\mathcal{E}_{\epsilon_0}^{\mathsf{con}}\right]$

$$\sum_{\substack{h\in[k]\\h\neq g}} n_{hg}^{(s+1)} \leq \frac{5\sum_{h\in[k],h\neq g} n_h^*}{4\log(1/G(\epsilon_0\Delta/(2\sigma)))} \leq \frac{5n}{4\log(1/G(\epsilon_0\Delta/(2\sigma)))}.$$

We define $\xi = \frac{5}{4\alpha\log(1/G(\epsilon_0\Delta/(2\sigma)))}$. In view of (35) this implies

$$n_{gg}^{(s+1)} = n_g^* - \sum_{\substack{h\in[k]\\h\neq g}} n_{gh}^{(s+1)} \geq n_g^* - n\alpha\xi \geq n_g^*(1-\xi). \tag{36}$$

Using the above and noticing that in addition to the points in $\cup_{h\in[k]}\left\{T_h^* \cap T_g^{(s+1)}\right\}$, $T_g^{(s+1)}$ can at most have $n\alpha(1-\psi)$ many extra points, accounting for the outliers, we get

$$\frac{n_{gg}^{(s+1)}}{n_g^{(s+1)}} \geq \frac{n_{gg}^{(s+1)}}{n_{gg}^{(s+1)} + n\alpha\xi + n\alpha(1-\psi)} \geq \frac{1}{1+\frac{n_g^*(1-\psi+\xi)}{n_{gg}^{(s+1)}}} \geq \frac{1}{1+\frac{1-\psi+\xi}{1-\xi}} = \frac{1}{2} + \frac{\psi-2\xi}{2(2-\psi)}.$$

Combining the last display with (36) we get

$$H_{s+1} = \min_{g\in[k]}\left\{\min\left\{\frac{n_{gg}^{(s)}}{n_g^*}, \frac{n_{gg}^{(s)}}{n_g^{(s)}}\right\}\right\} \geq \frac{1}{2} + \min\left\{\frac{\psi-2\xi}{2(2-\psi)}, \frac{1}{2}-\xi\right\}.$$

As $\psi < 1$, we get $\frac{\psi-2\xi}{2(2-\psi)} \leq \frac{\psi}{2} - \xi \leq \frac{1}{2} - \xi$. This finishes the proof.

The proof of Lemma 18(ii) is similar to the proof of Lemma 16(ii). □

B. *Proof of Theorem 11*

For $c_1 > 0$ to be chosen later, we define

$$\epsilon_0 = \frac{G^{-1}\left(e^{-\frac{c_1}{\alpha\psi}}\right)}{\Delta/(2\sigma)}, \quad \gamma_0 = \gamma.$$

Then from Lemma 18 it follows that we can choose $c_2 > 0$ such that if $\tau = \frac{\gamma_0}{1+2\gamma_0}$ and

$$\Delta \geq c_2\sigma\max\left\{G^{-1}\left(\exp\left\{-\frac{1+2/\alpha}{\tau}\right\}\right), G^{-1}\left(e^{-\frac{c_1}{\alpha\psi}}\right)\right\},$$

then on the event $\mathcal{E}_{\epsilon_0,\gamma_0}$, as $\delta = \frac{1}{2} - \min\left\{\frac{\gamma_0}{4}, \delta/24\right\}$, we have that the COD_δ algorithm guarantees

- if $\Lambda_0 \leq \frac{1}{2} - \epsilon_0$ then $H_1 \geq \frac{1}{2} + \frac{\psi}{6}$.
- if $H_0 \geq \frac{1}{2} + \gamma_0$ then $\Lambda_0 \leq 0.3$,

A second application of Lemma 18, with $\epsilon_1 = 0.2, \gamma_1 = \frac{\psi}{6}$ and the above lower bound on Δ for large enough c_1, c_2, implies that the COD_δ algorithm guarantees on the event $\mathcal{E}_{\epsilon_1,\gamma_1}$ for all $s \geq 1$,

(P1) if $\Lambda_s \leq \frac{1}{2} - \epsilon_1$ then $H_{s+1} \geq \frac{1}{2} + \frac{\psi}{6}$,

(P2) if $H_s \geq \frac{1}{2} + \gamma_1$ then

$$\Lambda_s \leq \frac{8\sigma}{\Delta}G^{-1}\left(\exp\left\{-\frac{(1+2/\alpha)(1+2\gamma_1)}{\gamma_1}\right\}\right) \leq \frac{G^{-1}(e^{-c_4/(\alpha\psi)})}{\mathsf{SNR}} \leq 0.2,$$

where c_4 is an absolute constant. Combining the above displays we get that on the event

$$\mathcal{E} = \mathcal{E}_{\epsilon_0,\gamma_0} \cap \mathcal{E}_{\epsilon_1,\gamma_1} \text{ with } \mathbb{P}[\mathcal{E}] \geq 1 - 8ke^{-\frac{n\alpha}{4}} \tag{37}$$

we have

$$\Lambda_s \le \frac{G^{-1}(e^{-c_4/(\alpha\psi)})}{\mathsf{SNR}} \le 0.2,\ \ H_s \ge \frac{1}{2}+\frac{\psi}{6} \quad \text{for all } s \ge 2. \tag{38}$$

Now, it is sufficient to show that $\mathbb{P}\left[z_i \neq \widehat{z}_i^{(s+1)} \middle| \mathcal{E}\right]$ is small for each $i \in [n]$. This will imply that on the event $\mathcal{E}$, $\ell(\widehat{z}^{(s+1)}, z) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\left\{z_i \neq \widehat{z}_i^{(s+1)}\right\}}$ is also small in probability. In view of the above, using a Markov inequality we will conclude the result. From this point onward the proof is again similar to the proof in Theorem 2 for showing that $\mathbb{P}\left[z_i \neq \widehat{z}_i^{(s+1)} \middle| \mathcal{E}\right]$ is small. The only difference is that we replace the term $G^{-1}(e^{-c_4/\alpha})$ by $G^{-1}(e^{-c_4/(\alpha\psi)})$. This finishes the proof.

Appendix C
Proof of mislabeling lower bound (Theorem 3)

We will consider a smaller set of labels to perform the analysis. For a simplicity of notations, let $m = k\lceil n\alpha\rceil \le k(n\alpha+1)$ and $n-m$ is divisible by 3. Define

$$\begin{aligned}
\mathcal{Z}^* &= \bar{z} \times \{1,2\}^{\frac{n-m}{3}} \subseteq [k]^n,\\
m = k\lceil n\alpha\rceil, &\quad \bar{z} \in [k]^{m+\frac{2(n-m)}{3}},\\
\bar{z}_i = u, u \in [k],\ i &\in (u-1)\frac{m}{k}+1, \ldots, u\frac{m}{k},\\
\bar{z}_{m+1} &= \cdots = \bar{z}_{m+\frac{n-m}{3}} = 1,\\
\bar{z}_{m+\frac{n-m}{3}+1} &= \cdots = \bar{z}_{m+\frac{2(n-m)}{3}} = 2.
\end{aligned} \tag{39}$$

In other words, for each $z \in \mathcal{Z}^*$, we already know the labels corresponding to the first $m+\frac{2(n-m)}{3}$ entries. For the rest of the entries the labels can either be 1 or 2, and we will put a prior on the distribution of those labels. Note that for each label vector $z \in \mathcal{Z}^*$ we have $|\{i : z_i = g\}| \ge \lceil n\alpha\rceil$ for each $g = 1, \ldots, k$. Under the above choice of label set, it suffices to show

$$\begin{aligned}
&\inf_{\widehat{z}\in\mathcal{Z}^*} \sup_{(z,\{\theta_i\}_{i=1}^k)\in\mathcal{P}_0 : z\in\mathcal{Z}^*} \mathbb{E}\left[\ell(\widehat{z}, z)\right]\\
&\ge \frac{1-k\alpha-k/n}{12}\cdot G(\mathsf{SNR}+C_G), \quad \Delta \ge \sigma C_G,
\end{aligned}$$

for a suitable choice of C_G.

To show the above, we first note that for any $z, z' \in \mathcal{Z}^*, z \neq z'$, we have

$$\sum_{i=1}^n \mathbf{1}_{\{z_i \neq z_i'\}} = \sum_{i=m+\frac{2(n-m)}{3}+1}^n \mathbf{1}_{\{z_i\neq z_i'\}} \le \frac{n-m}{3},$$

and for any $\pi \in \mathcal{S}_k$ that is not an identity map, $\sum_{i=1}^n \mathbf{1}_{\{\pi(z_i)\neq z_i'\}}$ is bounded by

- $2\lceil n\alpha\rceil + \sum_{i=m+\frac{2(n-m)}{3}+1}^n \mathbf{1}_{\{z_i\neq z_i'\}}$ if $\pi(1)=1, \pi(2)=2$,
- $\frac{n-m}{3}$, if $\pi(1)\neq 1$ or $\pi(2)\neq 2$.

This implies the loss $\mathbb{E}\left[\ell(\widehat{z}, z)\right]$ reduces to $\mathbb{E}\left[\sum_{i=1}^n \frac{1}{n}\mathbf{1}_{\{z_i\neq z_i'\}}\right]$. In view of the above deductions, with the parameter (label) set $\mathcal{Z}^*$ we will apply Assouad's Lemma [57] to bound

$$\inf_{\widehat{z}\in\mathcal{Z}^*} \sup_{(z,\{\theta_i\}_{i=1}^k)\in\mathcal{P}_0 : z\in\mathcal{Z}^*} \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n \mathbf{1}_{\{z_i\neq z_i'\}}\right].$$

Lemma 19 (Assouad)**.** *Let $r \ge 1$ be an integer and $\mathcal{F}_r = \{\mathcal{Q}_z : z \in \mathcal{Z}\}$ contains 2^r probability measures. Let $\widehat{f}$ be an estimator of $f(\mathcal{Q}_z)$ taking values in a metric space $(\mathcal{D}, d)$, for some $\mathcal{Q}_z \in \mathcal{F}_r$. Write $v \sim v'$ if v and v' differ in only one coordinate, and write $v \sim_j v'$ when the coordinate is the j-th. Suppose that there are r pseudo-distances $d_1, \ldots, d_r$ on $\mathcal{D}$ such that for any $x, y \in \mathcal{D}$*

$$d(x,y) = \sum_{j=1}^r d_j(x,y),$$

and further assume that, if $v \sim_j v'$ then $d_j(f(\mathcal{Q}_z), f(\mathcal{Q}_z')) \ge \delta$. Then

$$\begin{aligned}
&\max_{\mathcal{Q}_z\in\mathcal{F}_r} \mathbb{E}_z\left[d(\widehat{f}, f(\mathcal{Q}_z))\right]\\
&\ge r\cdot\frac{\delta}{2}\cdot\min\{1-\mathrm{TV}(\mathcal{Q}_z, \mathcal{Q}_{z'}) : z \sim z'\}.
\end{aligned}$$

To apply the above lemma, define the data distribution $\mathcal{Q}_z$ given any label vector $z \in \mathcal{Z}^*$

$$\begin{aligned} z \in \mathcal{Z}^*, \quad \mathcal{Q}_{z_i} &= \text{Distribution of } \theta_{z_i} + w_i,\ i = 1, \ldots, n. \\ \overline{\mathcal{Q}} &= \mathcal{Q}_{\bar{z}_1} \times \cdots \times \mathcal{Q}_{\bar{z}_{m+\frac{2(n-m)}{3}}}, \\ Q_z &= \bar{\mathcal{Q}} \times \mathcal{Q}_{z_{m+\frac{2(n-m)}{3}+1}} \times \cdots \times \mathcal{Q}_{z_n}. \end{aligned}$$

In view of the above definition, to apply Lemma 19, we choose

$$\begin{aligned} \mathcal{Z} = \mathcal{Z}^*, \quad r = \frac{n-m}{3}, \quad \delta = 1 \\ f(\mathcal{Q}_z) = z, \quad d_j(z, z') = \mathbf{1}_{\left\{ z_{m+\frac{2(n-m)}{3}+j} \neq z'_{m+\frac{2(n-m)}{3}+j} \right\}}. \end{aligned}$$

Hence, using Lemma 19 we get that given any estimator $\widehat{z}$ (which we can choose to be in $\mathcal{Z}^*$), it satisfies

$$\begin{aligned} &\max_{\mathcal{Q}_z \in \mathcal{F}_r} \mathbb{E}_z \left[\sum_{i=1}^{n} \mathbf{1}_{\{\widehat{z}_i \neq z_i\}} \right] \\ &\geq \frac{n-m}{6} \min\{1 - \mathrm{TV}(\mathcal{Q}_z, \mathcal{Q}_{z'}) : z \sim z' \in \mathcal{Z}^*\}, \end{aligned} \tag{40}$$

We further specify the error distributions corresponding to the labels $\{z_{m+\frac{2(n-m)}{3}+1}, \ldots, z_n\}$ based on the decay function G. As G is already differentiable on $(\sigma c_G, \infty)$ we can extend G on $(0, \sigma c_G]$ such that it is differentiable throughout with $G(0) = 1, G'(0) = 0$. Then $1 - G(\cdot)$ is a distribution function with a density $-G'$. We define

$$\begin{aligned} \{w_i\}_{i=m+\frac{2(n-m)}{3}+1}^{n} &\overset{\text{iid}}{\sim} R \cdot V, \\ \mathbb{P}[R \geq x] &= 1 - G(x/\sigma), \\ \mathbb{P}\left[V = \frac{\theta_1 - \theta_2}{\|\theta_1 - \theta_2\|}\right] &= \frac{1}{2} = \mathbb{P}\left[V = \frac{\theta_2 - \theta_1}{\|\theta_1 - \theta_2\|}\right]. \end{aligned} \tag{41}$$

In view of the above we can simplify (40) as

$$\max_{\mathcal{Q}_z \in \mathcal{F}_r} \mathbb{E}_z \left[\sum_{i=1}^{n} \mathbf{1}_{\{\widehat{z}_i \neq z\}} \right] \geq \frac{n-m}{6} (1 - \mathrm{TV}(P_1, P_2)), \tag{42}$$

where P_i denotes the distribution of $\theta_i + R \cdot V$ for $i = 1, 2$. To analyze the total variation term in the above formula, we first note that without a loss of generality we can assume that θ_1, θ_2 lie on the real line with $\theta_1 = -\frac{\Delta}{2}, \theta_2 = \frac{\Delta}{2}$. This is because the total variation distance is invariant under location shifts and rotational transformations. Then we simplify the distributions in (41) in terms of the density of w_i-s as

$$f_{w_i}(x) = -\frac{1}{2\sigma} G'(|x|/\sigma), \quad i = m + \frac{2(n-m)}{3} + 1, \ldots, n.$$

Hence, using a location shift argument, we get the densities of P_1, P_2 on $(-\infty, \infty)$ as

$$\begin{aligned} dP_1(y) &= -\frac{1}{2\sigma} G'\left(\frac{|y + \Delta/2|}{\sigma}\right) dy, \\ dP_2(y) &= -\frac{1}{2\sigma} G'\left(\frac{|y - \Delta/2|}{\sigma}\right) dy. \end{aligned} \tag{43}$$

Then the total variation distance between P_1, P_2 can be bounded as

$$\begin{aligned}
&\mathrm{TV}(P_1, P_2) \\
&= \frac{1}{2}\int_{-\infty}^{\infty} |dP_1(y) - dP_2(y)| \\
&= \frac{1}{4\sigma}\int_{-\infty}^{\infty} \left| G'\left(\frac{\left|y+\frac{\Delta}{2}\right|}{\sigma}\right) - G'\left(\frac{\left|y-\frac{\Delta}{2}\right|}{\sigma}\right)\right| dy \\
&\overset{(a)}{=} \frac{1}{2\sigma}\int_{-\infty}^{0} \left| G'\left(\frac{\left|y+\frac{\Delta}{2}\right|}{\sigma}\right) - G'\left(\frac{\left|y-\frac{\Delta}{2}\right|}{\sigma}\right)\right| dy \\
&= \frac{1}{2\sigma}\left(\int_{-\infty}^{-\frac{\Delta}{2}-\sigma c_G} + \int_{-\frac{\Delta}{2}-\sigma c_G}^{-\frac{\Delta}{2}+\sigma c_G} + \int_{-\frac{\Delta}{2}+\sigma c_G}^{0}\right) \\
&\quad \left| -G'\left(\frac{\left|y+\frac{\Delta}{2}\right|}{\sigma}\right) - \left(-G'\left(\frac{\left|y-\frac{\Delta}{2}\right|}{\sigma}\right)\right)\right| dy \\
&\overset{(b)}{\leq} -\frac{1}{2\sigma}\left(\int_{-\infty}^{-\frac{\Delta}{2}-\sigma c_G} + \int_{-\frac{\Delta}{2}+\sigma c_G}^{0}\right) G'\left(\frac{\left|y+\frac{\Delta}{2}\right|}{\sigma}\right) dy \\
&\quad - \frac{1}{2\sigma}\int_{-\frac{\Delta}{2}-\sigma c_G}^{-\frac{\Delta}{2}+\sigma c_G} \left(G'\left(\frac{\left|y+\frac{\Delta}{2}\right|}{\sigma}\right) + G'\left(\frac{\left|y-\frac{\Delta}{2}\right|}{\sigma}\right)\right) dy
\end{aligned}$$

where $c_G > 0$ is as prescribed in (Q) and

- (a) followed as $\left|G'\left(\frac{\left|y+\frac{\Delta}{2}\right|}{\sigma}\right) - G'\left(\frac{\left|y-\frac{\Delta}{2}\right|}{\sigma}\right)\right|$, as a function of y, is symmetric about 0
- (b) followed as $G'(y)$ is negative for all $y > 0$ and we allow $\Delta \geq 2\sigma c_G$ implies for $y \in (-\infty, -\frac{\Delta}{2} - \sigma c_G) \cup (-\frac{\Delta}{2} + \sigma c_G, 0)$

$$-G'\left(\frac{\left|y+\frac{\Delta}{2}\right|}{\sigma}\right) \geq -G'\left(\frac{\left|y-\frac{\Delta}{2}\right|}{\sigma}\right) \geq 0$$

We continue the last inequality on $\mathrm{TV}(P_1, P_2)$ to get

$$\begin{aligned}
&\mathrm{TV}(P_1, P_2) \\
&\leq -\int_{-\infty}^{0} \frac{1}{2\sigma} G'\left(\frac{\left|y+\frac{\Delta}{2}\right|}{\sigma}\right) dy \\
&\quad - \int_{-\frac{\Delta}{2}-\sigma c_G}^{-\frac{\Delta}{2}+\sigma c_G} \frac{1}{2\sigma} G'\left(\frac{\left|y-\frac{\Delta}{2}\right|}{\sigma}\right) dy \\
&\overset{(a)}{=} 1 + \int_{0}^{\infty} \frac{1}{2\sigma} G'\left(\frac{y+\frac{\Delta}{2}}{\sigma}\right) dy \\
&\quad - \int_{-\frac{\Delta}{2}-\sigma c_G}^{-\frac{\Delta}{2}+\sigma c_G} \frac{1}{2\sigma} G'\left(\frac{-\left(y-\frac{\Delta}{2}\right)}{\sigma}\right) dy \\
&\overset{(b)}{=} 1 + \int_{\Delta/2}^{\infty} \frac{1}{2\sigma} G'\left(\frac{z}{\sigma}\right) dz - \int_{\Delta-\sigma c_G}^{\Delta+\sigma c_G} \frac{1}{2\sigma} G'\left(\frac{z}{\sigma}\right) dz \\
&\overset{(c)}{\leq} 1 + \int_{\Delta/2}^{\infty} \frac{1}{2\sigma} G'\left(\frac{z}{\sigma}\right) dz - \int_{\Delta/2}^{\Delta/2+2\sigma c_G} \frac{1}{2\sigma} G'\left(\frac{z}{\sigma}\right) dz \\
&= 1 - \frac{1}{2} G\left(\frac{\Delta}{2\sigma} + 2c_G\right),
\end{aligned}$$

where

- (a) followed as $-\frac{1}{2\sigma} G'\left(\frac{\left|y+\frac{\Delta}{2}\right|}{\sigma}\right)$ is a density on $(-\infty, \infty)$ from (43)
- (b) followed by change of variables
- (c) followed as $-G'(y) = |G'(y)|$ is decreasing over (c_G, ∞) and $\Delta \geq 2\sigma c_G$.

Plugging the last display in (42) completes the proof.

APPENDIX D
TECHNICAL RESULTS

Proof of Lemma 15. It suffices to show that for any $S \subseteq [n]$

$$\mathbb{P}\left[\sum_{i\in S} \mathbf{1}_{\left\{\|w_i\| > \sigma G^{-1}\left(\exp\left\{-\frac{1+1/\beta}{\tau}\right\}\right)\right\}} \geq \tau|S|\right] \leq e^{-n},$$

and then use the union bound over different choices of S to get the result. We define

$$V_i = \mathbf{1}_{\left\{\|w_i\| > \sigma G^{-1}\left(\exp\left\{-\frac{1+1/\beta}{\tau}\right\}\right)\right\}}.$$

In view of the above definitions it is enough to show that

$$\mathbb{P}\left[\sum_{i\in S} V_i \geq \tau|S|\right] \leq e^{-n} \text{ for all } S \subseteq [n]. \tag{44}$$

Note that

$$\begin{aligned}&\mathbb{P}\left[\|w_i\| > \sigma G^{-1}\left(\exp\left\{-\frac{1+1/\beta}{\tau}\right\}\right)\right]\\ &\leq \exp\left\{-\frac{1+1/\beta}{\tau}\right\} \leq \frac{\tau}{1+1/\beta} < \tau.\end{aligned}$$

We note the following result on stochastic dominance.

Lemma 20. *[58, Chapter 4.2] Given two random variables $X, Y \in \mathbb{R}$, suppose that we call X to be stochastically smaller than Y if $\mathbb{P}[X \geq a] \leq \mathbb{P}[Y \geq a], \quad \forall a \in \mathbb{R}$. Then, for $X \sim$ Binom$(n,p), Y \sim$ Binom(n,q) for $0 < p < q < 1$, we get that X is stochastically smaller than Y.*

The above result implies that $\sum_{i\in S} V_i$ is stochastically smaller than any $Z \sim \text{Binom}(|S|, \exp\{-\frac{1+1/\beta}{\tau}\})$. We continue to analyze (44) using Chernoff's inequality for the Binomial random variable given below

Lemma 21. *[59, Section 2.2] For a random variable $Z \sim$ Binom(m,q), we have*

$$\begin{aligned}\mathbb{P}\left[Z \geq ma\right] &\leq \exp\left(-mh_q(a)\right); \quad q < a < 1,\\ h_q(a) &= a\log\frac{a}{q} + (1-a)\log\frac{1-a}{1-q}.\end{aligned}$$

Using

$$Z \sim \text{Binom}(|S|, q), q = \exp\left\{-\frac{1+1/\beta}{\tau}\right\}, a = \tau, m = |S|$$

in the above lemma and $x\log x \geq -0.5$ for $x \in (0,1)$ we get

$$\begin{aligned}&\mathbb{P}\left[\sum_{i\in S} V_i \geq \tau|S|\right]\\ &\leq \mathbb{P}\left[Z \geq \tau|S|\right]\\ &\leq \exp\left(-m\left(\tau\log\frac{\tau}{q} + (1-\tau)\log\frac{1-\tau}{1-q}\right)\right)\\ &\leq \exp\left(-m\left\{(1+1/\beta) + \tau\log\tau + (1-\tau)\log(1-\tau)\right\}\right)\\ &\leq e^{-n}.\end{aligned}$$

Finally taking an union bound over all choices of S we get the desired bound. □

APPENDIX E
PROOF OF THE TWO CLUSTER INITIALIZATION RESULT (THEOREM 6)

A. Preparation

For the proofs involving the initialization results, let $\mathcal{B}(x,R)$ denote the Euclidean ball of radius R around the point x. Our proofs rely on the following high probability guarantees.

Lemma 22. *There is an event $\widetilde{\mathcal{E}}$ with $\mathbb{P}[\widetilde{\mathcal{E}}] \geq 1 - 4e^{-\frac{\min_{g=1,2} n_g^*}{4}}$ on which the following statements hold for the 2-cluster problem, for any given $\beta \in (0,1)$:*

(i) $\left|\mathcal{B}(\theta_i, \sigma G^{-1}(e^{-\frac{5}{4\beta^2}})) \cap \{Y_i : i \in T_i^*\}\right| \geq n_i^*(1-\beta^2)$ *for each* $i = 1, 2$,

(ii) $\left|\mathcal{B}(\theta_i, \frac{\Delta}{32}) \cap \{Y_i : i \in T_i^*\}\right| \geq n_i^* \left(1 - \frac{5}{4\log(1/G(\Delta/(32\sigma)))}\right)$ *for each* $i = 1, 2$.

Proof. The proof of part (i) follows by choosing $\epsilon_0 = \frac{\sigma}{\Delta} G^{-1}(e^{-\frac{5}{4\beta^2}})$ in Lemma 14. The proof of part (ii) follows by choosing by choosing $\epsilon_0 = \frac{1}{32}$ in Lemma 14. □

B. *Proof of Theorem 6*

In the proof below, we assume that all the mentioned constants depend on G, α, σ, unless otherwise specified. We will provide a proof of the result involving the outliers, as in Theorem 12 as the proof in absence of the outliers is very similar. In summary, we show the following:

Suppose that out of the n many observed data points, n_i^* many are from cluster $T_i^*, i = 1, 2$ and n^{out} many are adversarial outliers (i.e., $n_1^* + n_2^* + n^{\mathsf{out}} = n$). Also, assume that for some constant $\alpha > 0$ the counts satisfy $n_1^*, n_2^* > n\alpha$, $n^{\mathsf{out}} \leq \frac{n\alpha^2}{32}$. Then the following holds with probability at least $1 - 4e^{-n\alpha/4}$. There are constants $c_1, c_2 > 0$ such that if $\Delta \geq c_1 \sigma G^{-1}\left(e^{-\frac{c_2}{\alpha^2}}\right)$ then the centroid outputs from $\mathsf{IOD}_{2, m_1, m, \beta}$ *algorithm satisfy* $\max_{i=1,2} \|\theta_{\pi(i)} - \mu_i^*\| \leq \Delta/3$ *for a permutation* π *on* $\{1, 2\}$ *with*

$$m_1 = \left\lceil \frac{n\alpha}{4} \right\rceil, m = \max\{1, \left\lfloor \frac{n\alpha^2}{16} \right\rfloor\}, \beta = \frac{\alpha}{4}.$$

We prove the above statement here. For our entire analysis we will assume that the event $\widetilde{\mathcal{E}}$ in Lemma 22 holds, which has a high probability guarantee. We will extensively use the following definition of order statistics: Given any set V of real numbers and fraction $0 < p < 1$, define $V^{\{p\}}$ as the $\lceil p|V| \rceil$-th smallest number in V. Then, the proof is a combination of the following results.

Lemma 23. *There is one θ_i, such that*

$$\|\theta_i - \mu_1^{(1)}\| \leq 3\sigma G^{-1}(e^{-\frac{5}{4\beta^2}})$$

Lemma 24. *There is a stage $\ell + 1$, with $\ell \geq 1$, such that* $\mathsf{dist}_1^{(\ell+1)} > \frac{\Delta}{16}$.

Lemma 25. *Suppose that* $\ell = \min\left\{r \geq 1 : \mathsf{dist}_1^{(r+1)} > \frac{\Delta}{16}\right\}$. *Then* $\mathsf{totdist}_\ell \leq \Delta/8$.

Lemma 26. *If* $\mathsf{totdist}_\ell \leq \frac{\Delta}{8}$, *then there is a permutation π of $\{1, 2\}$ such that*

$$\max_{i=1,2} \|\mu_i^{(\ell)} - \theta_{\pi(i)}\| \leq \frac{\Delta}{3}.$$

Lemma 23, Lemma 24 and Lemma 25 together implies, provided Δ is large enough, that among all of the iterations of our algorithm there is an instance on which the $\mathsf{totdist}_\ell$ measure becomes smaller than $\frac{\Delta}{8}$. As our algorithm finally picks the iteration step $\ell = \ell^*$ with the lowest $\mathsf{totdist}_\ell$ measure, it ensures that $\mathsf{totdist}_{\ell^*} \leq \frac{\Delta}{8}$. In view of Lemma 26 this implies $\max_{i=1,2} \|\theta_{\pi(i)} - \mu_i^*\| \leq \Delta/3$ as required. Below we provide the proofs.

Proof of Lemma 23. In view of Lemma 22, there is a constant $c_1 = \sigma G^{-1}(e^{-\frac{5}{4\beta^2}})$ such that

$$|\{j \in [n] : Y_j \in \mathcal{B}(\theta_i, c_1)\}| \geq n_i^* \left(1 - \beta^2\right), \quad i \in \{1, 2\}. \tag{45}$$

As we have $n_1^*, n_2^* > n\alpha$ by assumption, it follows that there is a point Y_i such that

$$|\{j \in [n] : Y_j \in \mathcal{B}(Y_i, 2c_1)\}| \geq m_1 \geq \frac{n\alpha}{4}.$$

Hence, the tightest neighborhood around any point $Y_i, i \in [n]$, that contains at least $n\alpha/4$ points from $Y_1, \ldots, Y_n$, has a radius of at most $2c_1$ around that Y_i. Define

$$D(x, S) = \{\|x - Y_i\| : i \in S\}, \quad x \in \mathbb{R}^d, S \subseteq [n]. \tag{46}$$

Let i^* be one such index in $[n]$ that satisfies

$$D(Y_{i^*}, [n])^{\{\frac{m_1}{n}\}} = \min_{j \in [n]} D(Y_j, [n])^{\{\frac{m_1}{n}\}}. \tag{47}$$

Then $\mathcal{B}(Y_{i^*}, 2c_1)$ and $\cup_{i=1,2}\mathcal{B}(\theta_i, c_1)$ can not be disjoint, as in view of (45) the disjointedness will imply that their union will contain more than n points from $Y_1, \ldots, Y_n$

$$\begin{aligned}
&\Big|\{i \in [n] : Y_i \in \mathcal{B}(Y_{i^*}, 2c_1)\} \\
&\cup \left[\cup_{j=1,2}\left\{i \in [n] : Y_i \in \mathcal{B}(\theta_j, c_1)\right\}\right]\Big| \\
&\geq m_1 + \sum_{i=1,2} n_i^* \left(1-\beta^2\right) = \frac{n\alpha}{4} + (n - n^{\mathsf{out}})(1-\beta^2) \\
&\geq n + \frac{n\alpha}{4} - n\beta^2 - n^{\mathsf{out}} \geq n + \frac{n\alpha^2}{8},
\end{aligned}$$

where we use the fact that $\{i \in [n] : Y_i \in \mathcal{B}(\theta_i, c_1)\}, i = 1, 2$ are disjoint sets as $\|\theta_1 - \theta_2\| \geq \Delta, \beta = \frac{\alpha}{4}, n_1^* + n_2^* + n^{\mathsf{out}} = n$ and $n^{\mathsf{out}} \leq \frac{n\alpha^2}{16}$. Hence, Y_{i^*} is at a distance at most $3c_1$ from one of the true centroids θ_1, θ_2. Without a loss of generality we can pick $\mu_1^{(1)} = Y_{i^*}$ and we assume that $\mu_1^{(1)}$ is closer to θ_1 than θ_2. □

Proof of Lemma 24. In view of the proof of Lemma 23, let us assume that θ_1 is the closest centroid to $\mu_1^{(1)}$ and define $c_1 = \sigma G^{-1}(e^{-\frac{5}{4\beta^2}})$ as before to have for $i \in \{1, 2\}$

$$\mu_1^{(1)} \in \mathcal{B}\left(\theta_1, 3c_1\right), \quad \left|\{Y_j : j \in T_i^*\} \cap \mathcal{B}\left(\theta_i, c_1\right)\right| \geq n_i^*\left(1-\beta^2\right). \tag{48}$$

We observe the following:

- In view of $\mathcal{B}(\mu_1^{(1)}, 4c_1) \supset \mathcal{B}(\theta_1, c_1)$ we get

  $$\begin{aligned}
  &\left|\{Y_i : i \in [n]\} \cap \mathcal{B}(\mu_1^{(1)}, 4c_1)\right| \geq \left|\{Y_i : i \in [n]\} \cap \mathcal{B}(\theta_1, c_1)\right| \\
  &\geq n\alpha(1-\beta^2) \geq \frac{n\alpha}{2}.
  \end{aligned}$$

  As the size of $\mathcal{P}_1^{(1)}$ is at most $m_1 = \lceil \frac{n\alpha}{4} \rceil$, the distance of $\mu_1^{(1)}$ to any point in $\mathcal{P}_1^{(1)}$ is less than $4c_1$. As $\Delta \geq 64c_1$, the last statement implies $\mathsf{dist}_1^{(1)} \leq \frac{\Delta}{16}$.
- At the last step, say $\widetilde{\ell}$, in our algorithm, $\mathcal{P}_1^{(\widetilde{\ell})}$ will have at least $n - m = n - \frac{n\alpha^2}{16}$ many points. In view of (48) we also have

  $$\begin{aligned}
  &\left|\left\{Y_j : j \in [n], Y_j \in \mathcal{B}(\theta_2, c_1)\right\} \cap \mathcal{P}_1^{(\widetilde{\ell})}\right| \\
  &\geq \left|\left\{Y_j : j \in [n], Y_j \in \mathcal{B}(\theta_2, c_1)\right\}\right| - \overline{|\mathcal{P}_1^{(\widetilde{\ell})}|} \\
  &\geq n_2^*\left(1-\beta^2\right) - \frac{n\alpha^2}{16} \geq n\alpha - \frac{n\alpha^2}{16} - n\alpha\beta^2.
  \end{aligned}$$

As we have

- the tightest neighborhood in the data set around $\mu_1^{(1)}$ with a size at least $(1-\beta)|\mathcal{P}_1^{(\widetilde{\ell})}|$, say, N, will include at least $(1-\beta)(n - \frac{n\alpha^2}{16}) \geq n - n\beta - \frac{n\alpha^2}{16}$ points, and
- (48) implies that $\{Y_j : j \in T_2^*\} \cap \mathcal{B}\left(\theta_2, c_1\right)$ will contain at least $\frac{n\alpha}{2}$ points

we get that

$$|N| + \left|\{Y_j : j \in T_2^*\} \cap \mathcal{B}\left(\theta_2, c_1\right)\right| \geq n + \frac{n\alpha}{4}.$$

This implies $N \cap \{Y_j : j \in T_2^*, \mathcal{B}\left(\theta_2, c_1\right)\}$ is nonempty. Suppose that y is an element in the above set. Then we have that the distance of y from $\mu_1^{(\widetilde{\ell})}$ is at least $\Delta - 4c_1$,

$$\|\mu_1^{(\widetilde{\ell})} - y\| \geq \|\theta_1 - \theta_2\| - \|\mu_1^{(\widetilde{\ell})} - \theta_1\| - \|\theta_2 - y\| \geq \Delta - 4c_1.$$

Hence we get $\mathsf{dist}_1^{(\widetilde{\ell})} \geq \Delta - 4c_1$. As $\Delta > 64c_1$ we get $\mathsf{dist}_1^{(\widetilde{\ell})} \geq \frac{\Delta}{2}$ as required. □

Proof of Lemma 25. In view of Lemma 23, without a loss of generality we assume that θ_1 is the closest centroid to $\mu_1^{(1)}$ and (48). We first prove the following claims:

$$\left|\mathcal{P}_1^{(\ell)} \cap \{Y_i : i \in T_1^*\}\right| \geq n_1^* - m - \frac{5n}{4\log(1/(G(\Delta/32\sigma)))} \tag{49}$$

$$\left|\mathcal{P}_1^{(\ell)} \cap \{Y_i : i \in T_2^*\}\right| \leq n\beta + \frac{5n}{4\log(1/(G(\Delta/32\sigma)))}. \tag{50}$$

The first claim (49) follows from the following sequence of arguments

- Note that $\mu_1^{(\ell)} = \mu_1^{(\ell+1)}$ from the description in Algorithm 4. This implies

$$\mathsf{dist}_1^{(\ell+1)} = \left\{ D(\mu_1^{(\ell)}, \mathcal{P}_1^{(\ell+1)}) \right\}^{\{1-\beta\}} > \frac{\Delta}{16}.$$

As $\mathcal{P}_1^{(\ell+1)}$ is constructed by including the data points according to increasing Euclidean distances from $\mu_1^{(\ell)} = \mu_1^{(\ell+1)}$ we get

$$\mathcal{P}_1^{(\ell+1)} \supseteq \{Y_i : i \in [n]\} \cap \mathcal{B}\left(\mu_1^{(\ell)}, \frac{\Delta}{16}\right). \tag{51}$$

As we have $\mathcal{P}_1^{(\ell)} \subset \mathcal{P}_1^{(\ell+1)}$ and $|\mathcal{P}_1^{(\ell)}| \leq |\mathcal{P}_1^{(\ell+1)}| = |\mathcal{P}_1^{(\ell)}| + m$, we get that there is a set $A \subseteq \{Y_i : i \in [n]\}$ that satisfies

$$\mathcal{P}_1^{(\ell)} \supseteq \mathcal{P}_1^{(\ell+1)}/A, \quad |A| \leq m \leq \frac{n\alpha^2}{16},$$

for large enough n such that $\frac{n\alpha^2}{16} > 1$. In view of (51) the last display implies

$$\begin{aligned}
&\mathcal{P}_1^{(\ell)} \cap \{Y_i : i \in T_1^*\} \supseteq \mathcal{P}_1^{(\ell+1)} \cap \{Y_i : i \in T_1^*\}/A \\
&\supseteq \mathcal{B}\left(\mu_1^{(\ell)}, \frac{\Delta}{16}\right) \cap \{Y_i : i \in T_1^*\}/A,
\end{aligned}$$

and hence

$$\begin{aligned}
&\left|\mathcal{P}_1^{(\ell)} \cap \{Y_i : i \in T_1^*\}\right| \\
&\geq \left|\mathcal{B}\left(\mu_1^{(\ell)}, \frac{\Delta}{16}\right) \cap \{Y_i : i \in T_1^*\}\right| - |A| \\
&\geq \left|\mathcal{B}\left(\mu_1^{(\ell)}, \frac{\Delta}{16}\right) \cap \{Y_i : i \in T_1^*\}\right| - m.
\end{aligned} \tag{52}$$

- As we have from (48), with $\Delta \geq 96 c_1$:

$$\mathcal{B}\left(\mu_1^{(\ell)}, \frac{\Delta}{16}\right) \supseteq \mathcal{B}\left(\theta_1, \frac{\Delta}{16} - 3c_1\right) \supseteq \mathcal{B}\left(\theta_1, \frac{\Delta}{32}\right), \tag{53}$$

in view of Lemma 22(ii) we get

$$\begin{aligned}
&\left|\mathcal{B}\left(\mu_1^{(\ell)}, \frac{\Delta}{16}\right) \cap \{Y_i : i \in T_1^*\}\right| \\
&\geq \left|\mathcal{B}\left(\theta_1, \frac{\Delta}{32}\right) \cap \{Y_i : i \in T_1^*\}\right| \\
&\geq n_1^* \left(1 - \frac{5}{4\log\left(1/G\left(\frac{\Delta}{32\sigma}\right)\right)}\right) \\
&\geq n_1^* - \frac{5n}{4\log(1/(G(\Delta/32\sigma)))}.
\end{aligned} \tag{54}$$

Combining (52) and (54) we get (49). Next, to prove the claim (50), we note that:

- In view of Lemma 22 there are at most $\frac{5n}{4\log(1/(G(\Delta/32\sigma)))}$ many points from $\{Y_i : i \in T_2^*\}$ outside $\mathcal{B}(\theta_2, \frac{\Delta}{32})$. As (48) implies $\mathcal{B}\left(\mu_1^{(\ell)}, \frac{\Delta}{16}\right) \subseteq \mathcal{B}\left(\theta_1, \frac{\Delta}{16} + 3c_1\right)$, and, $\mathcal{B}\left(\theta_1, \frac{\Delta}{16} + 3c_1\right)$ and $\mathcal{B}(\theta_2, \frac{\Delta}{32})$ are disjoint, we have

$$\begin{aligned}
&\left|\mathcal{P}_1^{(\ell)} \cap \{Y_i : i \in T_2^*\} \cap \mathcal{B}\left(\mu_1^{(\ell)}, \frac{\Delta}{16}\right)\right| \\
&\leq \left|\{Y_i : i \in T_2^*\}/\mathcal{B}\left(\theta_2, \frac{\Delta}{32}\right)\right| \\
&\leq \frac{5n}{4\log(1/(G(\Delta/32\sigma)))}.
\end{aligned} \tag{55}$$

- On the other hand, $\mathsf{dist}_1^{(\ell)} = \left\{ D(\mu_1^{(\ell)}, \mathcal{P}_1^{(\ell)}) \right\}^{\{1-\beta\}} \leq \frac{\Delta}{16}$ implies that

$$\left|\mathcal{P}_1^{(\ell)} \cap \{Y_i : i \in T_2^*\}/\mathcal{B}\left(\mu_1^{(\ell)}, \frac{\Delta}{16}\right)\right| \leq n\beta. \tag{56}$$

Combining (55) and (56) we get (50).

Hence, we have proven the inequalities (49) and (50). These inequalities together imply

$$\left|\overline{\mathcal{P}_1^{(\ell)}} \cap \{Y_i : i \in T_2^*\}\right| \geq n_2^* - n\beta - \frac{5n}{4\log(1/(G(\Delta/32\sigma)))} \tag{57}$$

$$\left|\overline{\mathcal{P}_1^{(\ell)}} \cap \{Y_i : i \in T_1^*\}\right| \leq m + \frac{5n}{4\log(1/(G(\Delta/32\sigma)))}. \tag{58}$$

In view of $\left|\{Y_i : i \in T_2^*\}/\mathcal{B}(\theta_2, \Delta/32)\right| \leq \frac{5n}{4\log(1/(G(\Delta/32\sigma)))}$ from Lemma 22, we have

$$\begin{aligned}
&\left|\overline{\mathcal{P}_1^{(\ell)}}/\mathcal{B}(\theta_2, \Delta/32)\right| \\
&\leq \left|\left\{\overline{\mathcal{P}_1^{(\ell)}} \cap \{Y_i : i \in T_2^*\}/\mathcal{B}(\theta_2, \Delta/32)\right\}\right. \\
&\quad \left. \cup \left\{\overline{\mathcal{P}_1^{(\ell)}} \cap \{Y_i : i \in T_1^*\}/\mathcal{B}(\theta_2, \Delta/32)\right\}\right| + n^{\mathsf{out}} \\
&\leq \left|\{Y_i : i \in T_2^*\}/\left\{\mathcal{B}(\theta_2, \Delta/32)\right\}\right| \\
&\quad + \left|\overline{\mathcal{P}_1^{(\ell)}} \cap \{Y_i : i \in T_1^*\}\right| + n^{\mathsf{out}} \\
&\leq \frac{5n}{4\log(1/(G(\Delta/32\sigma)))} + m + \frac{5n}{4\log(1/(G(\Delta/32\sigma)))} \\
&\quad + \frac{n\alpha^2}{32} \\
&\leq \frac{3n\alpha^2}{32} + \frac{5n}{2\log(1/(G(\Delta/32\sigma)))} \leq \frac{5n\alpha^2}{32},
\end{aligned} \tag{59}$$

where the last inequality followed from $\frac{5n}{2\log(1/(G(\Delta/32\sigma)))} \leq \frac{n\alpha^2}{16}$ as $\Delta \geq 32\sigma G^{-1}(e^{-\frac{40}{\alpha^2}})$. Then we make the following observations.

- As $\left|\overline{\mathcal{P}_1^{(\ell)}}\right| \geq n\alpha - n\beta - \frac{n\alpha}{16} \geq \frac{11n\alpha}{16}$ from (57), any subset of $\overline{\mathcal{P}_1^{(\ell)}}$ with size $(1-\beta)\left|\overline{\mathcal{P}_1^{(\ell)}}\right|$, discards a set of size at least $\frac{11n\alpha\beta}{16} \geq \frac{n\alpha^2}{6}$ (note that $\beta = \frac{\alpha}{4}$).
- From (59) we get $\left|\overline{\mathcal{P}_1^{(\ell)}}/\mathcal{B}(\theta_2, \Delta/32)\right| \leq \frac{n\alpha^2}{6.4}$. In view of the last argument this implies that the set $\overline{\mathcal{P}_1^{(\ell)}} \cap \mathcal{B}(\theta_2, \Delta/32)$, which has a diameter at most $\frac{\Delta}{16}$, contains more points than any subset of $\overline{\mathcal{P}_1^{(\ell)}}$ with size $(1-\beta)\left|\overline{\mathcal{P}_1^{(\ell)}}\right|$.
- Hence, the diameter of the tightest subset of $\overline{\mathcal{P}_1^{(\ell)}}$ with size $(1-\beta)\left|\overline{\mathcal{P}_1^{(\ell)}}\right|$ is at most $\frac{\Delta}{16}$.

This implies $\mathsf{dist}_2^{(\ell)} \leq \Delta/16$ and concludes our proof. □

Proof of Lemma 26. As $\mathsf{totdist}_\ell \leq \frac{\Delta}{8}$, we have $\mathsf{dist}_i^{(\ell)} \leq \frac{\Delta}{8}$ for $i \in \{1,2\}$. First we show that both $\mu_1^{(\ell)}$ and $\mu_2^{(\ell)}$ lie in $\cup_{i=1,2}\mathcal{B}(\theta_i, \Delta/3)$. If not, without a loss of generality let $\mu_2^{(\ell)}$ lie outside $\cup_{i=1,2}\mathcal{B}(\theta_i, \Delta/3)$. Then we have

$$\mathcal{B}\left(\mu_2^{(\ell)}, \frac{\Delta}{8}\right) \cap \left\{\cup_{i\in\{1,2\}}\mathcal{B}\left(\theta_i, \frac{\Delta}{8}\right)\right\} = \phi, . \tag{60}$$

As we have $\mathsf{dist}_2^{(\ell)} \leq \frac{\Delta}{8}$, we get that

$$\left|\overline{\mathcal{P}_1^{(\ell)}} \cap \mathcal{B}\left(\mu_2^{(\ell)}, \frac{\Delta}{8}\right)\right| \geq (1-\beta)\left|\overline{\mathcal{P}_1^{(\ell)}}\right|. \tag{61}$$

Using Lemma 22, we get that

$$\begin{aligned}
&\left|\{Y_i : i \in [n]\}/\left\{\cup_{i\in\{1,2\}}\mathcal{B}\left(\theta_i, \frac{\Delta}{8}\right)\right\}\right| \\
&\leq \frac{5n}{4\log(1/(G(\Delta/32\sigma)))} + n^{\mathsf{out}}.
\end{aligned} \tag{62}$$

In view of the last display, using (60) and(61) we get

$$\begin{aligned}\left|\overline{\mathcal{P}_1^{(\ell)}}\right| &\le \frac{1}{1-\beta}\left|\overline{\mathcal{P}_1^{(\ell)}} \cap \mathcal{B}\left(\mu_2^{(\ell)}, \frac{\Delta}{8}\right)\right| \\ &\le \frac{1}{1-\beta}\left|\{Y_i : i \in [n]\} / \left\{\cup_{i\in\{1,2\}} \mathcal{B}\left(\theta_i, \frac{\Delta}{8}\right)\right\}\right| \\ &\le \frac{5n}{2\log(1/(G(\Delta/32\sigma)))} + 2n^{\mathsf{out}}.\end{aligned}$$

The last display implies

$$\left|\mathcal{P}_1^{(\ell)}\right| \ge n - \frac{5n}{2\log(1/(G(\Delta/32\sigma)))} - 2n^{\mathsf{out}} \ge n - \frac{n\alpha^2}{8}, \tag{63}$$

where the last inequality followed using $n^{\mathsf{out}} \le \frac{n\alpha^2}{32}$, provided

$$\frac{5n}{2\log(1/(G(\Delta/32\sigma)))} \le \frac{n\alpha^2}{16}, \quad \text{i.e., } \Delta \ge 32\sigma G^{-1}\left(e^{-\frac{40}{\alpha^2}}\right).$$

In view of (62) with $\cap_{i\in\{1,2\}}\mathcal{B}\left(\theta_i, \frac{\Delta}{8}\right) = \phi$, $n^{\mathsf{out}} \le \frac{n\alpha^2}{32}$ and $n_1^*, n_2^* \ge n\alpha$, we get that given any set $\mathcal{S} \subseteq \{Y_i : i \in [n]\}$ of size at least $n - \frac{n\alpha}{2}$, there will be at least two points in $\mathcal{S}$ that are at least $\Delta - \frac{\Delta}{4}$ distance away. Choose $\mathcal{S} = \{i \in [n] : Y_i \in \mathcal{B}(\mu_1^{(\ell)}, \mathsf{dist}_1^{(\ell)})\}$. Using (63) we get

$$|\mathcal{S}| \ge (1-\beta)\left|\mathcal{P}_1^{(\ell)}\right| \ge n(1-\alpha/4)(1-\alpha^2/8) \ge n - \frac{n\alpha}{2},$$

Suppose x, y are the farthest away points in $\mathcal{S}$. This leads to a contradiction as

$$\Delta - \frac{\Delta}{4} \le \|x-y\| \le \|x-\mu_1^{(\ell)}\| + \|y-\mu_1^{(\ell)}\| \le 2\cdot \mathsf{dist}_1^{(\ell)} \le \frac{\Delta}{4}.$$

Now it remains to show that $\mu_1^{(\ell)}, \mu_2^{(\ell)}$ lie in different balls among $\mathcal{B}(\theta_1, \Delta/3)$ and $\mathcal{B}(\theta_2, \Delta/3)$. If not, then suppose that both lie in $\mathcal{B}(\theta_1, \Delta/3)$. Note that either $\mathcal{P}_1^{(\ell)}$ or $\overline{\mathcal{P}_1^{(\ell)}}$ will contain more than half the points from $\{Y_i : i \in T_2^*\} \cap \mathcal{B}\left(\theta_2, \frac{\Delta}{8}\right)$. We deal with the case where $\overline{\mathcal{P}_1^{(\ell)}}$ is the partition with more than half of the points in $\{Y_i : i \in T_2^*\} \cap \mathcal{B}\left(\theta_2, \frac{\Delta}{8}\right)$. The case with $\mathcal{P}_1^{(\ell)}$ can be worked out similarly. Then we have the following

- As $\mathsf{dist}_2^{(\ell)} \le \frac{\Delta}{8}$ and $\mu_2^{(\ell)} \in \mathcal{B}(\theta_1, \Delta/3)$ we get

$$\begin{aligned}&\mathcal{B}\left(\mu_2^{(\ell)}, \mathsf{dist}_2^{(\ell)}\right) \subseteq \mathcal{B}\left(\mu_2^{(\ell)}, \frac{\Delta}{8}\right) \subseteq \mathcal{B}\left(\theta_1, \frac{\Delta}{3} + \frac{\Delta}{8}\right) \\ &= \mathcal{B}\left(\theta_1, \frac{11\Delta}{24}\right)\end{aligned}$$

- In view of the last argument we have

$$\begin{aligned}&\mathcal{B}\left(\theta_2, \frac{\Delta}{8}\right) \cap \mathcal{B}\left(\mu_2^{(\ell)}, \mathsf{dist}_2^{(\ell)}\right) \\ &\subseteq \mathcal{B}\left(\theta_2, \frac{\Delta}{8}\right) \cap \mathcal{B}\left(\theta_1, \frac{11\Delta}{24}\right) = \phi,\end{aligned} \tag{64}$$

- From Lemma 22, whenever $\frac{5n}{8\log(1/(G(\Delta/32\sigma)))} \le \frac{n\alpha}{6}$, i.e., $\Delta \ge 32\sigma G^{-1}\left(e^{-\frac{3}{4\alpha}}\right)$, we get

$$\begin{aligned}&\frac{1}{2}\left|\{Y_i : i \in T_2^*\} \cap \mathcal{B}\left(\theta_2, \frac{\Delta}{8}\right)\right| \\ &\ge \frac{n\alpha}{2} - \frac{5n}{8\log(1/(G(\Delta/32\sigma)))} \ge \frac{n\alpha}{3}.\end{aligned} \tag{65}$$

However, this leads to a contradiction, as in view of (64) we have

$$\begin{aligned}&\left|\overline{\mathcal{P}_1^{(\ell)}} \cap \mathcal{B}\left(\theta_2, \frac{\Delta}{8}\right)\right| \\ &\le \left|\overline{\mathcal{P}_1^{(\ell)}} / \left\{Y_i : i \in [n] Y_i \in \mathcal{B}\left(\mu_2^{(\ell)}, \mathsf{dist}_2^{(\ell)}\right)\right\}\right| \\ &\le n\beta \le \frac{n\alpha}{4},\end{aligned}$$

but on the other hand, using the fact that $\overline{\mathcal{P}_1^{(\ell)}}$ contains more than half of the points in $\{Y_i : i \in T_2^*\} \cap \mathcal{B}\left(\theta_2, \frac{\Delta}{8}\right)$, we get from (65)

$$\left|\overline{\mathcal{P}_1^{(\ell)}} \cap \mathcal{B}\left(\theta_2, \frac{\Delta}{8}\right)\right| \geq \frac{1}{2}\left|\{Y_i : i \in T_2^*\} \cap \mathcal{B}\left(\theta_2, \frac{\Delta}{8}\right)\right| \geq \frac{n\alpha}{3}.$$

□

APPENDIX F
PROOF OF INITIALIZATION RESULTS WITH k-CLUSTERS (THEOREM 8)

We will use the following high probability guarantee for proving our initialization result. The proof is identical to that of Lemma 22 and is omitted.

Lemma 27. *Given any $\beta \in (0,1)$, there is an event $\widetilde{\mathcal{E}}_k$ with $\mathbb{P}\left[\widetilde{\mathcal{E}}_k\right] \geq 1 - 2ke^{-\frac{\min_{g\in[k]} n_g^*}{4}}$ on which the following holds for the k-cluster problem:*

(i) $\left|\mathcal{B}(\theta_i, \sigma G^{-1}(e^{-\frac{5}{4\beta^2}})) \cap \{Y_i : i \in T_i^*\}\right| \geq n_i^*(1-\beta^2)$ *for each $i \in [k]$,*

(ii) For each $i \in [k]$,

$$\left|\mathcal{B}(\theta_i, \frac{\Delta}{16k}) \cap \{Y_i : i \in T_i^*\}\right| \geq n_i^*\left(1 - \frac{5}{4\log\left(1/G\left(\Delta/(16\sigma k)\right)\right)}\right)$$

In the proof below, we assume that all the mentioned constants depend on G, α, σ, k, unless otherwise specified. In addition, for our entire analysis we will assume that the event $\widetilde{\mathcal{E}}_k$ mentioned in Lemma 27 holds, which has a high probability. We will provide a proof of the result involving outliers, as in Theorem 12 as the proof in absence of outliers is very similar. For the sake of simplifying our analysis we will also assume that $\min_{h\in[k]} n_h^* = \frac{n\alpha}{k}$. In summary, the above modifications are equivalent to showing the following:

Suppose that out of the n many observed data points, n_i^* many are from cluster $T_i^*, i = 1, \ldots, k$ and n^{out} many are adversarial outliers (i.e., $\sum_{i=1}^k n_i^* + n^{\mathsf{out}} = n$). Also assume for some constant $1 > \alpha > 0$ the counts satisfy $n_i^* > \frac{n\alpha}{k}, i = 1, \ldots, k$, $n^{\mathsf{out}} \leq \frac{n\alpha^2}{64k^3}$. Then there are constants c_1, c_2 such that the following is satisfied. Whenever $\Delta > c_1 k\sigma G^{-1}\left(e^{-c_2/\beta^2}\right)$, there is a permutation π of the set $[k]$ that satisfies $\max_{i\in[k]} \|\theta_{\pi(i)} - \mu_i^*\| \leq \Delta/3$ with probability at least $1 - 2ke^{-n\alpha/4k}$, where the $\{\mu_i^*\}$ are centroid estimates generated via the $\mathsf{IOD}_{k,m_1,m,\beta}$ algorithm with

$$m_1 = \left\lceil \frac{n\alpha}{4k} \right\rceil, m = \max\left\{1, \left\lfloor \frac{n\beta^2}{2} \right\rfloor\right\}, \beta = \frac{\alpha}{4k^2}.$$

Similar to before, we will extensively use the following definition of order statistics: Given any set V of real numbers and fraction $0 < p < 1$, let $V^{\{p\}}$ define the $\lceil p|V| \rceil$-th smallest number in V. We make the following observations for simplifying the notation. Whenever we call the IOD algorithm to find j centroids from the remaining data set, it contains a *for-loop* with the loop counter denoted by ℓ_j. As a result, whenever we find a set of centroids $\widehat{\mu}_k, \ldots, \widehat{\mu}_2, \widehat{\mu}_2$ it corresponds to a set of loop counts $\widetilde{\ell}_k, \ldots, \widetilde{\ell}_2$

$$(\widehat{\mu}_k, \ldots, \widehat{\mu}_2, \widehat{\mu}_1) = (\mu_k^{(k,\widetilde{\ell}_k)}, \ldots, \mu_2^{(2,\widetilde{\ell}_2)}, \mu_1^{(1,\widetilde{\ell}_2)}),$$

and vice-versa. In view of this relation, in the proofs below we will interchangeably use the centroids and the indices.

The proof is a combination of the following results.

Lemma 28. *There is one θ_i, such that*

$$\|\theta_i - \mu_k^{(k,1)}\| \leq 3\sigma G^{-1}\left(e^{-\frac{5}{4\beta^2}}\right).$$

Lemma 29. *There is a stage $\bar{\ell}_k + 1$, with $\bar{\ell}_k \geq 1$, such that $\mathsf{dist}_k^{(\bar{\ell}_k+1)} > \frac{\Delta}{8k}, \mathsf{dist}_k^{(\bar{\ell}_k)} \leq \frac{\Delta}{8k}$.*

Lemma 30. *There exists steps $\ell_k, \ldots, \ell_2$ such that for each $i = 2, \ldots, k$, at the ℓ_i-th step the distance to the $(1-\beta)|\mathcal{P}_i^{(\ell_i)}|$-th closest point from $\mu_i^{(i,\ell_i)}$, within the set $\mathcal{P}_i^{(\ell_i)}$ will all be smaller than $\frac{\Delta}{8k}$ and the $(1-\beta)|\overline{\mathcal{P}_2^{(\ell_2)}}|$-th closest point from $\mu_1^{(1,1)}$, within the set $\overline{\mathcal{P}_2^{(\ell_2)}}$ will be smaller than $\frac{\Delta}{8k}$.*

Lemma 31. *If $\mathsf{totdist}_k^{(\ell_k)} \leq \frac{\Delta}{8}$ for some ℓ_k, then for the loop-index $\ell_k, \ldots, \ell_2$ achieving the above we get that there is a permutation π of $[k]$ such that (with ℓ_1 being set as ℓ_2)*

$$\max_{i\in[k]} \|\mu_i^{(i,\ell_i)} - \theta_{\pi(i)}\| \leq \frac{\Delta}{3}.$$

Lemma 28, Lemma 29 and Lemma 30 together implies, provided Δ is large enough, that among all of the iterations of our algorithm there is an instance on which the $\mathsf{totdist}_k^{(\ell)}$ measure becomes smaller than $\frac{\Delta}{8}$. As our algorithm finally picks the iteration step $\ell = \ell^*$ with the lowest $\mathsf{totdist}_k^{(\ell)}$ measure, it ensures that $\mathsf{totdist}_k^{(\ell_k^*)} \leq \frac{\Delta}{8}$. In view of Lemma 31 this implies $\max_{i\in[k]} \|\theta_{\pi(i)} - \mu_i\| \leq \Delta/3$, for the centroid estimates $\mu_k, \ldots, \mu_1$ generated at that iteration stage, as required. Now we prove below Lemma 28, Lemma 29, Lemma 30 and Lemma 31.

Proof of Lemma 28. In view of Lemma 27, there is a constant $c_1 > 0$ such that

$$|\{j \in [n] : Y_j \in \mathcal{B}(\theta_i, c_1)\}| \geq n_i^* \left(1 - \beta^2\right), \quad i \in [k]. \tag{66}$$

As we have $n_i^* \geq \frac{n\alpha}{k}$ by assumption, it follows that there is a point Y_i such that

$$|\{j \in [n] : Y_j \in \mathcal{B}(Y_i, 2c_1)\}| \geq \frac{n\alpha}{4k} (= m_1).$$

Hence, the tightest neighborhood around any point $Y_i, i \in [n]$, that contains at least $\frac{n\alpha}{4k}$ points from $Y_1, \ldots, Y_n$, has a radius of at most $2c_1$ around that Y_i. Using the definition (46)

$$D(x, S) = \{\|x - Y_i\| : i \in S\}, \quad x \in \mathbb{R}^d, S \subseteq [n],$$

pick $i^* \in [n]$ that satisfies

$$D(Y_{i^*}, [n])^{\{\frac{m_1}{n}\}} = \min_{j\in[n]} D(Y_j, [n])^{\{\frac{m_1}{n}\}}.$$

Then $\mathcal{B}(Y_{i^*}, 2c_1)$ and $\cup_{j\in[k]}\mathcal{B}(\theta_j, c_1)$ can not be disjoint, as in view of (66) it will imply that their union will contain more than n points from $Y_1, \ldots, Y_n$

$$\begin{aligned}
&\Big|\{i \in [n] : Y_i \in \mathcal{B}(Y_{i^*}, 2c_1)\} \\
&\quad \cup \left[\cup_{j\in[k]} \{i \in [n] : Y_i \in \mathcal{B}(\theta_j, c_1)\}\right]\Big| \\
&\geq \frac{n\alpha}{4k} + \sum_{j\in[k]} n_j^* \left(1 - \beta^2\right) \\
&\geq (n - n^{\mathsf{out}})(1 - \beta^2) + \frac{n\alpha}{4k} > n + \frac{n\alpha}{8k},
\end{aligned}$$

where we assume that

$$n^{\mathsf{out}} \leq \frac{n\alpha}{16k}, \tag{67}$$

and we use the fact that $\{i \in [n] : Y_i \in \mathcal{B}(\theta_j, c_1)\}, j \in [k]$ are disjoint sets as $\min_{g\neq h\in[k]} \|\theta_g - \theta_h\| \geq \Delta$. Hence, Y_{i^*} is at a distance at most $3c_1$ from one of the centroids. Without a loss of generality we can pick $\mu_k^{(k,1)} = Y_{i^*}$ and we assume that θ_k is the closest true centroid to $\mu_k^{(k,1)}$ than any of the other centroids. □

Proof of Lemma 29. In view of Lemma 28 we have for $c_1 = \sigma G^{-1}\left(e^{-\frac{5}{4\beta^2}}\right)$ and $j \in [k]$

$$\begin{aligned}
\mu_k^{(k,1)} &\in \mathcal{B}\left(\theta_k, 3c_1\right), \\
\left|\{Y_i : i \in T_j^*\} \cap \mathcal{B}\left(\theta_j, c_1\right)\right| &\geq n_j^* \left(1 - \beta^2\right).
\end{aligned} \tag{68}$$

We observe the following:

- The set $\mathcal{B}(\mu_k^{(k,1)}, 4c_1)$ contains $\mathcal{B}(\theta_k, c_1)$, which contains at least $\frac{n\alpha}{k}(1 - \beta^2)$ points from $\{Y_i : i \in T_k^*\}$. As the size of $\mathcal{P}_k^{(1)}$ is at most $\lceil \frac{n\alpha}{4k} \rceil$ the distance of $\mu_k^{(k,1)}$ to any point in $\mathcal{P}_k^{(1)}$ is less than $4c_1$. As we have $\Delta \geq 32kc_1$ the last statement implies $\mathsf{dist}_k^{(1)} \leq \frac{\Delta}{8k}$.
- At the last step, say $\widetilde{\ell}_k$, in the for loop indexed by ℓ_k, $\mathcal{P}_k^{(\widetilde{\ell}_k)}$ will have at least $n - m$ many points (recall that $m = \left\lfloor \frac{n\alpha^2}{32k^4} \right\rfloor \geq 1$ for large n). This implies:
  (a) The tightest neighborhood (say N) around $\mu_k^{(k,1)}$ with a size at least $(1-\beta)|\mathcal{P}_k^{(\widetilde{\ell}_k)}|$ will include at least $(1-\beta)(n-m) \geq n - n\beta - m$ points,
  (b) (68) implies that $\cup_{j\in[k-1]} \left\{\{Y_i : i \in T_j^*\} \cap \mathcal{B}\left(\theta_j, c_1\right)\right\}$ will contain at least $\frac{n\alpha}{2k}$ points.

Hence we get that the sets N and $\cup_{j\in[k-1]}\{Y_i : i\in[n], Y_i\in\mathcal{B}(\theta_j,c_1)\}$ can not be disjoint as their union will contain at least n points. The above implies that the neighborhood N will contain at least one point y from the set $\cup_{j\in[k-1]}\{Y_i : i\in[n], Y_i\in\mathcal{B}(\theta_j,c_1)\}$. Suppose that $y\in N$ is such that

$$y\in\{Y_i : i\in[n], Y_i\in\mathcal{B}(\theta_j,c_1)\} \text{ for some } j\in[k-1].$$

Then the distance of y from $\mu_k^{(\bar{\ell}_k)}$ is at least $\Delta-4c_1$,

$$\begin{aligned}&\|\mu_k^{(\bar{\ell}_k)}-y\|\\ &\geq\|\theta_k-\theta_j\|-\|\mu_k^{(\bar{\ell}_k)}-\theta_k\|-\|\theta_j-y\|\geq\Delta-4c_1.\end{aligned}\tag{69}$$

As $\Delta-4c_1\geq\frac{\Delta}{8k}$ we have that there exist some $1\leq\ell_k\leq n-1$ such that $\mathsf{dist}_k^{(\ell_k+1)}>\frac{\Delta}{8k}$. Then the following choice of $\bar{\ell}_k$ finishes the proof

$$\bar{\ell}_k=\min\left\{r\geq 1:\mathsf{dist}_k^{(r+1)}>\frac{\Delta}{8k}\right\}.$$

□

Proof of Lemma 30. We will verify the result using an induction argument: The following is satisfied for each $i=k,k-1,\ldots,2$ (induction variable). There exists an index value $\bar{\ell}_j$ corresponding the the j-th loop count ℓ_j, for $j=k,\ldots,2$ such that the corresponding centroids $\mu_k^{(k,\bar{\ell}_k)},\ldots,\mu_i^{(i,\bar{\ell}_i)}$ satisfy

(Q1) For each $g=k,\ldots,2$, there is one θ_g, such that $\|\theta_g-\mu_g^{(g,\ell_g)}\|\leq 3\sigma G^{-1}\left(e^{-\frac{5}{4\beta^2}}\right)$.

(Q2) At the ℓ_i-th step the distance to the $(1-\beta)|\mathcal{P}_i^{(\ell_i)}|$-th closest point from $\mu_i^{(i,1)}$, within the set $\mathcal{P}_i^{(\ell_i)}$ will all be smaller than $\frac{\Delta}{8k}$.

(Q3) For $h=1,\ldots,i-1$

$$\begin{aligned}&\left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}}\cap\{Y_j : j\in T_h^*\}\right|\\ &\geq n_h^*-(k-i+1)n\beta-\frac{5(k-i+1)n_h^*}{4\log(1/(G(\Delta/16\sigma k)))}.\end{aligned}$$

(Q4)

$$\begin{aligned}&\left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}}\cap\{\cup_{g=i}^k\{Y_j : j\in T_g^*\}\}\right|\\ &\leq(k-i+1)m+\frac{5\sum_{g=i}^k n_g^*}{4\log(1/G(\Delta/(16k\sigma)))}.\end{aligned}$$

Base case $i=k$: Note that our algorithm starts by picking the tightest neighborhood with m_1 points, and then we keep adding m points from $\overline{\mathcal{P}_k^{(\ell_k)}}$ to $\mathcal{P}_k^{(\ell_k)}$ at each step. In view of Lemma 28 we get that the estimate $\mu_k^{(k,1)}$ lies within a radius $3c_1$ of θ_k, with $c_1=\sigma G^{-1}\left(e^{-\frac{5}{4\beta^2}}\right)$. Hence (Q1) is satisfied. In view of Lemma 29, when we run the k-th for loop at the iteration $\bar{\ell}_k$, we get $\mathsf{dist}_k^{(\bar{\ell}_k)}\leq\frac{\Delta}{8k}$, and hence (Q2) is satisfied.

Let $\bar{\ell}_k$ be as in Lemma 29. Without a loss of generality we assume that θ_k is the closest centroid to $\mu_k^{(k,1)}$ and (68) holds. We first prove the following claims:

$$\left|\mathcal{P}_k^{(\bar{\ell}_k)}\cap\{Y_i : i\in T_k^*\}\right|\geq n_k^*-m-\frac{5n_k^*}{4\log(1/(G(\Delta/16\sigma k)))}\tag{70}$$

$$\left|\mathcal{P}_k^{(\bar{\ell}_k)}\cap\{Y_i : i\in T_j^*\}\right|\leq n\beta+\frac{5n_j^*}{4\log(1/(G(\Delta/16\sigma k)))},\tag{71}$$

where $j\in[k-1]$. The first claim (70) follows from the following sequence of arguments (note the definition in (46))

- Using $\mu_k^{(k,\bar{\ell}_k)}=\mu_k^{(k,\bar{\ell}_k+1)}$ and from Lemma 29 we get

$$\begin{aligned}&\left\{D(\mu_k^{(k,\bar{\ell}_k)},\mathcal{P}_k^{(\bar{\ell}_k+1)})\right\}^{\{1-\beta\}}\\ &=\left\{D(\mu_k^{(k,\bar{\ell}_k+1)},\mathcal{P}_k^{(\bar{\ell}_k+1)})\right\}^{\{1-\beta\}}=\mathsf{dist}_k^{(\bar{\ell}_k+1)}>\frac{\Delta}{8k},\end{aligned}$$

which implies

$$\mathcal{P}_k^{(\bar{\ell}_k+1)}\supseteq\{Y_i : i\in[n]\}\cap\mathcal{B}\left(\mu_k^{(k,\bar{\ell}_k)},\frac{\Delta}{8k}\right).\tag{72}$$

As we have $\mathcal{P}_k^{(\bar{\ell}_k)} \subset \mathcal{P}_k^{(\bar{\ell}_k+1)}$ and $|\mathcal{P}_k^{(\bar{\ell}_k)}| \le |\mathcal{P}_k^{(\bar{\ell}_k+1)}| \le |\mathcal{P}_k^{(\bar{\ell}_k)}| + m$, we get that there is a set $A \subseteq \{Y_i : i \in [n]\}$ that satisfies

$$\mathcal{P}_k^{(\bar{\ell}_k)} \supseteq \mathcal{P}_k^{(\bar{\ell}_k+1)}/A, \quad |A| \le m.$$

In view of (72) the last display implies

$$\begin{aligned}
&\left\{\mathcal{P}_k^{(\bar{\ell}_k)} \cap \{Y_i : i \in T_k^*\}\right\} \\
&\supseteq \left\{\mathcal{P}_k^{(\bar{\ell}_k+1)} \cap \{Y_i : i \in T_k^*\}/A\right\} \\
&\supseteq \left\{\mathcal{B}\left(\mu_k^{(k,\bar{\ell}_k)}, \frac{\Delta}{8k}\right) \cap \{Y_i : i \in T_k^*\}/A\right\},
\end{aligned}$$

and hence

$$\begin{aligned}
&\left|\mathcal{P}_k^{(\bar{\ell}_k)} \cap \{Y_i : i \in T_k^*\}\right| \\
&\ge \left|\mathcal{B}\left(\mu_k^{(k,\bar{\ell}_k)}, \frac{\Delta}{8k}\right) \cap \{Y_i : i \in T_k^*\}\right| - |A| \\
&\ge \left|\mathcal{B}\left(\mu_k^{(k,\bar{\ell}_k)}, \frac{\Delta}{8k}\right) \cap \{Y_i : i \in T_k^*\}\right| - \frac{n\beta^2}{2}.
\end{aligned} \tag{73}$$

- Note that we have from (68) and $\Delta \ge 48\sigma k$:

$$\mathcal{B}\left(\mu_k^{(k,\bar{\ell}_k)}, \frac{\Delta}{8k}\right) \supseteq \mathcal{B}\left(\theta_k, \frac{\Delta}{8k} - 3c_1\right) \supseteq \mathcal{B}\left(\theta_k, \frac{\Delta}{16k}\right). \tag{74}$$

In view of Lemma 27(ii) this implies

$$\begin{aligned}
&\left|\mathcal{B}\left(\mu_k^{(k,\bar{\ell}_k)}, \frac{\Delta}{8k}\right) \cap \{Y_i : i \in T_k^*\}\right| \\
&\ge \left|\mathcal{B}\left(\theta_k, \frac{\Delta}{16k}\right) \cap \{Y_i : i \in T_k^*\}\right| \\
&\ge n_k^* \left(1 - \frac{5}{4\log\left(1/G\left(\frac{\Delta}{16\sigma k}\right)\right)}\right) \\
&\ge n_k^* - \frac{5n_k^*}{4\log(1/(G(\Delta/16\sigma k)))}.
\end{aligned} \tag{75}$$

Combining (73) and (75) we get (70).

Next, to prove the claim (71), we note that:

- In view of Lemma 27, for each $j = 1, \ldots, k-1$, there are at most $\frac{5n_j^*}{4\log(1/(G(\Delta/16\sigma k)))}$ many points from $\{Y_i : i \in T_j^*\}$ outside $\mathcal{B}(\theta_j, \frac{\Delta}{16k})$. In view of (68) and $\mu_k^{(k,\bar{\ell}_k)} = \mu_k^{(k,1)}$ we get $\mathcal{B}\left(\mu_k^{(k,\bar{\ell}_k)}, \frac{\Delta}{8k}\right)$ is a subset of $\mathcal{B}\left(\theta_k, \frac{\Delta}{8k} + 3c_1\right)$ as $\Delta \ge 24c_1 k$. Hence we get $\mathcal{B}\left(\mu_k^{(k,\bar{\ell}_k)}, \frac{\Delta}{8k}\right)$ and $\cup_{j=1}^{k-1}\mathcal{B}(\theta_j, \frac{\Delta}{16k})$ are disjoint, and hence we have for each $j = 1, \ldots, k-1$

$$\begin{aligned}
&\left|\mathcal{P}_k^{(\bar{\ell}_k)} \cap \{Y_i : i \in T_j^*\} \cap \mathcal{B}\left(\mu_k^{(k,\bar{\ell}_k)}, \frac{\Delta}{8k}\right)\right| \\
&\le \left|\{Y_i : i \in T_j^*\}/\cup_{j=1}^{k-1}\mathcal{B}\left(\theta_j, \frac{\Delta}{16k}\right)\right| \\
&\le \frac{5n_j^*}{4\log(1/(G(\Delta/16\sigma k)))}.
\end{aligned} \tag{76}$$

- On the other hand, from Lemma 29 we have

$$\mathsf{dist}_k^{(\bar{\ell}_k)} = \left\{D(\mu_k^{(k,\bar{\ell}_k)}, \mathcal{P}_k^{(\bar{\ell}_k)})\right\}^{\{1-\beta\}} \le \frac{\Delta}{8k}.$$

This implies that for each $j \in [k-1]$

$$\left|\mathcal{P}_k^{(\bar{\ell}_k)} \cap \{Y_i : i \in T_j^*\}/\mathcal{B}\left(\mu_k^{(k,\bar{\ell}_k)}, \frac{\Delta}{8k}\right)\right| \le n\beta. \tag{77}$$

Combining (76) and (77) we get (71).

Hence, we have proven the inequalities (70) and (71). These inequalities together imply for $j \in [k-1]$

$$\left|\overline{\mathcal{P}_k^{(\bar{\ell}_k)}} \cap \{Y_i : i \in T_j^*\}\right| \geq n_j^* - n\beta - \frac{5n_j^*}{4\log(1/(G(\Delta/16\sigma k)))}, \tag{78}$$

$$\left|\overline{\mathcal{P}_k^{(\bar{\ell}_k)}} \cap \{Y_i : i \in T_k^*\}\right| \leq m + \frac{5n_k^*}{4\log(1/(G(\Delta/16\sigma k)))}. \tag{79}$$

The first inequality above verifies (Q3) and the second inequality verifies (Q4).

Induction step from i ***to*** $i-1$*:* To complete the induction argument, let us suppose that the statement holds for some $3 \leq i \leq k$ and we intend to prove the case of $i-1$. The proof of (Q1) follows from the following general result. The proof is essentially a repetition of argument as in the proof of Lemma 28, and is presented at the end of this section.

Lemma 32. *Suppose that we have for $i \geq 3$ and $h = 1, \ldots, i-1$*

$$\left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_h^*\}\right| \geq \frac{3n_h^*}{5},$$
$$\left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \left\{\cup_{g=i}^k \{Y_j : j \in T_g^*\}\right\}\right| \leq \frac{n\alpha\beta}{5k}.$$

Then there is a centroid θ_{i-1} such that $\|\mu_{i-1}^{(i-1,1)} - \theta_{i-1}\| \leq 3\sigma G^{-1}\left(e^{-\frac{5}{4\beta^2}}\right)$ if $\Delta \geq 16\sigma k G^{-1}\left(e^{-\frac{5k}{\alpha}}\right)$.

Next we prove (Q2) for the loop indexed by ℓ_{i-1}. In view of the Lemma 32 and Lemma 27 we note that for $h \in [i-1]$

$$\begin{aligned} \mu_{i-1}^{(i-1,1)} &\in \mathcal{B}\left(\theta_{i-1}, 3c_1\right), \\ \left|\{Y_j : j \in T_h^*\} \cap \mathcal{B}\left(\theta_h, c_1\right)\right| &\geq n_h^*\left(1-\beta^2\right). \end{aligned} \tag{80}$$

where $c_1 = \sigma G^{-1}\left(e^{-\frac{5}{4\beta^2}}\right)$. In view of a reasoning similar as in the proof of Lemma 29 we note the following. As we keep adding m points from $\overline{\mathcal{P}_{i-1}^{(\ell_{i-1})}}$ to $\mathcal{P}_{i-1}^{(\ell_{i-1})}$ at each step $\ell_{i-1} = 2, \ldots, \left\lfloor \frac{n'-m_1}{m} \right\rfloor$, note that at some stage ℓ_{i-1}, before we exhaust all the points, the distance to the $(1-\beta)|\mathcal{P}_{i-1}^{(\ell_{i-1})}|$-th closest point from $\mu_{i-1}^{(i-1,1)}$, within the set $\mathcal{P}_{i-1}^{(\ell_{i-1})}$ will exceed $\frac{\Delta}{8k}$. Hence, to prove our claim (Q2) we observe the following:

- In view of (80) we have $\mathcal{B}(\mu_{i-1}^{(i-1,1)}, 4c_1) \supseteq \mathcal{B}(\theta_{i-1}, c_1)$, which implies

$$\begin{aligned} &\left|\{Y_j : j \in T_{i-1}^*\}/\mathcal{B}(\mu_{i-1}^{(i-1,1)}, 4c_1)\right| \\ &\leq \left|\{Y_j : j \in T_{i-1}^*\}/\mathcal{B}(\theta_{i-1}, c_1)\right| \leq n_{i-1}^*\beta^2. \end{aligned} \tag{81}$$

  In view of the assumption (Q3) at the induction step i the last display implies

$$\begin{aligned} &\left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_{i-1}^*\} \cap \mathcal{B}(\mu_{i-1}^{(i-1,1)}, 4c_1)\right| \\ &= \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_{i-1}^*\}\right| \\ &\quad - \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_{i-1}^*\}/\mathcal{B}(\mu_{i-1}^{(i-1,1)}, 4c_1)\right| \\ &\geq \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_{i-1}^*\}\right| \\ &\quad - \left|\{Y_j : j \in T_{i-1}^*\}/\mathcal{B}(\mu_{i-1}^{(i-1,1)}, 4c_1)\right| \\ &\overset{(a)}{\geq} n_{i-1}^* - (k-i+1)n\beta - n_{i-1}^*\beta^2 \\ &\quad - \frac{5(k-i+1)n_{i-1}^*}{4\log(1/(G\left(\frac{\Delta}{16\sigma k}\right)))} \\ &\overset{(b)}{\geq} \frac{n\alpha}{2k}, \end{aligned}$$

  where (a) used the inequality (81) and (b) holds whenever $\Delta \geq 16\sigma k G^{-1}\left(e^{-\frac{10k}{\alpha}}\right)$ as $(k-i+1)n\beta \leq \frac{n\alpha}{4k}, n\beta^2 \leq \frac{n\alpha}{16k^4}$. As the size of $\mathcal{P}_{i-1}^{(1)}$ is at most $\lceil \frac{n\alpha}{4k} \rceil$ the distance of $\mu_{i-1}^{(i-1,1)}$ to any point in $\mathcal{P}_{i-1}^{(1)}$ is less than $4c_1$, which implies $\mathsf{dist}_{i-1}^{(1)} \leq \frac{\Delta}{8k}$.

- We first note that at the last step, say $\widetilde{\ell}_{i-1}$, in the for loop indexed by ℓ_{i-1}, $\overline{\mathcal{P}_{i-1}^{(\widetilde{\ell}_{i-1})}}$ will have at most m many points and

$$\overline{\mathcal{P}_{i}^{(\widetilde{\ell}_{i})}} = \overline{\mathcal{P}_{i-1}^{(\widetilde{\ell}_{i-1})}} \cup \mathcal{P}_{i-1}^{(\widetilde{\ell}_{i-1})}, \quad \left|\overline{\mathcal{P}_{i-1}^{(\widetilde{\ell}_{i-1})}}\right| \leq m. \tag{82}$$

  Hence, in view of (Q3) and $n_1^* \geq 4n\beta$ we get

$$\begin{aligned}
&\left|\mathcal{P}_{i-1}^{(\widetilde{\ell}_{i-1})} \cap \{Y_j : j \in T_1^*\}\right| \\
&\geq \left|\overline{\mathcal{P}_{i}^{(\widetilde{\ell}_{i})}} \cap \{Y_j : j \in T_1^*\}\right| - m \\
&\overset{(a)}{\geq} n_1^* - kn\beta - \frac{n\beta^2}{2} - \frac{5kn_1^*}{4\log(1/G(\frac{\Delta}{16\sigma k}))} \overset{(b)}{\geq} 2n\beta,
\end{aligned} \tag{83}$$

  where (a) followed from (82) and (b) followed as $\Delta \geq 16\sigma k G^{-1}\left(e^{-\frac{5k^2}{\alpha}}\right)$ and $kn\beta \leq \frac{n\alpha}{4k}, n\beta^2 \leq \frac{n\alpha}{16k^4}$. As the tightest neighborhood (say N) around $\mu_{i-1}^{(i-1,1)}$ with a size at least $(1-\beta)|\mathcal{P}_{i-1}^{(\widetilde{\ell}_{i-1})}|$ will exclude at most $n\beta$ points from $\mathcal{P}_{i-1}^{(\widetilde{\ell}_{i-1})}$, in view of (83) we get that the neighborhood N will include at least $n\beta$ points from $\{Y_j : j \in T_1^*\}$. Now, (68) implies that $\{Y_i : i \in T_1^*\} \cap \mathcal{B}(\theta_1, c_1)$ will contain at lest $n_1^*(1-\beta^2)$ points from $\{Y_j : j \in T_1^*\}$, hence we get that the above neighborhood N will contain at least one point $y \in \{Y_j : j \in [n], Y_i \in \mathcal{B}(\theta_1, c_1)\}$. Then the distance of y from $\mu_{i-1}^{(i-1,\widetilde{\ell}_{i-1})}$ is at least $\Delta - 4c_1$,

$$\begin{aligned}
&\|\mu_{i-1}^{(i-1,\widetilde{\ell}_{i-1})} - y\| \\
&\geq \|\theta_1 - \theta_{i-1}\| - \|\mu_{i-1}^{(i-1,\widetilde{\ell}_{i-1})} - \theta_{i-1}\| - \|\theta_1 - y\| \\
&\geq \Delta - 4c_1.
\end{aligned}$$

Hence we have that there exist some $1 \leq \ell_{i-1} \leq n-1$ such that $\mathsf{dist}_{i-1}^{(\ell_{i-1}+1)} > \frac{\Delta}{8k}$. Choose $\bar{\ell}_{i-1}$ as

$$\bar{\ell}_{i-1} = \min\left\{r \geq 1 : \mathsf{dist}_{i-1}^{(r+1)} > \frac{\Delta}{8k}\right\}. \tag{84}$$

to satisfy the condition (Q2).

Next we establish (Q3) and (Q4) for the induction level $i-1$. Let $\bar{\ell}_{i-1}$ is as in the last definition. We prove the following claims: For $h = 1, \ldots, i-2$

$$\begin{aligned}
&\left|\overline{\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})}} \cap \{Y_j : j \in T_h^*\}\right| \\
&\geq n_h^* - (k-i+2)n\beta - \frac{5(k-i+2)n_h^*}{4\log(1/(G(\Delta/16\sigma k)))},
\end{aligned} \tag{85}$$

$$\left|\overline{\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})}} \cap \{Y_j : j \in T_{i-1}^*\}\right| \leq m + \frac{5n_{i-1}^*}{4\log(1/(G(\Delta/16\sigma k)))}. \tag{86}$$

To prove the claim (85), we note that:

- In view of Lemma 27, for each $h = 1, \ldots, i-2$, there are at most $\frac{5n_h^*}{4\log(1/(G(\frac{\Delta}{16\sigma k})))}$ many points from $\{Y_j : j \in T_h^*\}$ outside $\mathcal{B}(\theta_h, \frac{\Delta}{16k})$. In view of (80) we get

$$\mathcal{B}\left(\mu_{i-1}^{(i-1,\bar{\ell}_{i-1})}, \frac{\Delta}{8k}\right) \subseteq \mathcal{B}\left(\theta_{i-1}, \frac{\Delta}{8k} + 4c_1\right).$$

  This implies $\mathcal{B}\left(\mu_{i-1}^{(i-1,\bar{\ell}_{i-1})}, \frac{\Delta}{8k}\right)$ and $\cup_{h=1}^{i-2}\mathcal{B}(\theta_h, \frac{\Delta}{16k})$ are disjoint. Hence for $h \in [i-2]$

$$\begin{aligned}
&\left|\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})} \cap \{Y_j : j \in T_h^*\} \cap \mathcal{B}\left(\mu_{i-1}^{(i-1,\bar{\ell}_{i-1})}, \frac{\Delta}{8k}\right)\right| \\
&\leq \left|\{Y_j : j \in T_h^*\}/\cup_{h=1}^{i-2}\mathcal{B}\left(\theta_h, \frac{\Delta}{16k}\right)\right| \\
&\leq \frac{5n_h^*}{4\log(1/(G(\Delta/16\sigma k)))},
\end{aligned} \tag{87}$$

  where the last inequality followed from Lemma 27.

- On the other hand, in view of already proven (Q2) at the induction step $i-1$ we get $\mathsf{dist}_{i-1}^{(\bar\ell_{i-1})} = \left\{D(\mu_{i-1}^{(i-1,\bar\ell_{i-1})}, \mathcal{P}_{i-1}^{(\bar\ell_{i-1})})\right\}^{\{1-\beta\}} \le \frac{\Delta}{8k}$, which implies that for each $h \in [i-2]$

$$\left|\mathcal{P}_{i-1}^{(\bar\ell_{i-1})} \cap \{Y_j : j \in T_h^*\}/\mathcal{B}\left(\mu_{i-1}^{(i-1,\bar\ell_{i-1})}, \frac{\Delta}{8k}\right)\right| \le n\beta. \tag{88}$$

Combining (87) and (88) we get

$$\begin{aligned}
&\left|\mathcal{P}_{i-1}^{(\bar\ell_{i-1})} \cap \{Y_j : j \in T_h^*\}\right| \\
&\le \left|\mathcal{P}_{i-1}^{(\bar\ell_{i-1})} \cap \{Y_j : j \in T_h^*\} \cap \mathcal{B}\left(\mu_{i-1}^{(i-1,\bar\ell_{i-1})}, \frac{\Delta}{8k}\right)\right| \\
&\quad + \left|\mathcal{P}_{i-1}^{(\bar\ell_{i-1})} \cap \{Y_j : j \in T_h^*\}/\mathcal{B}\left(\mu_{i-1}^{(i-1,\bar\ell_{i-1})}, \frac{\Delta}{8k}\right)\right| \\
&\le n\beta + \frac{5n_h^*}{4\log(1/(G(\Delta/16\sigma k)))}.
\end{aligned}$$

Combining the above display with (Q3) at the induction level i we get for each $h \in [i-2]$

$$\begin{aligned}
&\left|\overline{\mathcal{P}_{i-1}^{(\bar\ell_{i-1})}} \cap \{Y_j : j \in T_h^*\}\right| \\
&= \left|\overline{\mathcal{P}_{i}^{(\bar\ell_{i})}} \cap \{Y_j : j \in T_h^*\}\right| - \left|\mathcal{P}_{i-1}^{(\bar\ell_{i-1})} \cap \{Y_j : j \in T_h^*\}\right| \\
&\ge n_h^* - (k-i+2)n\beta - \frac{5(k-i+2)n_h^*}{4\log(1/(G(\Delta/16\sigma k)))}.
\end{aligned}$$

This completes the verification of (Q3) for the level $i-1$.

The claim (86) follows from the following sequence of arguments (note the definition in (46))

- Using $\mu_{i-1}^{(i-1,\bar\ell_{i-1})} = \mu_{i-1}^{(i-1,\bar\ell_{i-1}+1)}$ and (84) we get

$$\begin{aligned}
&\left\{D(\mu_{i-1}^{(i-1,\bar\ell_{i-1})}, \mathcal{P}_{i-1}^{(\bar\ell_{i-1}+1)})\right\}^{\{1-\beta\}} \\
&= \left\{D(\mu_{i-1}^{(i-1,\bar\ell_{i-1}+1)}, \mathcal{P}_{i-1}^{(\bar\ell_{i-1}+1)})\right\}^{\{1-\beta\}} \\
&= \mathsf{dist}_{i-1}^{(\bar\ell_{i-1}+1)} > \frac{\Delta}{8k},
\end{aligned}$$

  which implies

$$\mathcal{P}_{i-1}^{(\bar\ell_{i-1}+1)} \supseteq \overline{\mathcal{P}_{i}^{(\bar\ell_{i})}} \cap \mathcal{B}\left(\mu_{i-1}^{(i-1,\bar\ell_{i-1})}, \frac{\Delta}{8k}\right). \tag{89}$$

  As we have $\mathcal{P}_{i-1}^{(\bar\ell_{i-1})} \subset \mathcal{P}_{i-1}^{(\bar\ell_{i-1}+1)}$ and $|\mathcal{P}_{i-1}^{(\bar\ell_{i-1})}| \le |\mathcal{P}_{i-1}^{(\bar\ell_{i-1}+1)}| \le |\mathcal{P}_{i-1}^{(\bar\ell_{i-1})}| + \frac{n\beta^2}{2}$, we get that there is a set $A \subseteq \{Y_i : i \in [n]\}$ that satisfies

$$\mathcal{P}_{i-1}^{(\bar\ell_{i-1})} \supseteq \mathcal{P}_{i-1}^{(\bar\ell_{i-1}+1)}/A, \quad |A| \le m.$$

  In view of (89) the last display implies

$$\begin{aligned}
&\left\{\mathcal{P}_{i-1}^{(\bar\ell_{i-1})} \cap \{Y_j : j \in T_{i-1}^*\}\right\} \\
&\supseteq \left\{\mathcal{P}_{i-1}^{(\bar\ell_{i-1}+1)} \cap \{Y_j : j \in T_{i-1}^*\}/A\right\} \\
&\supseteq \left\{\overline{\mathcal{P}_{i}^{(\bar\ell_{i})}} \cap \mathcal{B}\left(\mu_{i-1}^{(i-1,\bar\ell_{i-1})}, \frac{\Delta}{8k}\right) \cap \{Y_j : j \in T_{i-1}^*\}/A\right\},
\end{aligned}$$

  and hence

$$\begin{aligned}
&\left|\mathcal{P}_{i-1}^{(\bar\ell_{i-1})} \cap \{Y_j : j \in T_{i-1}^*\}\right| \\
&\ge \left|\overline{\mathcal{P}_{i}^{(\bar\ell_{i})}} \cap \mathcal{B}\left(\mu_{i-1}^{(i-1,\bar\ell_{i-1})}, \frac{\Delta}{8k}\right) \cap \{Y_j : j \in T_{i-1}^*\}\right| - |A| \\
&\ge \left|\overline{\mathcal{P}_{i}^{(\bar\ell_{i})}} \cap \mathcal{B}\left(\mu_{i-1}^{(i-1,\bar\ell_{i-1})}, \frac{\Delta}{8k}\right) \cap \{Y_j : j \in T_{i-1}^*\}\right| - m.
\end{aligned} \tag{90}$$

- As we have from (80) and $\Delta \geq 48c_1k$:

$$\mathcal{B}\left(\mu_{i-1}^{(i-1,\bar{\ell}_{i-1})}, \frac{\Delta}{8k}\right) \supseteq \mathcal{B}\left(\theta_{i-1}, \frac{\Delta}{8k} - 3c_1\right) \supseteq \mathcal{B}\left(\theta_{i-1}, \frac{\Delta}{16k}\right),$$

in view of Lemma 27(ii) we get

$$\begin{aligned}
&\left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \mathcal{B}\left(\mu_{i-1}^{(i-1,\bar{\ell}_{i-1})}, \frac{\Delta}{8k}\right) \cap \{Y_j : j \in T_{i-1}^*\}\right| \\
&\geq \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \mathcal{B}\left(\theta_{i-1}, \frac{\Delta}{16k}\right) \cap \{Y_j : j \in T_{i-1}^*\}\right| \\
&\geq \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_{i-1}^*\}\right| \\
&\quad - \left|\{Y_j : j \in T_{i-1}^*\}/\mathcal{B}\left(\theta_{i-1}, \frac{\Delta}{16k}\right)\right| \\
&\geq \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_{i-1}^*\}\right| - \frac{5n_{i-1}^*}{4\log(1/(G(\Delta/16\sigma k)))}. \qquad (91)
\end{aligned}$$

where the last inequality followed from Lemma 27.

Combining (90), (91) and $\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} = \mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})} \cup \overline{\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})}}$ we get

$$\begin{aligned}
&\left|\overline{\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})}} \cap \{Y_j : j \in T_{i-1}^*\}\right| \\
&= \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_{i-1}^*\}\right| - \left|\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})} \cap \{Y_j : j \in T_{i-1}^*\}\right| \\
&\leq m + \frac{5n_{i-1}^*}{4\log(1/(G(\Delta/16\sigma k)))}.
\end{aligned}$$

Using (Q4) for the induction level i, with $\overline{\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})}} \subseteq \overline{\mathcal{P}_i^{(\bar{\ell}_i)}}$ we get

$$\begin{aligned}
&\left|\overline{\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})}} \cap \left\{\cup_{g=i-1}^{k}\{Y_j : j \in T_g^*\}\right\}\right| \\
&\leq \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \left\{\cup_{g=i}^{k}\{Y_j : j \in T_g^*\}\right\}\right| \\
&\quad + \left|\overline{\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})}} \cap \{Y_j : j \in T_{i-1}^*\}\right| \\
&\leq (k-i+2)m + \frac{5\sum_{g=i-1}^{k} n_g^*}{4\log(1/G(\Delta/(16k\sigma)))}.
\end{aligned}$$

This concludes the verification of (Q4) for the induction level $i-1$. This also concludes the proof of the induction results.

In view of the induction arguments, we have that

$$\mathsf{dist}_i^{(\bar{\ell}_i)} \leq \frac{\Delta}{8k}, \quad i = k, k-1, \ldots, 2. \qquad (92)$$

Finally, to complete the proof of Lemma 30 it remains to show that $\mathsf{dist}_1^{(\ell_2)} \leq \frac{\Delta}{8k}$. In view of the induction statement we have that

1)

$$\begin{aligned}
&\left|\overline{\mathcal{P}_2^{(\bar{\ell}_2)}} \cap \{Y_j : j \in T_1^*\}\right| \\
&\geq n_1^* - (k-1)n\beta - \frac{5(k-1)n_1^*}{4\log(1/(G(\Delta/16\sigma k)))}.
\end{aligned}$$

2)

$$\begin{aligned}&\left|\overline{\mathcal{P}_2^{(\bar{\ell}_2)}}\cap\left\{\cup_{g=2}^k\{Y_j:j\in T_g^*\}\right\}\right|\\&\le(k-1)m+\frac{5\sum_{g=2}^k n_g^*}{4\log(1/G(\Delta/(16\sigma k)))}\\&\le\frac{n\alpha\beta}{8k}+\frac{5\sum_{g=2}^k n_g^*}{4\log(1/G(\Delta/(16\sigma k)))}.\end{aligned}$$

According to Algorithm 5, in the final stage, to find $\mu_1^{(1,\bar{\ell}_2)}$ we deploy the $\mathsf{HDP}_{1-\beta}$ algorithm. In view of

$$\left|\{Y_j:j\in T_1^*\}/\mathcal{B}\left(\theta_1,\frac{\Delta}{16k}\right)\right|\le\frac{5n_1^*}{4\log(1/(G(\Delta/16\sigma k)))}$$

from Lemma 27, we have

$$\begin{aligned}&\left|\overline{\mathcal{P}_2^{(\bar{\ell}_2)}}/\mathcal{B}\left(\theta_1,\frac{\Delta}{16k}\right)\right|\\&=\left|\left\{\overline{\mathcal{P}_2^{(\bar{\ell}_2)}}\cap\{Y_j:j\in T_1^*\}/\mathcal{B}\left(\theta_1,\frac{\Delta}{16k}\right)\right\}\right.\\&\quad\left.\cup\left\{\overline{\mathcal{P}_2^{(\bar{\ell}_2)}}\cap\left\{\cup_{g=2}^k\{Y_j:j\in T_g^*\}\right\}/\mathcal{B}\left(\theta_1,\frac{\Delta}{16k}\right)\right\}\right|+n^{\mathsf{out}}\\&\le\left|\{Y_j:j\in T_1^*\}/\mathcal{B}\left(\theta_1,\frac{\Delta}{16k}\right)\right|\\&\quad+\left|\overline{\mathcal{P}_2^{(\bar{\ell}_2)}}\cap\left\{\cup_{g=2}^k\{Y_j:j\in T_g^*\}\right\}\right|+n^{\mathsf{out}}\\&\le\frac{5n_1^*}{4\log(1/(G(\Delta/16\sigma k)))}+(k-1)\frac{n\beta^2}{2}\\&\quad+\frac{5\sum_{g=2}^k n_g^*}{4\log(1/G(\Delta/(16\sigma k)))}+n^{\mathsf{out}}\\&\le\frac{n\alpha\beta}{8k}+\frac{5n}{2\log(1/(G(\Delta/(16\sigma k))))}+n^{\mathsf{out}}\le\frac{n_1^*\beta}{4},\end{aligned}$$

where for the last inequality we assume that

$$\Delta\ge16k\sigma G^{-1}\left(c^{-\frac{40}{\alpha\beta}}\right),\quad n^{\mathsf{out}}\le\frac{n\alpha\beta}{16k}.\tag{93}$$

As we have

$$\left|\overline{\mathcal{P}_2^{(\bar{\ell}_2)}}\right|\ge\left|\overline{\mathcal{P}_2^{(\bar{\ell}_2)}}\cap\{Y_j:j\in T_1^*\}\right|\ge\frac{n_1^*}{2},\tag{94}$$

any subset of $\overline{\mathcal{P}_2^{(\bar{\ell}_2)}}$ with size $(1-\beta)\left|\overline{\mathcal{P}_2^{(\bar{\ell}_2)}}\right|$, discards a set of size at least $\frac{n_1^*\beta}{2}$ from $\overline{\mathcal{P}_2^{(\bar{\ell}_2)}}$. Hence the tightest subset of $\overline{\mathcal{P}_2^{(\bar{\ell}_2)}}$ with size $(1-\beta)\left|\overline{\mathcal{P}_1^{(\ell)}}\right|$ will have a diameter of at most $\frac{\Delta}{16k}$. This implies $\mathsf{dist}_1^{(\bar{\ell}_2)}\le\frac{\Delta}{16k}$. In view of (92) this proves that there is a path of indices $\bar{\ell}_k,\dots\bar{\ell}_2$ such that $\mathsf{dist}_1^{(\bar{\ell}_2)}+\sum_{h=2}^k\mathsf{dist}_h^{(\bar{\ell}_h)}\le\frac{\Delta}{8}$. Hence, when we pick the indices to optimize totdist, we get $\min_{\ell_k}\mathsf{totdist}_k^{(\ell_k)}\le\frac{\Delta}{8}$, as required. □

Proof of Lemma 30. Note that the term $\mathsf{totdist}_k^{(\ell_k)}$ is given by $\sum_{i=1}^k\mathsf{dist}_i^{(\ell_i)}$ for some sequence of indices originating from the inbuilt *for-loops* at different levels $\ell_k,\dots,\ell_2$ and $\ell_2=\ell_1$. Hence it suffices to prove that if the sum $\sum_{i=1}^k\mathsf{dist}_i^{(\ell_i)}$ is smaller than $\frac{\Delta}{8}$ for any sequence of the loop counts we have good centroid approximations. This is summarized in the following result.

Lemma 33. *Suppose that for a sequence of indices $\ell_1,\dots,\ell_k$ we have $\sum_{i=1}^k\mathsf{dist}_i^{(\ell_i)}\le\frac{\Delta}{8}$. Then if the corresponding centroids are $\{\mu_i^{(i,\ell_i)}\}_{i=1}^k$, with $\ell_2=\ell_1$, we get that there is a permutation π of $[k]$ such that $\mu_i^{i,\ell_i}\in\mathcal{B}(\theta_{\pi(i)},\Delta/3)$ for each $i\in1,\dots,k$.*

Proof. First we show that all of the centroids lie in $\cup_{i=1}^{k}\mathcal{B}(\theta_i, \Delta/3)$. If not, without a loss of generality let $\mu_1^{(1,\ell_1)}$ lie outside $\cup_{i=1}^{k}\mathcal{B}(\theta_i, \Delta/3)$. Then we have

$$\begin{aligned}
&\left|\{Y_i : i \in [n]\} \cap \mathcal{B}\left(\mu_1^{(1,\ell_1)}, \frac{\Delta}{8}\right)\right| \\
&\leq \left|\{Y_i : i \in [n]\}/ \cup_{i=1}^{k} \mathcal{B}(\theta_i, \frac{\Delta}{8})\right| \\
&\leq \frac{5n}{4\log(1/(G(\Delta/16\sigma k)))} + n^{\mathsf{out}},
\end{aligned} \tag{95}$$

where the last inequality followed from Lemma 27. For an ease of notation, throughout the proof we define

$$\ell_1 \triangleq \ell_2, \quad \mathcal{P}_1^{(\ell_1)} \triangleq \overline{\mathcal{P}_1^{(\ell_2)}}, \quad \mu_1^{(1,\ell_1)} \triangleq \mu_1^{(1,\ell_2)}. \tag{96}$$

Note that in terms of the indices $\ell_2, \ldots, \ell_k$ we have the partition of $\{Y_i : i \in [n]\}$ as

$$\{Y_i : i \in [n]\} = \cup_{g=1}^{k}\mathcal{P}_g^{(\ell_g)}, \quad \mathcal{P}_g^{(\ell_g)} \cap \mathcal{P}_h^{(\ell_h)} = \phi, g \neq h \in [k]. \tag{97}$$

In view of the assumption $\mathsf{dist}_1^{(\ell_1)} \leq \frac{\Delta}{8}$ and the fact that the $\left|\mathcal{P}_1^{(\ell_1)} \cap \mathcal{B}\left(\mu_1^{(1,\ell_1)}, \mathsf{dist}_1^{(\ell_1)}\right)\right| \geq (1-\beta)\left|\mathcal{P}_1^{(\ell_1)}\right|$ we have

$$\left|\mathcal{P}_1^{(\ell_1)}/\mathcal{B}\left(\mu_1^{(1,\ell_1)}, \frac{\Delta}{8}\right)\right| \leq \left|\mathcal{P}_1^{(\ell_1)}/\mathcal{B}\left(\mu_1^{(1,\ell_1)}, \mathsf{dist}_1^{(\ell_1)}\right)\right| \leq n\beta. \tag{98}$$

In view of (95) the last display implies

$$\begin{aligned}
&\left|\mathcal{P}_1^{(\ell_1)}\right| \\
&\leq \left|\mathcal{P}_1^{(\ell_1)}/\mathcal{B}\left(\mu_1^{(1,\ell_1)}, \frac{\Delta}{8}\right)\right| \\
&\quad + \left|\{Y_i : i \in [n]\} \cap \mathcal{B}\left(\mu_1^{(1,\ell_1)}, \frac{\Delta}{8}\right)\right| \\
&\leq n\beta + \frac{3n}{4\log(1/(G(\Delta/16\sigma k)))} + n^{\mathsf{out}}.
\end{aligned}$$

As $\cup_{i=2}^{k}\mathcal{P}_i^{(\ell_i)}$ and $\mathcal{P}_1^{(\ell_1)}$ are disjoint and their union covers all the data points, the last display implies for any $j = 2, \ldots, k$

$$\begin{aligned}
&\left|\{Y_i : i \in T_j^*\} \cap \left\{\cup_{g=2}^{k}\mathcal{P}_g^{(\ell_g)}\right\}\right| \\
&= \left|\{Y_i : i \in T_j^*\}/\mathcal{P}_1^{(\ell_1)}\right| \\
&\geq n_j^* - n\beta - \frac{5n}{4\log(1/(G(\Delta/16\sigma k)))} - n^{\mathsf{out}} \geq \frac{7n\alpha}{8k},
\end{aligned} \tag{99}$$

where the last inequality follows from $\beta \leq \frac{\alpha}{12k}$ as $k \geq 3$, $\Delta \geq 16\sigma k G^{-1}\left(e^{-\frac{5k^2}{\alpha}}\right)$ and we assume that

$$n^{\mathsf{out}} \leq \frac{n\beta}{2}. \tag{100}$$

Then we have for $j = 1, \ldots, k$

$$\begin{aligned}
&\left|\{Y_i : i \in T_j^*\} \cap \left[\cup_{g=2}^k \left\{\mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\}\right]\right| \\
&= \Bigg|\{Y_i : i \in T_j^*\} \cap \left\{\cup_{g=2}^k \mathcal{P}_g^{(\ell_g)}\right\} \\
&\quad \cap \left[\cup_{g=2}^k \left\{\mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\}\right]\Bigg| \\
&= \left|\{Y_i : i \in T_j^*\} \cap \left\{\cup_{g=2}^k \mathcal{P}_g^{(\ell_g)}\right\}\right| \\
&\quad - \Bigg|\{Y_i : i \in T_j^*\} \cap \left\{\cup_{g=2}^k \mathcal{P}_g^{(\ell_g)}\right\} / \\
&\quad \left[\cup_{g=2}^k \left\{\mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\}\right]\Bigg| \\
&\geq \left|\{Y_i : i \in T_j^*\} \cap \left\{\cup_{g=2}^k \mathcal{P}_g^{(\ell_g)}\right\}\right| \\
&\quad - \left|\left\{\cup_{g=2}^k \mathcal{P}_g^{(\ell_g)}\right\} / \left[\cup_{g=2}^k \left\{\mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\}\right]\right| \\
&\overset{(a)}{\geq} \frac{7n\alpha}{8k} - \left|\cup_{g=2}^k \left\{\mathcal{P}_g^{(\ell_g)} / \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\}\right| \\
&= \frac{7n\alpha}{8k} - \sum_{g=2}^k \left|\mathcal{P}_g^{(\ell_g)} / \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right| \\
&\overset{(b)}{\geq} \frac{7n\alpha}{8k} - n\beta \geq \frac{3n\alpha}{4k},
\end{aligned} \tag{101}$$

where (a) followed (99) and the fact that $\{\mathcal{P}_g^{(\ell_g)}\}_{g=2}^k$ are disjoint and (b) followed from as

$$\sum_{g=2}^k \left|\mathcal{P}_g^{(\ell_g)} / \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right| \leq \beta \sum_{g=2}^k \left|\mathcal{P}_g^{(\ell_g)}\right| \leq n\beta.$$

As we have for each $j = 1, \ldots, k$

$$\begin{aligned}
&\{Y_i : i \in T_j^*\} \cap \left[\cup_{g=2}^k \left\{\mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\}\right] \\
&= \cup_{g=2}^k \left\{\{Y_i : i \in T_j^*\} \cap \mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\}
\end{aligned}$$

with the union on right side of the above display is disjoint, by the pigeon hole principle there exist indices g, j_1, j_2 such that

$$\begin{aligned}
&\left|\left\{\mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\} \cap \{Y_i : i \in T_j^*\}\right| \\
&\geq \frac{\min_{j=1}^k \left|\{Y_i : i \in T_j^*\} \cap \left[\cup_{g=2}^k \left\{\begin{smallmatrix}\mathcal{P}_g^{(\ell_g)} \cap \\ \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\end{smallmatrix}\right\}\right]\right|}{k} \\
&\geq \frac{3n\alpha}{4k^2},
\end{aligned}$$

for $j = j_1, j_2$, where the last display followed using (101). However, for $j = j_1, j_2$, as $\Delta \geq 16\sigma k G^{-1}\left(e^{-\frac{5k^2}{\alpha}}\right)$ implies

$$\left|\{Y_i : i \in T_j^*\} / \mathcal{B}(\theta_j, \Delta/3)\right| \leq \frac{5n_j^*}{4\log(1/G(\Delta/16\sigma k))} \leq \frac{n\alpha}{4k^2},$$

we get that

$$\begin{aligned}
&\left|\left\{\mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\} \cap \{Y_i : i \in T_j^*\} \cap \mathcal{B}(\theta_j, \Delta/3)\right| \\
&\geq \left|\left\{\mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\} \cap \{Y_i : i \in T_j^*\}\right| \\
&\quad - \left|\{Y_i : i \in T_j^*\} / \mathcal{B}(\theta_j, \Delta/3)\right| \geq \frac{n\alpha}{2k^2}.
\end{aligned}$$

Hence, there exists $x, y \in \{Y_i : i \in [n]\}$ such that

$$\begin{aligned} x \in & \left\{\mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\} \\ & \cap \{Y_i : i \in T_{j_1}^*\} \cap \mathcal{B}(\theta_{j_1}, \Delta/3), \\ y \in & \left\{\mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\} \\ & \cap \{Y_i : i \in T_{j_2}^*\} \cap \mathcal{B}(\theta_{j_2}, \Delta/3). \end{aligned}$$

As we have $\mathsf{dist}_g^{(\ell_g)} \leq \frac{\Delta}{8}$ and $\|\theta_{j_1} - \theta_{j_2}\| \geq \Delta$ we get a contradiction

$$\begin{aligned} \|x - y\| &\geq \|\theta_{j_1} - \theta_{j_2}\| - \|x - \theta_{j_1}\| - \|y - \theta_{j_2}\| \geq \frac{\Delta}{3}. \\ \|x - y\| &\leq \|x - \mu_g^{(g,\ell_g)}\| + \|y - \mu_g^{(g,\ell_g)}\| \leq 2\mathsf{dist}_g^{(\ell_g)} \leq \frac{\Delta}{4}. \end{aligned}$$

Hence, all of the centroids lie in $\cup_{i=1}^k \mathcal{B}(\theta_i, \Delta/3)$.

Now it remains to show that $\left\{\mu_g^{(g,\ell_g)}\right\}_{g=1}^k$ lie in different balls among $\{\mathcal{B}(\theta_g, \Delta/3)\}_{g=1}^k$. If not, then without a loss of generality let $\mathcal{B}\left(\theta_1, \frac{\Delta}{3}\right)$ contains two of the centroids, say $\mu_{j_1}^{(j_1,\ell_{j_1})}, \mu_{j_2}^{(j_2,\ell_{j_2})}$. Also, as $\mathcal{B}\left(\theta_1, \frac{\Delta}{3}\right)$ contains two centroids, by the pigeonhole principle we get that there is an index $g \neq 1$ such that

$$\mu_j^{(j,\ell_j)} \notin \mathcal{B}\left(\theta_g, \frac{\Delta}{3}\right), \quad j = 1, \ldots, k.$$

In view of the disjoint union $\left\{\mathcal{B}\left(\theta_g, \frac{\Delta}{8}\right)\right\}_{g=1}^k$,and $\mathsf{dist}_j^{(\ell_j)} \leq \frac{\Delta}{8}$ the above implies

$$\mathcal{B}\left(\theta_g, \frac{\Delta}{8}\right) \cap \mathcal{B}\left(\mu_j^{(j,\ell_j)}, \mathsf{dist}_j^{(\ell_j)}\right) = \phi, \quad j = 1, \ldots, k.$$

Note that by Lemma 27

$$\left|\{Y_i : i \in [n]\} \cap \mathcal{B}\left(\theta_g, \frac{\Delta}{8}\right)\right| \geq \frac{n\alpha}{2k}. \tag{102}$$

As the disjoint union $\cup_{m=1}^k \mathcal{P}_m^{(\ell_m)}$ is the entire set of data points, we get

$$\begin{aligned} & \{Y_i : i \in [n]\} \cap \mathcal{B}\left(\theta_g, \frac{\Delta}{8}\right) \\ &= \left\{\cup_{j=1}^k \mathcal{P}_j^{(\ell_j)}\right\} \cap \mathcal{B}\left(\theta_g, \frac{\Delta}{8}\right) \\ &= \cup_{j=1}^k \left\{\mathcal{P}_j^{(\ell_j)} \cap \mathcal{B}\left(\theta_g, \frac{\Delta}{8}\right)\right\} \\ &\subseteq \cup_{j=1}^k \left\{\mathcal{P}_j^{(\ell_j)} / \mathcal{B}\left(\mu_j^{(j,\ell_j)}, \mathsf{dist}_j^{(\ell_j)}\right)\right\}, \end{aligned}$$

which implies

$$\begin{aligned} & \left|\{Y_i : i \in [n]\} \cap \mathcal{B}\left(\theta_g, \frac{\Delta}{8}\right)\right| \\ &\leq \sum_{j=1}^k \left\{\mathcal{P}_j^{(\ell_j)} / \mathcal{B}\left(\mu_j^{(j,\ell_j)}, \mathsf{dist}_j^{(\ell_j)}\right)\right\} \\ &\leq \sum_{j=1}^k \beta \left|\mathcal{P}_j^{(\ell_j)}\right| \leq n\beta = \frac{n\alpha}{4k^2}. \end{aligned}$$

This provides a contradiction to (102). Hence, all the centroids must lie in different $\mathcal{B}\left(\theta_j, \frac{\Delta}{3}\right)$ sets. □

□

Proof of Lemma 32. In view of Lemma 27, there is a constant $c_1 = \sigma G^{-1}\left(e^{-\frac{5}{4\beta^2}}\right)$ such that

$$|\{Y_j : j \in T_h^*\}/\mathcal{B}(\theta_h, c_1)| \leq n_h^* \beta^2, \quad h \in [k]. \tag{103}$$

Hence we get that for $h \in [i-1]$

$$\begin{aligned}
&\left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_h^*\} \cap \mathcal{B}(\theta_h, c_1)\right| \\
&= \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_h^*\}\right| \\
&\quad - \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_h^*\}/\mathcal{B}(\theta_h, c_1)\right| \\
&\geq \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_h^*\}\right| - |\{Y_j : j \in T_h^*\}/\mathcal{B}(\theta_h, c_1)| \\
&\geq \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_h^*\}\right| - n_h^*\beta^2 \\
&\geq n_h^* - (k-i+1)n\beta - \frac{5(k-i+1)n_h^*}{4\log(1/(G(\Delta/16\sigma k)))} - n_h^*\beta^2 \\
&\geq m_1,
\end{aligned} \tag{104}$$

where the last inequality holds whenever $\Delta \geq 16\sigma k G^{-1}\left(e^{-\frac{5k}{\alpha}}\right)$ as $(k-i+1)n\beta, n\beta^2 \leq \frac{n\alpha}{4k}$. As we have from the lemma statement

$$\left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \left\{\cup_{g=i}^k \{Y_j : j \in T_g^*\}\right\}\right| \leq \frac{2n\alpha\beta}{5k} \leq \frac{m_1}{2}, \tag{105}$$

we get that the tightest neighborhood around any point in $\overline{\mathcal{P}_i^{(\bar{\ell}_i)}}$ with a size m_1 will have a radius of at most $2c_1$ around that Y_i. Let $\mu_{i-1}^{(i-1,1)} = Y_{i^*}$ be the chosen centroid. Hence $\left|\left\{\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \mathcal{B}(Y_{i^*}, 2c_1)\right\}\right| \geq m_1$. Then $\mathcal{B}(Y_{i^*}, 2c_1)$ and $\cup_{j\in[i-1]}\mathcal{B}(\theta_j, c_1)$ can not be disjoint, as in view of (104) it will imply that

$$\begin{aligned}
&\left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}}\right| \\
&\geq \left|\left\{\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \mathcal{B}(Y_{i^*}, 2c_1)\right\} \cup \left\{\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \left\{\cup_{h\in[i-1]}\mathcal{B}(\theta_h, c_1)\right\}\right\}\right| \\
&\overset{(a)}{\geq} m_1 + \sum_{h\in[i-1]} \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_h^*\} \cap \mathcal{B}(\theta_h, c_1)\right| \\
&\overset{(b)}{\geq} m_1 - n\beta^2 + \sum_{h\in[i-1]} \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_h^*\}\right| \\
&= m_1 - n\beta^2 + \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \left[\cup_{h\in[i-1]}\{Y_j : j \in T_h^*\}\right]\right| \\
&\geq m_1 - n\beta^2 + \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}}\right| \\
&\quad - \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \left\{\cup_{g=i}^k \{Y_j : j \in T_g^*\}\right\}\right| - n^{\mathsf{out}} \\
&\overset{(c)}{=} \frac{m_1}{2} - n\beta^2 + \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}}\right| - n^{\mathsf{out}} \overset{(d)}{\geq} \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}}\right| + \frac{m_1}{8}.
\end{aligned}$$

where (a) follows from the fact that $\{i \in [n] : Y_i \in \mathcal{B}(\theta_j, c_1)\}$ are disjoint sets for $j \in [k]$ as $\min_{g\neq h\in[k]} \|\theta_g - \theta_h\| \geq \Delta$, (b) follows from (105), (c) followed from (104) and (d) follows from assuming

$$n^{\mathsf{out}} \leq \frac{n\alpha}{32k} \leq \frac{m_1}{8}. \tag{106}$$

Hence, Y_{i^*} is at a distance at most $3c_1$ from one of the centroids. Without a loss of generality, we can choose the closest centroid to be θ_{i-1}. □

Requirement on n^{out}: To summarize how many outliers our initialization technique can tolerate, we combine (67),(93),(100),(106) to get

$$n^{\mathsf{out}} \leq \min\left\{\frac{n\alpha}{16k}, \frac{n\alpha\beta}{16k}, \frac{n\beta}{2}, \frac{n\alpha}{32k}\right\} = \frac{n\alpha^2}{64k^3}.$$

ACKNOWLEDGMENT

Soham Jana thanks Debarghya Mukherjee and Sohom Bhattacharya for helpful discussions at the initial stage of the paper.

REFERENCES

[1] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.
[2] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.
[3] A. A. Abbasi and M. Younis, "A survey on clustering algorithms for wireless sensor networks," *Computer Communications*, vol. 30, no. 14, pp. 2826–2841, 2007, network Coverage and Routing Schemes for Wireless Sensor Networks. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366407002162
[4] P. Sasikumar and S. Khara, "K-means clustering in wireless sensor networks," in *2012 Fourth international conference on computational intelligence and communication networks*. IEEE, 2012, pp. 140–144.
[5] C. Di Nuzzo and S. Ingrassia, "Three-way spectral clustering. in "brito p., dias jg, lausen b., montanari a., nugent r.(eds.) classification and data science in the digital age"," *Studies in Classification, Data Analysis, and Knowledge Organization, Springer*, pp. 111–118, 2023.
[6] C. D. Maravelias, "Habitat selection and clustering of a pelagic fish: effects of topography and bathymetry on species dynamics," *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 56, no. 3, pp. 437–450, 1999.
[7] S. Pigolotti, C. López, and E. Hernández-García, "Species clustering in competitive lotka-volterra models," *Physical review letters*, vol. 98, no. 25, p. 258101, 2007.
[8] H. Ng, S. Ong, K. Foong, P.-S. Goh, and W. Nowinski, "Medical image segmentation using k-means clustering and improved watershed algorithm," in *2006 IEEE southwest symposium on image analysis and interpretation*. IEEE, 2006, pp. 61–65.
[9] A. Ajala Funmilola, O. Oke, T. Adedeji, O. Alade, and E. Adewusi, "Fuzzy kc-means clustering algorithm for medical image segmentation," *Journal of information Engineering and Applications, ISSN*, vol. 22245782, pp. 2225–0506, 2012.
[10] A. Beatty, S. Liao, and J. J. Yu, "The spillover effect of fraudulent financial reporting on peer firms' investments," *Journal of Accounting and Economics*, vol. 55, no. 2-3, pp. 183–205, 2013.
[11] J. Fan, Q. Liu, B. Wang, and K. Zheng, "Unearthing financial statement fraud: Insights from news coverage analysis." *Management Science (2025+). Available at SSRN 4338277*, 2025.
[12] Y. Lu and H. H. Zhou, "Statistical and computational guarantees of lloyd's algorithm and its variants," *arXiv preprint arXiv:1612.02099*, 2016.
[13] M. Löffler, A. Y. Zhang, and H. H. Zhou, "Optimality of spectral clustering in the gaussian mixture model," *The Annals of Statistics*, vol. 49, no. 5, pp. 2506–2530, 2021.
[14] E. Abbe, J. Fan, and K. Wang, "An ℓ_p theory of pca and spectral clustering," *The Annals of Statistics*, vol. 50, no. 4, pp. 2359–2385, 2022.
[15] M. Ndaoud, "Sharp optimal recovery in the two component gaussian mixture model," *The Annals of Statistics*, vol. 50, no. 4, pp. 2096–2126, 2022.
[16] M. Dreveton, A. Gözeten, M. Grossglauser, and P. Thiran, "Universal lower bounds and optimal rates: Achieving minimax clustering error in sub-exponential mixture models," in *The Thirty Seventh Annual Conference on Learning Theory*. PMLR, 2024, pp. 1451–1485.
[17] C. Di Nuzzo and S. Ingrassia, "A mixture model approach to spectral clustering and application to textual data," *Statistical methods & applications*, vol. 31, no. 5, pp. 1071–1097, 2022.
[18] X. Chen and Y. Yang, "Cutoff for exact recovery of gaussian mixture models," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 4223–4238, 2021.
[19] A. Kumar, Y. Sabharwal, and S. Sen, "A simple linear time (1+/spl epsiv/)-approximation algorithm for k-means clustering in any dimensions," in *45th Annual IEEE Symposium on Foundations of Computer Science*. IEEE, 2004, pp. 454–462.
[20] X. Chen and A. Y. Zhang, "Achieving optimal clustering in gaussian mixture models with anisotropic covariance structures," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
[21] S. Vassilvitskii and D. Arthur, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2006, pp. 1027–1035.
[22] D. Patel, H. Shen, S. Bhamidi, Y. Liu, and V. Pipiras, "Consistency of lloyd's algorithm under perturbations," *arXiv preprint arXiv:2309.00578*, 2023.
[23] A. Deshpande, P. Kacham, and R. Pratap, "Robust k-means++," in *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 799–808.
[24] A. Bhaskara, S. Vadgama, and H. Xu, "Greedy sampling for approximate clustering in the presence of outliers," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[25] R. Kannan, S. Vempala *et al.*, "Spectral algorithms," *Foundations and Trends® in Theoretical Computer Science*, vol. 4, no. 3–4, pp. 157–288, 2009.
[26] A. Bojchevski, Y. Matkovic, and S. Günnemann, "Robust spectral clustering for noisy data: Modeling sparse corruptions improves latent embeddings," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 737–746.
[27] Y. Zhang and K. Rohe, "Understanding regularized spectral clustering via graph conductance," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
[28] P. R. Srivastava, P. Sarkar, and G. A. Hanasusanto, "A robust spectral clustering algorithm for sub-gaussian mixture models with outliers," *Operations Research*, vol. 71, no. 1, pp. 224–244, 2023.
[29] S. Jana, K. Yang, and S. Kulkarni, "Adversarially robust clustering with optimality guarantees," *arXiv preprint arXiv:2306.09977*, 2023.
[30] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.
[31] P. Bradley, O. Mangasarian, and W. Street, "Clustering via concave minimization," *Advances in neural information processing systems*, vol. 9, 1996.
[32] S. Minsker and N. Strawn, "The geometric median and applications to robust mean estimation," *SIAM Journal on Mathematics of Data Science*, vol. 6, no. 2, pp. 504–533, 2024.
[33] M. Chen, C. Gao, and Z. Ren, "Robust covariance and scatter matrix estimation under huber's contamination model," *The Annals of Statistics*, vol. 46, no. 5, pp. 1932–1960, 2018.
[34] I. Diakonikolas, D. M. Kane, D. Kongsgaard, J. Li, and K. Tian, "Clustering mixture models in almost-linear time via list-decodable mean estimation," in *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, 2022, pp. 1262–1275.
[35] J. Depersin and G. Lecué, "Robust sub-gaussian estimation of a mean vector in nearly linear time," *The Annals of Statistics*, vol. 50, no. 1, pp. 511–536, 2022.
[36] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira, "Sub-gaussian mean estimators," 2016.
[37] R. I. Oliveira and L. Resende, "Trimmed sample means for robust uniform mean estimation and regression," *arXiv preprint arXiv:2302.06710*, 2023.
[38] J. C. Lee and P. Valiant, "Optimal sub-gaussian mean estimation in very high dimensions," in *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2022.
[39] G. Lugosi and S. Mendelson, "Robust multivariate mean estimation: the optimality of trimmed mean," 2021.
[40] ——, "Near-optimal mean estimators with respect to general norms," *Probability theory and related fields*, vol. 175, no. 3, pp. 957–973, 2019.
[41] ——, "Mean estimation and regression under heavy-tailed distributions: A survey," *Foundations of Computational Mathematics*, vol. 19, no. 5, pp. 1145–1190, 2019.
[42] J.-Y. Audibert and O. Catoni, "Robust linear least squares regression," 2011.
[43] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[44] P. Rousseeuw and P. Kaufman, "Clustering by means of medoids," in *Proceedings of the statistical data analysis based on the L1 norm conference, neuchatel, switzerland*, vol. 31, 1987.
[45] P. J. Huber, "A robust version of the probability ratio test," *The Annals of Mathematical Statistics*, pp. 1753–1758, 1965.
[46] ——, "Robust estimation of a location parameter," in *Breakthroughs in statistics: Methodology and distribution*. New York, NY: Springer, 1992, pp. 492–518.
[47] I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart, "Robust estimators in high-dimensions without the computational intractability," *SIAM Journal on Computing*, vol. 48, no. 2, pp. 742–864, 2019.
[48] A. Liu and A. Moitra, "Robustly learning general mixtures of gaussians," *Journal of the ACM*, 2023.
[49] S. Minsker, M. Ndaoud, and Y. Shen, "Minimax supervised clustering in the anisotropic gaussian mixture model: a new take on robust interpolation," *arXiv preprint arXiv:2111.07041*, 2021.
[50] I. Diakonikolas, S. B. Hopkins, D. Kane, and S. Karmalkar, "Robustly learning any clusterable mixture of gaussians," *arXiv preprint arXiv:2005.06417*, 2020.
[51] A. Bakshi, I. Diakonikolas, H. Jia, D. M. Kane, P. K. Kothari, and S. S. Vempala, "Robustly learning mixtures of k arbitrary gaussians," in *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, 2022, pp. 1234–1247.
[52] A. Bakshi and P. Kothari, "Outlier-robust clustering of non-spherical mixtures," *arXiv preprint arXiv:2005.02970*, 2020.
[53] M. G. Omran, A. P. Engelbrecht, and A. Salman, "An overview of clustering methods," *Intelligent Data Analysis*, vol. 11, no. 6, pp. 583–605, 2007.
[54] D. Hsu, S. Kakade, and T. Zhang, "A tail inequality for quadratic forms of subgaussian random vectors," *Electronic communications in Probability*, vol. 17, pp. 1–6, 2012.
[55] J. Fan and Y. Fan, "High dimensional classification using features annealed independence rules," *Annals of statistics*, vol. 36, no. 6, p. 2605, 2008.
[56] D. Slate, "Letter Recognition," UCI Machine Learning Repository, 1991, DOI: https://doi.org/10.24432/C5ZP40.
[57] B. Yu, "Assouad, fano, and le cam," in *Festschrift for Lucien Le Cam: research papers in probability and statistics*. Springer, 1997, pp. 423–435.
[58] S. Roch, *Modern discrete probability: An essential toolkit*. Cambridge University Press, 2024.
[59] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Soham Jana received the Ph.D. degree from Yale University. He was a postdoctoral researcher at Princeton University. He joined the University of Notre Dame, USA, in 2024, and is currently an Assistant Professor at the Department of Applied and Computational Mathematics and Statistics. His research interests include machine learning, statistics, and signal/image processing.

Jianqing Fan received the Ph.D. degree from the University of California at Berkeley. He was appointed as an Assistant Professor, an Associate Professor, and a Professor at the University of North Carolina at Chapel Hill, the University of California at Los Angeles, the Chinese University of Hong Kong, respectively. He was the past president of the Institute of Mathematical Statistics and the International Chinese Statistical Association. He is currently the Frederick L. Moore Professor of Princeton University. His research interests include statistics, machine learning, data science, economics, finance, and computational biology. He was a Co-Editor of Annals of Statistics, Probability Theory and Related Fields, Journal of Econometrics, and Journal of Business and Economics Statistics. He is currently co-editing Journal of the American Statistical Association.

Sanjeev Kulkarni (Fellow, IEEE) received the Ph.D. degree from MIT. He joined Princeton University, in 1991. He was the Dean of the Graduate School from 2014 to 2017 and the Dean of the Faculty from 2017 to 2021. He is currently the William R. Kenan Junior Professor of Princeton University. He is jointly affiliated with the Department of Electrical and Computer Engineering and the Department of Operations Research and Financial Engineering. His research interests include machine learning, statistics, information theory, and signal/image processing.