A GENERAL THEORY FOR ROBUST CLUSTERING VIA TRIMMED MEAN

BY SOHAM JANA[1], JIANQING FAN*[1], AND SANJEEV KULKARNI[1]

[1]*Department of Operations Research and Financial Engineering, Princeton University*

Clustering is a fundamental tool in statistical machine learning in the presence of heterogeneous data. Many 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 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 the optimal mislabeling even in the presence of adversarial outliers. Our results reduce to the sub-Gaussian case 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 the Lloyd algorithm is suboptimal for more than two clusters even when errors are Gaussian, and for two clusters when errors distributions have heavy tails. Both simulated data and real data examples lend further support to both of our robust initialization procedure and clustering algorithm.

1. Introduction.

1.1. *Problem.* Clustering is an essential task in statistics and machine learning (Hastie et al., 2009; Xu and Wunsch, 2005) that has diverse practical applications (e.g., wireless networks (Abbasi and Younis, 2007; Sasikumar and Khara, 2012), grouping biological species (Maravelias, 1999; Pigolotti, López and Hernández-García, 2007), medical imaging (Ng et al., 2006; Ajala Funmilola et al., 2012) and defining peer firms in finance (Beatty, Liao and Yu, 2013; Fan et al., 2023)). 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

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

for a given k. Here $Y_1, \dots, Y_n$ are the data, θ_g is the centroid corresponding to the g-th cluster, $z = \{z_1, \dots, z_n\}$ are the unknown labels of the data describing which clusters they belong to, and $w_1, \dots, w_n$ are unknown independent errors. Recent advances in the literature have focused on recovering the labels z. Given any estimate $\widehat{z} = \{\widehat{z}_1, \dots, \widehat{z}_n\}$ of z, define the mislabeling error as the proportion of label estimates that do not match the correct ones: $\ell(z, \widehat{z}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{z_i \neq \widehat{z}_i\}}$. Upon observing the data, the goal is to produce label estimates with a small mislabeling error, equivalent to correctly identifying the inherent cluster

*Partially supported by NSF grants DMS-2210833, DMS-2053832, DMS-2052926 and ONR grant N00014-22-1-2340

MSC2020 subject classifications: Primary 62H30; secondary 62G35,62G05.

Keywords and phrases: Heavy tail, adversarial outliers, initializaiton, mislabeling, adaptive methods.

structure. Accurate clustering is essential for performing many inference tasks relevant to individual clusters, such as parameter estimation and testing, where the consistency guarantees degrade with outlying observations.

Unfortunately, given any fixed set of centroids, even with a large data set, it is only possible to label every point correctly if the error distributions have compact supports and the centroids are well separated. Consequently, given any distributional assumptions on the clusters, it is essential to quantify the mislabeling error, which can serve as a benchmark for comparing the performance of clustering algorithms. Classical works aiming to address the above problem use the minimum separation of centroids $\min_{g\neq h\in[k]}\|\theta_g-\theta_h\|$ (usually denoted by Δ) to characterize the mislabeling error. In essence, given any fixed class of error distributions, we should be able to identify the clusters more accurately as Δ goes to infinity.

In addition to answering the above question, it is essential to construct algorithms that achieve the above mislabeling rate. Major works in this direction mostly study the sub-Gaussian mixture model with prescribed conditions on the covariance matrices. For example, in the Gaussian mixture model, with $\sigma^2>0$ variance in each coordinate, the minimax optimal mislabeling rate (Lu and Zhou, 2016) is approximately $\exp\left(-(1+o(1))\frac{\Delta^2}{8\sigma^2}\right)$. A follow-up work (Chen and Zhang, 2021) also studied the anisotropic covariance structure in this context. In most such works, the clustering techniques involve some variation of Lloyd's algorithm with a good initialization and often use spectral clustering in the initial phase for dimension reduction (Abbe, Fan and Wang, 2022). However, the restrictive sub-Gaussian model is often invalid in practice, and various real-world applications of clustering involve outliers and heavy-tailed error distributions (e.g., image segmentation (Sfikas, Nikou and Galatsanos, 2007), biology (Song et al., 2005; Ronan, Qi and Naegle, 2016) and financial market (de Miranda Cardoso, Ying and Palomar, 2021; Cai, Le-Khac and Kechadi, 2016)). Unfortunately, the Lloyd algorithm and traditional spectral clustering techniques are not well suited to handle such situations. For example, Lloyd's algorithm lacks robustness guarantees due to the use of sample mean in the centroid estimation step, a common problem for mean-based clustering algorithms (Charikar et al., 2001; Olukanmi and Twala, 2017; Gupta et al., 2017). The vanilla version of spectral methods is also vulnerable to noisy setups (Bojchevski, Matkovic and Günnemann, 2017; Zhang and Rohe, 2018). Robust modification for obtaining the mislabeling rate in the presence of adversarial outliers has been previously studied in the literature (Srivastava, Sarkar and Hanasusanto, 2023; Jana, Kulkarni and Yang, 2023), although particularly in the sub-Gaussian settings. Results in the heavy-tail regimes are lacking, which is the main focus of our work. Notably, (Diakonikolas et al., 2022) 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.

Additionally, many clustering techniques, such as the Lloyd algorithm and the Expectation Maximization (EM) algorithm, require a good initial approximation of the clusters to kick off the process. Finding a good initialization in itself is a very challenging task. For instance, the Lloyd algorithm might not produce consistent clustering with a bad initialization (Lu and Zhou, 2016). The widely-used algorithm of (Kumar, Sabharwal and Sen, 2004) for mislabeling analysis in sub-Gaussian clustering (Abbe, Fan and Wang, 2022; Löffler, Zhang and Zhou, 2021; Chen and Zhang, 2021; Srivastava, Sarkar and Hanasusanto, 2023) only provides a good initialization with a small constant probability (Kumar, Sabharwal and Sen, 2004, Theorem 4.1). Another important initialization technique is k-means++ (Arthur and Vassilvitskii, 2007) that starts with a randomly chosen data point as the first centroid and then selects the subsequent initial centroid estimates according to a probability distribution on the observations, which puts more weights on points that are farther away from the existing centroid

approximates. Recently (Patel et al., 2023) used the k-means++ method for centroid initialization; however, the work only addresses the case $k = 2$ with Gaussian error distributions, in which case they show that the centroid initialization is good enough to produce a consistent mislabeling. It is also popular to use spectral methods for centroid initialization; however, theoretical guarantees tend to exist only in the sub-Gaussian setup. For example, (Lu and Zhou, 2016) uses the spectral method presented in (Kannan et al., 2009, 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 exponential decay. We provide a general solution to this initialization problem in our work as well.

1.2. *Our contributions.* In this paper, we characterize the mislabeling rate for a general class of error distributions and provide clustering and initialization algorithms that provably achieve the mislabeling rates in these regimes. In particular, suppose that the error distributions satisfy $\mathbb{P}[\|w_i\| > x] \le G(x/\sigma)$ for $\sigma > 0$ and a decreasing function $G : \mathbb{R}_+ \to [0, 1]$, and the centroids satisfy the separation condition $\min_{g,h\in[k]} \|\theta_g - \theta_h\| \ge \Delta$. Then, we show that the optimal mislabeling rate is given by $G(\Delta/2\sigma)$, up to constant factors depending only on k and the minimum cluster proportion. This matches the results for sub-Gaussian clustering when Δ is larger than $\sigma\sqrt{d}$. To achieve the above mislabeling error, we first construct a clustering algorithm COD (Clustering via Ordered Differences) that achieves the goal, without knowing G, when a weak initialization condition is met. The COD algorithm follows an iterative clustering style, similar to the k-means and k-medians algorithms (Jana, Kulkarni and Yang, 2023). More specifically, given either initial centroid estimates or clustering, we repeat the following steps at each iteration $s \ge 1$:

- *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 trimmed mean estimator.

The trimmed mean estimator we use in our algorithm effectively adapts to any decay distribution G. We also deduce similar results in the presence of adversarial outliers. These outliers can be the data from missing clusters (underspecified k). Additionally, we also provide an adaptive centroid estimation algorithm IOD (Initialization via Ordered Differences), whose guarantees match the initialization condition required by the COD algorithm, leading to an optimal mislabeling error in this broader context. Our algorithm only requires knowledge about a lower bound α on the minimum cluster proportion. The IOD algorithm also allows for the presence of adversarial outliers. Our algorithm uses a recursive technique for finding data points with dense neighborhoods in the data set, which we use as the centroids. The runtime of both IOD and COD algorithms are at most $c_{k,\alpha}(dn^2 + n^2 \log n)$ for some constant c_k that depends on k and α. For comparisons with existing methods, we show that even with good initialization, the Lloyd algorithm fails with heavy-tail error distributions, even when adversarial outliers need not be present.

1.3. *Related works.* There is a long list of work that utilizes a robust centroid estimation technique in clustering. The classical partitioning around the medoid (PAM) algorithm (Kaufman and Rousseeuw, 2009; Rousseeuw and Kaufman, 1987) updates the centroid estimates using a point from the data set (these centroid estimates are referred to as the medoids of the clusters) based on some dissimilarity metric. For example, (Rousseeuw and Kaufman, 1987) used the ℓ_1 distance, and argued the robustness of the corresponding ℓ_1 based PAM

algorithm. The recent work (Jana, Kulkarni and Yang, 2023) uses the coordinatewise median for centroid estimation. However, given their analyses, a consistent clustering with the coordinatewise median will require at least $\sigma\sqrt{d}$ separation of the centroids, with some multiplicative factor depending on the relevant decay function. The coordinatewise median-based technique is suboptimal since given any constant proportion of mislabeling at a fixed step, the centroid estimation error of the coordinatewise median scales with $\sqrt{d}$ (constant approximation error in each coordinate that adds up). In contrast, the minimum requirement for our clustering technique is $\Delta \geq \sigma C_{G,\alpha}$ where $C_{G,\alpha}$ is a particular quantile of the decay distribution and α is the proportion of data in the smallest cluster.

Our work resolves the clustering problem for heavy-tail decay conditions based on the norm of the error vectors. As mentioned above, the minimum centroid separation our theoretical guarantees require depend on the quantiles of the decay functions and the minimum cluster proportion. However, this requirement might not be optimal for specific distribution classes. For example, in the classical sub-Gaussian mixture model, (Lu and Zhou, 2016) showed that for a fixed number of clusters k and balanced cluster sizes, it suffices to have $\Delta/\sigma \to \infty$ for a consistent clustering with a good initialization. In contrast, our results require $\Delta/(\sigma\sqrt{d}) \to \infty$. This can probably be remedied first by using a robust principal component analysis or spectral methods for dimension reduction, e.g., (Wang and Fan, 2022; Srivastava, Sarkar and Hanasusanto, 2023; Bojchevski, Matkovic and Günnemann, 2017), and then performing our robust clustering technique. However, analyses of such methods are left for future work.

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 (Huber, 1965, 1992), 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 (Lugosi and Mendelson, 2021; Diakonikolas et al., 2019) for robust mean estimation. For robust clustering of Gaussian mixtures, (Liu and Moitra, 2023) examines a similar contamination model. However, these works do not study adversarial outliers in the presence of general heavy-tail error distributions, as is being done in this paper.

Many relevant works on clustering (Lu and Zhou, 2016; Jana, Kulkarni and Yang, 2023; Vempala and Wang, 2004) try to produce guarantees for accurately estimating the centroids. In contrast, our clustering technique is geared towards lowering mislabeling and does not guarantee an efficient estimation of the actual centroids. The analysis of our algorithm shows that an unbiased centroid estimation is not crucial for achieving asymptotically optimal mislabeling error if the bias is vanishing compared to the separation of the centroid. Nonetheless, it might be interesting to determine whether incorporating other centroid estimators in our setup can provide better centroid estimation while preserving the mislabeling guarantees. A notable example of centroid estimators that guarantee consistency in the presence of adversarial outliers is Tukey's half-space median. In the Gaussian setup, this median is consistent even in the presence of adversarial outliers and produces optimal error rate (Chen, Gao and Ren, 2018, Theorem 2.1). However, the Tukey's median is computationally expensive. Robust and consistent mean estimation problem with heavy-tailed data has been studied in much existing literature (Fan, Li and Wang, 2017; Sun, Zhou and Fan, 2020; Lugosi and Mendelson, 2019); it might be interesting to see whether incorporating such estimation strategies improve our guarantees.

Another critical related direction is the clustering of 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 (Chen and Zhang, 2021) using a variant of the Lloyd algorithm, and (Diakonikolas et al., 2020; Bakshi et al., 2022) with the target of approximating the mixture distribution in Total Variation distance. Specific heavy-tail regimes with non-spherical mixtures are also discussed in (Bakshi and Kothari, 2020); 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 work in such asymmetric clustering paradigms as well.

1.4. *Organization.* The rest of the paper is organized as follows. In Section 2, we re-introduce our mathematical model and present the clustering algorithm we use, given a good initialization. The theoretical results, i.e., mislabeling rate upper bound under good initialization conditions and the worst case mislabeling lower bound are presented in Section 2.3 (the results involving adversarial outliers are presented in Section 2.4) and Section 2.5 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 3. We present our initialization algorithm and their theoretical guarantees in Section 4. In Section 5 we show that even with good initialization the Lloyd algorithm might produce non-converging mislabeling errors. We demonstrate the effectiveness of our algorithms with application on actual and simulated data sets in Section 6. Finally we present proofs of some lemmas related to the two center initialization problem in Section 7.2. All the other proofs and technical details have been provided in the appendix.

2. Robust clustering under mislabeling guarantees under good initialization.

2.1. *Mixture model.* We introduce again our model here. Fix a monotonically decreasing function 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

(P) $\mathbb{P}\left[\|w\| > \sigma x\right] < G(x)$ for all $x \geq 0$.

For our paper, any monotonic decay condition suffices. We observe independent samples $Y_1, \ldots, Y_n \in \mathbb{R}^d$ from a mixture of k many G_σ distributions as follows:

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

where $z = \{z_i\}_{i=1}^n \in [k]^n$ denote the underlying labels, $\theta_1, \ldots, \theta_k$ are the unknown centers of the data distributions. We study the mislabeling loss function between the estimated labels $\widehat{z} = \{\widehat{z}_i\}_{i=1}^n$ and true labels $z = \{z_i\}_{i=1}^n$ given by

$$\ell(\widehat{z}, z) = \left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\widehat{z}_i \neq z_i\}}\right]. \tag{3}$$

2.2. *Algorithm.* In this section, we present the algorithm with the assumption that either an initial approximation of either the centroids or the labels is available. Our algorithm follows the following two main steps at each iteration $s \geq 1$:

- *Labeling step:* Given an estimate of the centroids $\widehat{\theta}_h^{(s)}$, construct cluster estimates $T_h^{(s)}, h \in \{1, \ldots, k\}$ using a suitable cost function;
- *Estimation step:* For each of the clusters, compute the new centroid estimates $\widehat{\theta}_h^{(s+1)}$ using the points from the estimated cluster $\left\{Y_i : i \in T_h^{(s)}\right\}$.

This process is iterated until the clusters remain the same over subsequent iterations. For the labeling step, we rely on the Euclidean distance based Voronoi tessellations of the data points: Once centroid estimates $(\widehat{\theta}_1^{(s)}, \ldots, \widehat{\theta}_k^{(s)})$ are available, we compute the label estimates for Y_i as $\widehat{z}_i^{(s)} = \operatorname{argmin}_{h \in \{1,\ldots,k\}} \|Y_i - \widehat{\theta}_h^{(s)}\|$, with ties broken arbitrarily. For the centroid estimation step, we use a novel multivariate trimmed mean algorithm (TM_δ) based on the ordered distances between all points in the estimated clusters; see Algorithm 1.

Algorithm 1 The Trimmed Mean (TM_δ) estimator

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

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

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

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

Here is an intuitive explanation of the TM_δ estimator. We first aim to find out a centering 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 points within the set X. Then, the algorithm computes an average of data points from the tightest neighborhood; indeed, our analysis will go through even if the output is X_{i^*}. Notably, if all the points in X were independent and belonged to a certain cluster, say $\{Y_i : i \in T_h^*\}$, 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_h. 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_h. When δ is very small, the estimator is approximately the sample mean, which is unbiased for θ_h.

In the main clustering algorithm, we apply the above estimator on approximations of the true cluster T_h^*. In this approximated cluster, say T_h with a size m, the data points are not necessarily independent, and there are misclustered points. In such scenarios, we will require to contain at least half of the points in $\{Y_i : i \in T_h\}$ to come from $\{Y_i : i \in T_h^*\}$ to make the estimation meaningful. Let us assume $\lceil m(\frac{1}{2} + c) \rceil$ points from T_h belong to $\{Y_i : i \in T_h^*\}$. 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 θ_h, albeit with a bias that depends on G, σ, δ. Notably, with a large enough Δ, the mislabeling errors 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 initial centroid or label estimates 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. (Omran, Engelbrecht and Salman, 2007, Section 3). We will provide a novel robust centroid initialization technique later in Section 4 that will guarantee global optimal mislabeling when combined with the COD_δ algorithm.

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

Input: Data $\{Y_1, \dots, Y_n\}$. Initial centroid estimates $(\widehat{\theta}_1^{(0)}, \dots, \widehat{\theta}_k^{(0)})$ (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 $h \in \{1, 2, \dots, k\}$ **do**
    Labeling step:
    if $s = 1$ and initial label estimates are available **then**
        Compute clusters $T_h^{(s)} = \{i \in \{1, \dots, n\} : \widehat{z}_i^{(0)} = h\}$
    else
        Compute the clusters, with ties broken arbitrarily,

$$T_h^{(s)} = \left\{ i \in \{1, \dots, n\} : \|Y_i - \widehat{\theta}_h^{(s-1)}\| \le \|Y_i - \widehat{\theta}_a^{(s-1)}\|,\ a \in \{1, \dots, k\}, a \neq h \right\},$$

    end if
    Estimation step: Update the new centroid 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: $(\widehat{\theta}_1^{(s)}, \dots, \widehat{\theta}_k^{(s)})$ and $\widehat{z}_i^{(s)} = \operatorname{argmin}_{h \in \{1,\dots,k\}} \|Y_i - \widehat{\theta}_h^{(s)}\|$.

2.3. *Mislabeling guarantees.* To better present our results, we first introduce some notations. For all $h, g \in [k]$, define

$$\begin{aligned} T_h^* &= \{i \in [n] : z_i = h\}, T_h^{(s)} = \left\{ i \in [n] : z_i^{(s-1)} = h \right\} \\ n_h^* &= |T_h^*|, n_h^{(s)} = \left| T_h^{(s)} \right|, n_{hg}^{(s)} = \left| T_h^* \cap T_g^{(s)} \right| \end{aligned} \tag{4}$$

Note that for $s \ge 1$ this implies

$$T_h^{(s)} = \left\{ i \in [n] : \|Y_i - \widehat{\theta}_h^{(s-1)}\| \le \|Y_i - \widehat{\theta}_a^{(s-1)}\|, a \in [k] \right\}. \tag{5}$$

with ties broken arbitrarily. Let us define the minimum fraction of points in the data set that come from a single component as

$$\alpha = \min_{g \in [k]} \frac{n_g^*}{n}. \tag{6}$$

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

$$H_s = \min_{g \in [k]} \left\{ \min \left\{ \frac{n_{gg}^{(s)}}{n_g^*}, \frac{n_{gg}^{(s)}}{n_g^{(s)}} \right\} \right\}. \tag{7}$$

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

$$\Lambda_s = \max_{h \in [k]} \frac{1}{\Delta} \|\widehat{\theta}_h^{(s)} - \theta_h\|. \tag{8}$$

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

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

When SNR is constant, given any data point generated from a two cluster mixture distribution, any algorithm will incorrectly determine the data generating cluster component with a non-vanishing probability. As a consequence, we only study the mislabeling guarantees as a function of the SNR only when the SNR is significantly large. We have the following result.

THEOREM 1. *Suppose that $\gamma \in (\frac{10}{n\alpha}, \frac{1}{2})$. Then there exist two constants $c_{G,\alpha} > 0$ and $C_{G,\alpha,\gamma} > 0$ such that the following hold. If the clustering initialization satisfies*

$$H_0 \geq \frac{1}{2} + \gamma \quad \textit{or} \quad \Lambda_0 \leq \frac{1}{2} - \sqrt{\frac{c_{G,\alpha}}{\mathsf{SNR}}},$$

then whenever $\mathsf{SNR} \geq C_{G,\alpha,\gamma}$ we have that the COD_δ algorithm with $\delta = \frac{1}{2} - \frac{\gamma}{4}$ achieves the expected mislabeling rate

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

REMARK 1. The constants mentioned in Theorem 1 are given by

$$c_{G,\alpha} = G^{-1}\left(\exp\left\{-\frac{c_0}{\alpha}\right\}\right) \quad \text{and} \quad C_{G,\alpha,\gamma} = G^{-1}\left(\exp\left\{-\frac{c_0}{\alpha\gamma}\right\}\right).$$

for an absolute constant c_0. Note that our proof shows that the dependency on γ in the condition $\mathsf{SNR} \geq C_{G,\alpha,\gamma}$ can be removed if the centroid based initialization condition on Λ_0 is satisfied.

2.4. *Clustering in presence of adversarial outliers.* We extend the above results to include adversarial outliers. We study the setup where an adversary, after accessing the original data set, adds n^{out} many new points of its choice. In essence, to retain the above mislabeling guarantee, we need to apply a higher value of the truncation parameter δ. We have the following guarantees.

THEOREM 2. *Suppose that $\gamma \in (\frac{10}{n\alpha}, \frac{1}{2})$ and an adversary, after analyzing the data $Y_1, \ldots, Y_n$ coming from the general mixture model (2), adds $n^{\mathsf{out}} = n\alpha(1-\psi)$ many outliers of its choice for some $\psi \in (0, 1]$. Then there exist two constants $c_{G,\alpha,\psi} > 0$ and $C_{G,\alpha,\gamma,\psi} > 0$ such that the following hold: If*

$$H_0 \geq \frac{1}{2} + \gamma \quad \textit{or} \quad \Lambda_0 \leq \frac{1}{2} - \sqrt{\frac{c_{G,\alpha,\psi}}{\mathsf{SNR}}},$$

then whenever $\mathsf{SNR} \geq C_{G,\alpha,\gamma,\psi}$ we have that the COD_δ algorithm with $\delta = \frac{1}{2} - \frac{1}{4}\min\left\{\gamma, \frac{\psi}{6}\right\}$ achieves the expected mislabeling rate

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

Here $\widehat{z}^{(s)}$ is the vector of estimated labels for the real data.

REMARK 2. The constants mentioned in Theorem 1 are given as

$$c_{G,\alpha,\psi} = G^{-1}\left(\exp\left\{-\frac{c_0}{\alpha\psi}\right\}\right) \quad \text{and} \quad C_{G,\alpha,\gamma} = G^{-1}\left(\exp\left\{-\frac{c_0}{\alpha \cdot \min\{\gamma, \psi\}}\right\}\right).$$

for an absolute constant c_0. Similar to before, the dependency on γ in the condition $\mathsf{SNR} \geq C_{G,\alpha,\gamma,\psi}$ can be removed if the centroid based initialization condition on Λ_0 is satisfied. Since the adversarial data are arbitrary, the result is also applicable to the case that the number of clusters is under determined. 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 are not too large. Then, Theorem 2 gives an accurate bound on the first k clusters.

2.5. *Optimality of mislabeling: lower bound.* In this section we show that it is impossible to achieve a smaller mislabeling error, up to constants depending on k. Notably, we show that even when we have a good centroid initialization, the mislabeling error can be as high as $\frac{1}{8}G(\mathsf{SNR})$ for any fixed k. Suppose that $\theta_1^*, \dots, \theta_k^*$ 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 labels

$$\mathcal{P}_0 = \left\{ z \in [k]^n, \{\theta_i\}_{i=1}^k \in \mathbb{R}^k : \quad \theta_i = \theta_i^*, i \in [k], \quad |\{i \in [n] : z_i = g\}| \geq \frac{n\alpha}{k}, g \in [k] \right\}$$

In addition, we assume that the decay function G satisfies the following 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. *Given any decay function G satisfying (Q), there exists $C_G > 0$ such that*

$$\inf_{\widehat{z}} \sup_{\mathcal{P}_0} \mathbb{E}\left[\ell(\widehat{z}, z)\right] \geq \frac{1 - \alpha - k/n}{4} \cdot G(\mathsf{SNR} + C_G), \quad \Delta \geq \sigma C_G.$$

3. Applications to specific distributions. In this section, we showcase our general results to two specific mixture models with error distributions having Gaussian tails and polynomial tails.

3.1. *Sub-Gaussian mixture model.* In this model, the observed data $Y_1, \dots, Y_n \in \mathbb{R}^d$ are distributed as

$$Y_i = \theta_{z_i} + w_i, \ i = 1, \dots, n, \tag{9}$$

where $\{z_i\}_{i=1}^n \in [k]^n$ denote the underlying unknown component labels 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.,

$$\mathbb{E}\left[e^{\langle a, w_i\rangle}\right] \leq e^{\frac{\sigma^2\|a\|^2}{2}}, \text{ for all } i \in \{1, \dots, n\} \text{ and } a \in \mathbb{R}^d. \tag{10}$$

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 (Hsu, Kakade and Zhang, 2012): given any $t > 0$ and $w \in \mathsf{SubG}(\sigma)$ we have

$$\mathbb{P}\left[\|w\|^2 > \sigma^2 \cdot (d + 2\sqrt{dt} + 2t)\right] \leq e^{-t}. \tag{11}$$

Simplifying the above, we get

$$\mathbb{P}\left[\|w\| > \sigma \cdot x\right] \leq \exp\left(-(\sqrt{x^2 - d/2} - \sqrt{d/2})^2/2\right), \quad x \geq \sqrt{d}. \tag{12}$$

Hence we apply Theorem 1 with

$$\begin{aligned} G(x) &= \exp\left(-(\sqrt{x^2 - d/2} - \sqrt{d/2})^2/2\right), \quad x \geq \sqrt{d}, \\ G^{-1}(y) &= \left(2\log(1/y) + 2\sqrt{d\log(1/y)} + d\right)^{1/2} \leq \sqrt{2\log(1/y)} + \sqrt{d}. \end{aligned} \tag{13}$$

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*

$$H_0 \geq \frac{1}{2} + \gamma \quad \text{or} \quad \Lambda_0 \leq \frac{1}{2} - \frac{c_1(d+1/\alpha)^{1/4}}{(\mathsf{SNR})^{1/2}}.$$

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

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

The implications of the above results 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; see (Lu and Zhou, 2016) for an example which achieves a similar error rate for the iterative procedure of Lloyd's algorithm for k-means. When d is fixed, the initialization conditions stated above are weaker than the conditions required for the Lloyd 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 the Lloyd algorithm requires a stronger initialization condition to guarantee the optimal mislabeling. Notably, this dependency of initialization condition on λ is necessary for the Lloyd 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 dependence 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.

3.2. *Mixture models with moment constraints on the norm.* In this section, we explore the clustering guarantees 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 $\mathbb{P}[\|w\| > x] < \frac{\sigma^p}{x^p}$ 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

$$Y_i = \theta_{z_i} + w_i,\ i = 1, \ldots, n, \quad w_i \in \mathcal{R}_p(\sigma), \quad z_i \in \{1, 2, \ldots, k\}, \quad \theta_h \in \mathbb{R}^d, h \in [k], \tag{14}$$

where $z = \{z_i\}_{i=1}^n \in [k]^n$ denote the underlying labels. The mislabeling proportion for the estimated labels $\widehat{z}$ 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 exists absolute constants $c_1, c_2 > 0$ such that the following hold. If the clustering initialization satisfies*

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

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, our result states that we should be able to produce good clustering guarantees. Note that the second term is usually negligible.

4. Provable initialization methods. In this section, we propose centroid initialization algorithms which guarantee that the conditions on Λ_0 required in Theorem 1 are met with a high probability. We deal with the cases of the two centroids and more than two centroids separately. The case of more than two centroids follows from a recursive structure which calls the two-cluster algorithm at the end.

4.1. *Two centroids.* We first present our initialization algorithm for the two-cluster setup. 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 approximations 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

1: Create distance matrix $D = \left\{D_{ij} : i, j \in [n], D_{i,j} = \|Y_i - Y_j\|_2\right\}$
2: **for** Each $i \in [n]$ **do**
3: Compute the R_i as the $\lceil nq \rceil$-th smallest number in $\left\{D_{ij}, j \in [n]\right\}$
4: **end for**
5: Find $i^* = \mathrm{argmin}_{i \in [n]} R_i$.

Output: Y_i^*

In view of the above, we present the two centroid initialization algorithm 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 $(1-\beta)m_1$-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 $(1-\beta)m_1$-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

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

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

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

THEOREM 6. *Suppose that out of the n many observed data points there are n_i^* many are from cluster $T_i^*, i = 1, 2$ and n^{out} many are adversarial outliers. Also assume that $n_1^* + n_2^* + n^{\mathsf{out}} = n$ and for some constant $\alpha > 0$ the counts satisfy $n_1^*, n_2^* > n\alpha$, $n^{\mathsf{out}} \leq \frac{n\alpha^2}{32}$. 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 $\mathsf{IOD}_{2,m_1,m,\beta}$ algorithm with $m_1 = \left\lceil \frac{n\alpha}{4} \right\rceil, m = \max\{1, \left\lfloor \frac{n\alpha^2}{16} \right\rfloor\}, \beta = \frac{\alpha}{4}$ 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 3. Our main result Theorem 1 states that for a large enough SNR, the COD algorithm obtains the optimal mislabeling for any initial centroid approximates that satisfies $\Lambda_0 \leq \frac{1}{2+c}$ for some constant $c > 0$. In other words, given centroid estimates μ_1^*, μ_2^* of θ_1, θ_2 respectively, it is sufficient to satisfy

$$\max_{i=1,2} \|\theta_{\pi(i)} - \mu_i^*\| = \Delta\Lambda_0 \leq \frac{\Delta}{2+c}, \tag{15}$$

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

We present the idea of the proof below. The details are presented in Section 7.2. For simplicity of notation, given $x \in \mathbb{R}^d$ and $R > 0$ let $\mathcal{B}(x, R)$ denote the Euclidean ball of radius R around the point x.

PROOF SKETCH OF THEOREM 6. We apply Algorithm 4 to the full data set $\{Y_1, \ldots, Y_n\}$. The first centroid is chosen by picking the index $i^* \in \{1, \ldots, n\}$ such that the tightest neighborhood around Y_{i^*} has the smallest radius

$$C_i = \min\{C > 0 : |\{Y_1, \ldots, Y_n\} \cap \mathcal{B}(Y_i, C)| \geq m_1\}, \quad i^* = \underset{i \in \{1,\ldots,n\}}{\operatorname{argmin}} C_i.$$

We show that $Y_{i^*} \in \cup_{i=1,2} \mathcal{B}(\theta_i, \sigma \widetilde{C}_{G,\alpha})$ for some constant $\widetilde{C}_{G,\alpha}$ depending on the decay function G and minimum cluster proportion α. We start by noting the following results about concentration on the number of points that reside in a neighborhood around the true centroids. A more general version (Lemma 25) with proof has been provided in Section 7.1.

LEMMA 7. *Suppose that $n_i^* \geq n\alpha$ for $i \in \{1, 2\}$. There is an event $\widetilde{\mathcal{E}}$ with $\mathbb{P}\left[\widetilde{\mathcal{E}}\right] \geq 1 - 4e^{-\frac{n\alpha}{4}}$ on which the following holds with a constant $C = \sigma C_{G,\alpha}$. For each $i \in \{1, 2\}$, $\mathcal{B}(\theta_i, C)$ contains at least $n_i^*(1 - \frac{\alpha^2}{16})$ data points from $\{Y_i : i \in T_i^*\}$.*

The above result implies that outside the set $\mathcal{B}(\theta_1, C) \cup \mathcal{B}(\theta_2, C)$ there can be at most $\frac{n\alpha^2}{16}$ points from the dataset. As the minimum distance between the sets $\mathcal{B}(\theta_1, C), \mathcal{B}(\theta_2, C)$ is $\Delta - 2C$, and each set $\mathcal{B}(\theta_1, C)$ contains at least $\frac{n\alpha}{2}$ points, we get that the point Y_{i^*}, having the tightest neighborhood with at least $m_1 = \frac{n\alpha}{4}$ points, will be inside either $\mathcal{B}(\theta_1, 2C)$ or $\mathcal{B}(\theta_2, 2C)$. Without a loss of generality, suppose that $\mathcal{B}(\theta_1, 2C)$ is the corresponding set, and we have found an approximation $\widehat{\theta}_1$ of θ_1 via Y_{i^*}. Denote the first m_1 points in the data set closest to $\widehat{\theta}_1 = Y_{i^*}$ as $\mathcal{P}_1$ and denote the complement set 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 approximation 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 approximate θ_2. Unfortunately, a direct application of the HDP on the set $\overline{\mathcal{P}_1}$ can not guarantee such a good approximation, as there are at least $n_1^* - \frac{n\alpha}{4} \geq \frac{3n\alpha}{4}$ points in $\overline{\mathcal{P}_1} \cap \{Y_i : i \in T_1^*\}$ and the tightest neighborhood found during the application of HDP can indeed belong to $\{Y_i : i \in T_1^*\}$, hence the corresponding output centroid can to be closer to θ_1 than θ_2. 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 a quantile of distances from the centroid estimates $\widehat{\theta}_1, \widehat{\theta}_2$ within the sets $\mathcal{P}_1$ and $\overline{\mathcal{P}_1}$ respectively. The reason behind using the quantiles of the intra-cluster distances rather than their sum, which is often used in the k-means type procedures, being 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 centroid estimate $\widehat{\theta}_2$. Hence, we end up with reasonable approximates of θ_1, θ_2 □

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

THEOREM 8. *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. Computing the tightest neighborhood of m_1 involves 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 which has a time complexity $O(n^2 \log n)$, and finally computing the minimum which takes $O(\log n)$ at most. 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, which takes another n unit time. Next 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. Combining the above, we get that the total runtime is $\frac{O(1)}{\beta^2}\left(n^2 d + n^2 \log n\right)$. □

A similar idea of approximate cluster centroid initialization has been proposed in (Kumar, Sabharwal and Sen, 2004). However, the method presented therein does not guarantee a low mislabeling error. More specifically, given a set of data points $\{Y_1, \ldots, Y_n\}$, the above paper proposes to produce $\widehat{\theta}_1, \widehat{\theta}_2$ that guarantees an approximate minimization of the k-means objective (i.e., partition the data into k sets and then minimize the total within-cluster variances) up to slack of $\epsilon > 0$

$$\sum_{i\in[n]} \left\{ d(Y_i, \{\widehat{\theta}_1, \widehat{\theta}_2\}) \right\}^2 \leq (1+\epsilon) \sum_{i\in[n]} \min_{\{\theta_1,\theta_2\}} \{d(Y_i, \{\theta_1, \theta_2\})\}^2, \quad d(x, S) = \min_{y\in S}\{\|x-y\|_2\}.$$

Such a minimization procedure can produce a reasonable centroid estimate when the data is highly concentrated around the centroids, for example, for sub-Gaussian errors. However, it is unknown whether such a method will work for heavy tail models as in those cases with a high probability, many observations might behave like outliers. In addition, the aforementioned paper uses an argument involving sampling a constant number of points from the data set and then using the sample to represent the entire data to perform the clustering. Consequently, the final theoretical guarantees the paper obtains hold with a probability γ^k for a constant γ much smaller than one. In contrast, our centroid estimation guarantees hold with a probability that approaches one as the size of the data set increases.

4.2. *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 initialize with computing a high-density point from the data set that has the tightest neighborhood (denote it by $\mathcal{P}$) of size m_1. This will serve as the first centroid estimate. Then for the remaining point set (call it $\overline{\mathcal{P}}$) we recursively apply the clustering 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 optimal clustering at that step. Finally, the centroids generated in the clustering step with the lowest distance measure is chosen to be the final output. Whenever we are left with the task of finding the best two centroid cluster estimates from $\overline{\mathcal{P}}$, we resort to $\mathsf{IOD}_{2,m_1,m,\beta}$. The details are provided in Algorithm 5.

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 $\overline{\mathcal{P}_k^{(1)}}$.
  for $\ell_k = 1,\dots,\left\lfloor \frac{n-m_1}{m} \right\rfloor$ **do**
    Set $\mu_k^{(k,\ell_k)} = \mu_k^{(k,1)}$ and compute

$$\mathsf{dist}_k^{(\ell_k)} = \text{the distance to the } (1-\beta)|\mathcal{P}_k^{(\ell_k)}|\text{-th closest point from } \mu_k^{(k,\ell_k)} \text{ in } \mathcal{P}_k^{(\ell_k)}.$$

    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 steps $\mathsf{IOD}_{2,m_1,m,\beta}$ algorithm and note the output as $\{\mu_1^*,\mu_2^*\}$ and $\mathsf{totdist}_k^*$.
end if

The following result describes a choice of the parameters m_1, m, β that guarantees a good initialization, sufficient to meet the requirements on Λ_0 in Theorem 1. Hence, our initialization algorithm, paired with the clustering technique COD, produces the optimal mislabeling starting from scratch.

THEOREM 9. *Suppose that out of the n many observed data points there are n_i^* many are from cluster $T_i^*, i = 1,\dots,k$ and n^{out} many are adversarial outliers. Also assume that $\sum_{i=1}^k n_i^* + n^{\mathsf{out}} = n$ and for some constant $\alpha > 0$ the counts satisfy $n_i^* > \frac{n\alpha}{k}, i = 1,\dots,k$, $n^{\mathsf{out}} \leq \frac{n\alpha^3}{64k^3}$. 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 approximations generated via the $\mathsf{IOD}_{k,m_1,m,\beta}$ algorithm with*

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

In view of Theorem 9, our initialization paired with the COD algorithm leads to an optimal mislabeling. Notably, the Lloyd algorithm (Lu and Zhou, 2016) and the hybrid k-median algorithm in (Jana, Kulkarni and Yang, 2023) also required the initialization condition $\Lambda_0 < 1/(2+c)$, for any constant $c > 0$, to produce the optimal mislabeling rate in the sub-Gaussian clustering problem. In view of Section 3, and the proof of Theorem 9 in Appendix E.2, we note that the constant $c_{k,\alpha,G}$ mentioned above in Theorem 9 is given by $\sigma c_\alpha \sqrt{d}$ for some constant c_α depending on α. This implies the following.

COROLLARY 10. *There is a constant c_α depending on α such that the following holds true. The Lloyd algorithm (Lu and Zhou, 2016) and the hybrid k-median algorithm in*

(Jana, Kulkarni and Yang, 2023) produce the optimal mislabeling rate in the sub-Gaussian error setup, provided $\Delta > \sigma c_\alpha \sqrt{d}$.

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

THEOREM 11. *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 computing 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 $1 \le \ell_k \le 2/\beta^2$ 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 in view of the above reasoning we have

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

Solving the above recursion we get

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

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

5. Suboptimality of the Lloyd algorithm. In this section, we establish that the Lloyd algorithm might produce a suboptimal mislabeling even when the initial labels are reasonably good, due to non-robust estimate of centroids. 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 poorly estimated centroids. In the two centroid setup, the suboptimality occurs when error distributions exhibits heavy tails.

5.1. *The case of at least three centroids.* For this section, we assume that whenever the Lloyd 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 12. *Given any* $\beta \in (0,1)$*, there exists a system of three centroids and an initialization with mislabeling proportion* β *such that the Lloyd 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

$$\begin{aligned} &Y_i = \theta_{z_i} + w_i,\ i = 1, \dots, n, \\ &w_i \overset{\text{iid}}{\sim} \text{Uniform}(-1,1),\ z_i \in \{1,2,3\},\ \theta_1 = -\frac{\Delta}{2}, \theta_2 = \frac{\Delta}{2}, \theta_3 = \frac{c\Delta}{2\beta}. \end{aligned} \tag{16}$$

Let $T_h^* = \{i \in [n] : z_i = h\}, h \in \{1,2,3\}$ 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 take the initialization

$$\widehat{z}_i^{(0)} = \begin{cases} 2 & \text{if } i \in S, \\ z_i & \text{otherwise.} \end{cases} \tag{17}$$

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

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

$$\widehat{\theta}_1^{(0)} \le -\frac{\Delta}{2} + 1, \quad \widehat{\theta}_2^{(0)} \in \left(\frac{(c+1)\Delta}{2(1+\beta)} - 1, \frac{(c+1)\Delta}{2(1+\beta)} + 1 \right), \quad \widehat{\theta}_3^{(0)} \ge \frac{c\Delta}{2\beta} - 1. \tag{18}$$

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 $\theta_2^{(s)}, \theta_3^{(s)}$ 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. □

5.2. *The case of two centroids.* We produce a counter example where k-means 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

$$G(x) = \frac{1}{1 + x^{1-\epsilon}}, \quad x > 0. \tag{19}$$

The model we use is

$$\begin{gathered} Y_i = \theta_{z_i} + w_i, \ i = 1, \dots, n, \\ w_i \overset{\text{iid}}{\sim} W, \quad \mathbb{P}[W > x] = G(x), x > 0, \quad z_i \in \{1,2\}, \quad \theta_1 = 0, \theta_2 = \Delta, \end{gathered} \tag{20}$$

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

$$\begin{aligned} &\mathbb{P}\left[\cup_{i=1}^n \left\{w_i > n^{1+\epsilon}\right\}\right] = 1 - \mathbb{P}\left[\cap_{i=1}^n \left\{w_i \le n^{1+\epsilon}\right\}\right] \\ &\qquad = 1 - \Pi_{i=1}^n \mathbb{P}\left[w_i \le n^{1+\epsilon}\right] = 1 - \left(1 - \frac{1}{1 + n^{1-\epsilon^2}}\right)^n \ge 1 - e^{-n^{\epsilon^2}}. \end{aligned} \tag{21}$$

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

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

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 result about concentration of counts sample quantiles.

LEMMA 13. *Fix $\epsilon_0 > 0$. Then there is an event $\mathcal{E}_{\epsilon_0}^{\mathsf{con}}$ with probability at least $1 - k \cdot e^{-\frac{\min_{g \in [k]} n_g^*}{4}}$ on which*

$$\sum_{i \in T_g^*} \mathbf{1}_{\{\epsilon^2 \Delta \le \|w_i\|\}} \le \frac{5 n_g^*}{4 \log(1/G(\epsilon_0^2 \Delta/\sigma))}, \quad \epsilon \ge \epsilon_0, \forall g \in [k].$$

A proof of the above result is presented at the end of this section. Note that in view of Lemma 13, for all large enough Δ and n we have

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

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

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

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 13. We define $B_i = \mathbf{1}_{\{\epsilon_0^2\Delta\leq\|w_i\|\}}$. As $\epsilon \geq \epsilon_0$, it is enough to find an event $\mathcal{E}_1$ with the said probability on which

$$\mathbb{P}\left[\sum_{i\in T_g^*} B_i \geq \frac{5n_g^*}{4\log(1/G(\epsilon_0^2\Delta/\sigma))}\right] \leq e^{-\frac{n_g^*}{4}} \text{ for each } g\in[k]. \tag{22}$$

Note that

$$\mathbb{P}\left[\epsilon_0^2\Delta \leq \|w_i\|\right] \leq G(\epsilon_0^2\Delta/\sigma). \tag{23}$$

This implies $\sum_{i\in T_g^*} B_i$ is stochastically smaller than Binom$(n_g^*, G(\epsilon_0^2\Delta/\sigma))$. We continue to analyze (22) via the Chernoff's inequality in Lemma 24 for the Binomial random variable Binom$(n_g^*, G(\epsilon_0^2\Delta/\sigma))$. Let

$$a = \frac{5}{4\log(1/G(\epsilon_0^2\Delta/\sigma))}, \quad m = n_g^*, \quad q = e^{-5/(4a)} = G(\epsilon_0^2\Delta/\sigma).$$

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

$$\begin{aligned}
\mathbb{P}\left[\sum_{i\in T_g^*} B_i \geq \frac{5n_g^*}{4\log(1/G(\epsilon_0^2\Delta/\sigma))}\right] &\leq \exp\left(-mh_q(a)\right) \\
&\leq \exp\left(-m\left(a\log\frac{a}{q} + (1-a)\log\frac{1-a}{1-q}\right)\right) \\
&\leq \exp\left(-m\left\{a\log\frac{a}{e^{-5/(4a)}} + (1-a)\log(1-a)\right\}\right) \\
&= \exp\left(-m\left\{a\log a + (1-a)\log(1-a) + \frac{5}{4}\right\}\right) \leq e^{-n_g^*/4}. \qquad (24)
\end{aligned}$$

□

6. Experiments.

6.1. *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 algorithm (e.g., the Lloyd–Forgy algorithm (Lloyd, 1982)). For initializing the Lloyd algorithm, we considers three methods:

- the proposed IOD algorithm
- the k-means++ algorithm (Vassilvitskii and Arthur, 2006)
- 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

$$f(x) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)\nu^{d/2}\pi^{d/2}\sigma}\left[1+\frac{\|x\|^2}{\sigma\nu}\right]^{-(\nu+d)/2}. \tag{25}$$

We study the effect of different dimension 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 ratio 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 95% confidence interval width. The average mislabeling errors are presented in Table 1, Table 2, Table 3 (along with the confidence interval widths within the parenthesis).

Results. We first present the numerical study describing the effect of ν Table 1. For the large values of $\nu = 10$ the data are supposed to be highly concentrated around the centroids, which should guarantee a low mislabeling error. In such a light tail setup, the Lloyd algorithm should work well, 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, the Lloyd 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 the Lloyd type methods perform equally bad, while our algorithm produces significantly lower mislabeling errors.

TABLE 1
Effect of degrees of freedom: $n = 200k, \sigma = 5, d = 5, \Delta = 25$

k	ν	COD + IOD	Lloyd + IOD	Lloyd + k-means++	Lloyd + random init
2	1	0.322 (0.011)	0.495 (0.002)	0.498 (0.000)	0.497 (0.000)
	1.5	0.128 (0.001)	0.322 (0.014)	0.48 (0.006)	0.366 (0.014)
	10	0.014 (0.000)	0.013 (0.000)	0.014 (0.000)	0.014 (0.000)
3	1	0.422 (0.005)	0.652 (0.004)	0.664 (0.000)	0.65 (0.005)
	1.5	0.364 (0.007)	0.411 (0.009)	0.576 (0.011)	0.403 (0.013)
	10	0.043 (0.008)	0.034 (0.007)	0.014 (0.000)	0.081 (0.013)

Next, we demonstrate the effect of the scale parameter σ in Table 2. For 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 values, and this is supported by our demonstrations. Additionally, in all the setups, our algorithm performs significantly better than its competitors.

TABLE 2
Effect of scale: $n = 200k, \nu = 1.5, d = 10, \Delta = 25$

k	σ	COD + IOD	Lloyd + IOD	Lloyd + k-means++	Lloyd + random init
2	1	0.014 (0.000)	0.029 (0.006)	0.274 (0.018)	0.1 (0.014)
	5	0.173 (0.003)	0.424 (0.006)	0.496 (0.001)	0.451 (0.005)
	10	0.352 (0.005)	0.492 (0.001)	0.497 (0.000)	0.495 (0.001)
3	1	0.161 (0.012)	0.169 (0.012)	0.27 (0.017)	0.169 (0.013)
	5	0.412 (0.001)	0.485 (0.006)	0.654 (0.003)	0.53 (0.008)
	10	0.509 (0.003)	0.628 (0.005)	0.664 (0.000)	0.647 (0.004)

In Table 3 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 increase proportionally to the square root of the dimension, multiplied with variability in each coordinate. Nonetheless, we see that our proposed clustering algorithm perform more robustly compared to the other methods in the simulation studies. It might be possible to improve all the clustering techniques by applying some sort of dimension reduction, for example, feature screening approaches (Fan and Fan, 2008) and the spectral methods in (Löffler, Zhang and Zhou, 2021), to the data set prior to applying the clustering methods. However, such analysis is beyond the scope of the current work.

TABLE 3
Effect of dimension: $n = 200k, \nu = 1.5, \sigma = 5, \Delta = 25$

k	d	COD + IOD	Lloyd + IOD	Lloyd + k-means++	k-means + random init
2	2	0.099 (0.001)	0.154 (0.007)	0.398 (0.01)	0.231 (0.01)
	10	0.174 (0.004)	0.414 (0.008)	0.495 (0.001)	0.445 (0.006)
	30	0.309 (0.01)	0.492 (0.002)	0.497 (0.000)	0.494 (0.002)
3	2	0.156 (0.008)	0.2 (0.009)	0.38 (0.009)	0.236 (0.009)
	10	0.41 (0.002)	0.479 (0.009)	0.655 (0.004)	0.528 (0.011)
	30	0.467 (0.002)	0.64 (0.002)	0.664 (0.000)	0.653 (0.004)

6.2. *Real data experiments.* Furthermore, we evaluated our proposed algorithm on the publicly available Letter Recognition dataset (Slate, 1991). 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 on this data with the aim of clustering data points corresponding to the same letters together. Additionally, we explore the robustness 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". In both the setups, we randomly sample 100 points from each cluster 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 (Jana, Kulkarni and Yang, 2023) for clustering in presence of adversarial outliers, initialized with the IOD algorithm
- the Lloyd algorithm initialized with the IOD algorithm
- the Lloyd algorithm initialized with the k-means++ algorithm
- the Lloyd algorithm initializations from the dataset.

The relevant parameters for the clustering are the same as those in the simulation studies section, with only modification being for the value of δ. This is in accordance with Theorem 2, 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, are independently repeated 150 times to measure the average mislabeling proportion and the corresponding 95% confidence bandwidths. The results are presented in Table 4 (the confidence bandwidths 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 yields better mislabeling rate even in absence of outliers. This probably 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, and we rather study the effect of a certain outlier class. This possibly indicates a similarity of the data points from the outlier class with one of the clusters, resulting in observing more points in the neighborhood of the corresponding cluster. When we observe more points in the clusters, the task of separating the clusters becomes much easier, resulting in a lower mislabeling.

TABLE 4
Results for clustering letters: $n = 100k, \delta = 0.48$, outlier proportion = 20%, outlier class = R

Classes	Outliers	COD + IOD	k-median + IOD	Lloyd + IOD	Lloyd + k-means++	Lloyd + random
W, V	without	0.276 (0.008)	0.32 (0.005)	0.391 (0.005)	0.355 (0.004)	0.402 (0.004)
	with	0.269 (0.008)	0.317 (0.004)	0.381 (0.004)	0.352 (0.004)	0.398 (0.004)
X, M, A	without	0.194 (0.010)	0.245 (0.007)	0.374 (0.004)	0.342 (0.006)	0.357 (0.007)
	with	0.264 (0.009)	0.275 (0.006)	0.388 (0.003)	0.354 (0.004)	0.379 (0.005)

7. Proof of the two cluster initialization result (Theorem 6).

7.1. *Preparation.* Our proofs rely on the following high probability guarantees.

LEMMA 14. *The following statements hold for the β in Theorem 6. There is an event $\widetilde{\mathcal{E}}$ with $\mathbb{P}\left[\widetilde{\mathcal{E}}\right] \geq 1 - 4e^{-\frac{\min_{g=1,2} n_g^*}{4}}$ on which the following holds for the 2-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* $i = 1, 2$,

(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 from the proof of Lemma 13 by choosing $\epsilon_0^2 = \frac{\sigma}{\Delta} G^{-1}(e^{-\frac{5}{4\beta^2}})$ in the lemma. The proof of part (ii) follows from the proof of Lemma 13 by choosing $\epsilon_0^2 = \frac{1}{16k}$ in the lemma. □

7.2. *Proof of Theorem 6.* In the proof below, we assume that all the mentioned constants depend on G, α, σ, unless otherwise specified. Let n^{out} be the total number of outliers, i.e., $n_1^* + n_2^* + n^{\mathsf{out}} = n$ and $n^{\mathsf{out}} \leq \frac{n\alpha^2}{32}$. In addition, for our entire analysis we will assume that the event $\widetilde{\mathcal{E}}$ 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. The proof is a combination of the following results.

LEMMA 15. *There is one θ_i, such that $\|\theta_i - \mu_1^{(1)}\| \leq 3\sigma G^{-1}(e^{-\frac{5}{4\beta^2}})$.*

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

LEMMA 17. *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 18. *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 15, Lemma 16 and Lemma 17 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 18 this implies $\max_{i=1,2} \|\theta_{\pi(i)} - \mu_i^*\| \leq \Delta/3$ as required. Below we prove Lemma 15, Lemma 16 and the rest of the proofs are diverted to Appendix E.1.

PROOF OF LEMMA 15. In view of Lemma 14, there is a constant $c_1 = \sigma G^{-1}(e^{-\frac{5}{4\beta^2}})$ such that

$$|\{j \in [n] : Y_j \in \mathcal{B}(\theta_i, c_1)\}| \geq n_i^*\left(1 - \beta^2\right), \quad i \in \{1, 2\}. \tag{26}$$

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

$$D(x, S) = \{\|x - Y_i\| : i \in S\}, \quad x \in \mathbb{R}^d, S \subseteq [n]. \tag{27}$$

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

$$\{D(Y_{i^*},[n])\}^{\{1-\frac{m_1}{n}\}} = \min_{j\in[n]} \{D(Y_j,[n])\}^{\{1-\frac{m_1}{n}\}}. \tag{28}$$

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 (26) the disjointedness will imply that their union will contain more than n points from $Y_1,\ldots,Y_n$

$$|\{i\in[n]: Y_i\in\mathcal{B}(Y_{i^*},2c_1)\}\cup[\cup_{j=1,2}\{i\in[n]: Y_i\in\mathcal{B}(\theta_j,c_1)\}]|$$

$$\geq m_1+\sum_{i=1,2} n_i^*\left(1-\beta^2\right) = \frac{n\alpha}{4}+(n-n^{\mathsf{out}})(1-\beta^2) \geq n+\frac{n\alpha}{4}-n\beta^2-n^{\mathsf{out}} \geq n+\frac{n\alpha^2}{8},$$

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 $\|\theta_1-\theta_2\|\geq\Delta$, $n_1^*+n_2^*+n^{\mathsf{out}}=n$ 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 θ_1 is the closer to $\mu_1^{(1)}$ than θ_2. □

PROOF OF LEMMA 16. In view of Lemma 15, 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 in the proof of Lemma 15 to have

$$\mu_1^{(1)}\in\mathcal{B}\left(\theta_1,3c_1\right), \quad |\{Y_j: j\in T_i^*\}\cap\mathcal{B}\left(\theta_i,c_1\right)|\geq n_i^*\left(1-\beta^2\right), i\in\{1,2\}. \tag{29}$$

We observe the following:

- In view of $\mathcal{B}(\mu_1^{(1)},4c_1)\supset\mathcal{B}(\theta_1,c_1)$ we get
$$\left|\{Y_i: i\in[n]\}\cap\mathcal{B}(\mu_1^{(1)},4c_1)\right| \geq |\{Y_i: i\in[n]\}\cap\mathcal{B}(\theta_1,c_1)| \geq n\alpha(1-\beta^2)\geq\frac{n\alpha}{2}.$$
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 (29) we also have
$$|\{Y_j: j\in[n], Y_j\in\mathcal{B}(\theta_2,c_1)\}\cap\mathcal{P}_1^{(\widetilde{\ell})}| \geq |\{Y_j: j\in[n], Y_j\in\mathcal{B}(\theta_2,c_1)\}| - |\overline{\mathcal{P}_1^{(\widetilde{\ell})}}|$$
$$\geq n_2^*\left(1-\beta^2\right)-\frac{n\alpha^2}{16}\geq n\alpha-\frac{n\alpha^2}{16}-n\alpha\beta^2.$$

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
- (29) 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

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

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 entry 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. □

REFERENCES

ABBASI, A. A. and YOUNIS, M. (2007). A survey on clustering algorithms for wireless sensor networks. *Computer Communications* **30** 2826-2841. Network Coverage and Routing Schemes for Wireless Sensor Networks. https://doi.org/10.1016/j.comcom.2007.05.024

ABBE, E., FAN, J. and WANG, K. (2022). An ℓ_p theory of PCA and spectral clustering. *The Annals of Statistics* **50** 2359–2385.

AJALA FUNMILOLA, A., OKE, O., ADEDEJI, T., ALADE, O. and ADEWUSI, E. (2012). Fuzzy kc-means clustering algorithm for medical image segmentation. *Journal of information Engineering and Applications, ISSN* **22245782** 2225–0506.

ARTHUR, D. and VASSILVITSKII, S. (2007). K-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* 1027–1035.

BAKSHI, A. and KOTHARI, P. (2020). Outlier-robust clustering of non-spherical mixtures. *arXiv preprint arXiv:2005.02970*.

BAKSHI, A., DIAKONIKOLAS, I., JIA, H., KANE, D. M., KOTHARI, P. K. and VEMPALA, S. S. (2022). Robustly learning mixtures of k arbitrary gaussians. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing* 1234–1247.

BEATTY, A., LIAO, S. and YU, J. J. (2013). The spillover effect of fraudulent financial reporting on peer firms' investments. *Journal of Accounting and Economics* **55** 183–205.

BOJCHEVSKI, A., MATKOVIC, Y. and GÜNNEMANN, S. (2017). 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* 737–746.

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

CAI, F., LE-KHAC, N.-A. and KECHADI, T. (2016). Clustering approaches for financial data analysis: a survey. *arXiv preprint arXiv:1609.08520*.

CHARIKAR, M., KHULLER, S., MOUNT, D. M. and NARASIMHAN, G. (2001). Algorithms for facility location problems with outliers. In *SODA* **1** 642–651. Citeseer.

CHEN, M., GAO, C. and REN, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics* **46** 1932–1960.

CHEN, X. and ZHANG, A. Y. (2021). Optimal clustering in anisotropic gaussian mixture models. *arXiv preprint arXiv:2101.05402*.

DE MIRANDA CARDOSO, J. V., YING, J. and PALOMAR, D. (2021). Graphical models in heavy-tailed markets. *Advances in Neural Information Processing Systems* **34** 19989–20001.

DIAKONIKOLAS, I., KAMATH, G., KANE, D., LI, J., MOITRA, A. and STEWART, A. (2019). Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing* **48** 742–864.

DIAKONIKOLAS, I., HOPKINS, S. B., KANE, D. and KARMALKAR, S. (2020). Robustly learning any clusterable mixture of gaussians. *arXiv preprint arXiv:2005.06417*.

DIAKONIKOLAS, I., KANE, D. M., KONGSGAARD, D., LI, J. and TIAN, K. (2022). 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* 1262–1275.

FAN, J. and FAN, Y. (2008). High dimensional classification using features annealed independence rules. *Annals of statistics* **36** 2605.

FAN, J., LI, Q. and WANG, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **79** 247–265.

FAN, J., LIU, Q., WANG, B. and ZHENG, K. (2023). Unearthing Financial Statement Fraud: Insights from News Coverage Analysis. *Available at SSRN 4338277*.

GUPTA, S., KUMAR, R., LU, K., MOSELEY, B. and VASSILVITSKII, S. (2017). Local search methods for k-means with outliers. *Proceedings of the VLDB Endowment* **10** 757–768.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. H. and FRIEDMAN, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* **2**. Springer.

HSU, D., KAKADE, S. and ZHANG, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic communications in Probability* **17** 1–6. https://doi.org/10.21214/ECP.v7-2079

HUBER, P. J. (1965). A robust version of the probability ratio test. *The Annals of Mathematical Statistics* 1753–1758.

HUBER, P. J. (1992). Robust estimation of a location parameter. *Breakthroughs in statistics: Methodology and distribution* 492–518.

JANA, S., KULKARNI, S. and YANG, K. (2023). Optimality and adversarial robustness of the Lloyd-median algorithm. *Preprint*.

KANNAN, R., VEMPALA, S. et al. (2009). Spectral algorithms. *Foundations and Trends® in Theoretical Computer Science* **4** 157–288.

KAUFMAN, L. and ROUSSEEUW, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

KUMAR, A., SABHARWAL, Y. and SEN, S. (2004). 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* 454–462. IEEE.

LIU, A. and MOITRA, A. (2023). Robustly Learning General Mixtures of Gaussians. *Journal of the ACM*.

LLOYD, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory* **28** 129–137.

LÖFFLER, M., ZHANG, A. Y. and ZHOU, H. H. (2021). Optimality of spectral clustering in the Gaussian mixture model. *The Annals of Statistics* **49** 2506–2530.

LU, Y. and ZHOU, H. H. (2016). Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*.

LUGOSI, G. and MENDELSON, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics* **19** 1145–1190.

LUGOSI, G. and MENDELSON, S. (2021). Robust multivariate mean estimation: the optimality of trimmed mean.

MARAVELIAS, C. D. (1999). Habitat selection and clustering of a pelagic fish: effects of topography and bathymetry on species dynamics. *Canadian Journal of Fisheries and Aquatic Sciences* **56** 437–450.

NG, H., ONG, S., FOONG, K., GOH, P.-S. and NOWINSKI, W. (2006). Medical image segmentation using k-means clustering and improved watershed algorithm. In *2006 IEEE southwest symposium on image analysis and interpretation* 61–65. IEEE.

OLUKANMI, P. O. and TWALA, B. (2017). K-means-sharp: modified centroid update for outlier-robust k-means clustering. In *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)* 14–19. IEEE.

OMRAN, M. G., ENGELBRECHT, A. P. and SALMAN, A. (2007). An overview of clustering methods. *Intelligent Data Analysis* **11** 583–605.

PATEL, D., SHEN, H., BHAMIDI, S., LIU, Y. and PIPIRAS, V. (2023). Consistency of Lloyd's Algorithm Under Perturbations. *arXiv preprint arXiv:2309.00578*.

PIGOLOTTI, S., LÓPEZ, C. and HERNÁNDEZ-GARCÍA, E. (2007). Species clustering in competitive Lotka-Volterra models. *Physical review letters* **98** 258101.

RONAN, T., QI, Z. and NAEGLE, K. M. (2016). Avoiding common pitfalls when clustering biological data. *Science signaling* **9** re6–re6.

ROUSSEEUW, P. and KAUFMAN, P. (1987). Clustering by means of medoids. In *Proceedings of the statistical data analysis based on the L1 norm conference, neuchatel, switzerland* **31**.

SASIKUMAR, P. and KHARA, S. (2012). K-means clustering in wireless sensor networks. In *2012 Fourth international conference on computational intelligence and communication networks* 140–144. IEEE.

SFIKAS, G., NIKOU, C. and GALATSANOS, N. (2007). Robust image segmentation with mixtures of Student's t-distributions. In *2007 IEEE International Conference on Image Processing* **1** I–273. IEEE.

SLATE, D. (1991). Letter Recognition. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5ZP40.

SONG, S., SJÖSTRÖM, P. J., REIGL, M., NELSON, S. and CHKLOVSKII, D. B. (2005). Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS biology* **3** e68.

SRIVASTAVA, P. R., SARKAR, P. and HANASUSANTO, G. A. (2023). A robust spectral clustering algorithm for sub-Gaussian mixture models with outliers. *Operations Research* **71** 224–244.

SUN, Q., ZHOU, W.-X. and FAN, J. (2020). Adaptive huber regression. *Journal of the American Statistical Association* **115** 254–265.

VASSILVITSKII, S. and ARTHUR, D. (2006). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* 1027–1035.

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

WANG, B. and FAN, J. (2022). Robust matrix completion with heavy-tailed noise. *arXiv preprint arXiv:2206.04276*.

XU, R. and WUNSCH, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks* **16** 645–678.

YU, B. (1997). Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics* 423–435. Springer.

ZHANG, Y. and ROHE, K. (2018). Understanding regularized spectral clustering via graph conductance. *Advances in Neural Information Processing Systems* **31**.

APPENDIX A: MISLABELING UPPER BOUND IN THEOREM 1

A.1. Preparation. The proof of the performance of the COD_δ algorithm is primarily based in the following two stages:

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

We obtain results on these steps separately and then combine them to prove Theorem 1. Our analysis depends on high-probability events $\mathcal{E}_\tau^{\mathsf{norm}}, \mathcal{E}_{\gamma_0,\epsilon_0}$ given in the following lemmas.

LEMMA 19. *Suppose that w_i-s are independent random variables satisfying the G_σ-decay condition 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*

$$\left\{ i \in S : \|w_i\|_2 \leq \sigma G^{-1}\left(\exp\left\{-\frac{1+1/\beta}{\tau}\right\}\right)\right\}$$

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

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

LEMMA 20. *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 $\frac{5}{2\alpha\log(1/G(\epsilon_0^2\Delta/\sigma))} < \frac{1}{2}$ 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 $H_{s+1} \geq 1 - \frac{5}{2\alpha\log(1/G(\epsilon_0^2\Delta/\sigma))}$,*
(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 20. We first prove part (i). For any $g \neq h \in [k] \times [k]$,

$$\begin{aligned}
\mathbf{1}_{\left\{z_i=g, \widehat{z}_i^{(s+1)}=h\right\}} &\leq \mathbf{1}_{\left\{\|Y_i - \widehat{\theta}_h^{(s)}\|^2 \leq \|Y_i - \widehat{\theta}_g^{(s)}\|^2,\ i \in T_g^*\right\}} \\
&= \mathbf{1}_{\left\{\|\theta_g + w_i - \widehat{\theta}_h^{(s)}\|^2 \leq \|\theta_g + w_i - \widehat{\theta}_g^{(s)}\|^2\right\}} = \mathbf{1}_{\left\{\|\theta_g - \widehat{\theta}_h^{(s)}\|^2 - \|\theta_g - \widehat{\theta}_g^{(s)}\|^2 \leq 2\langle w_i, \widehat{\theta}_h^{(s)} - \widehat{\theta}_g^{(s)}\rangle\right\}}. \qquad (30)
\end{aligned}$$

The triangle inequality and the fact $\|\theta_h - \widehat{\theta}_h^{(s)}\| \leq \Lambda_s\Delta \leq \Lambda_s\|\theta_g - \theta_h\|$ implies

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

This implies, using the fact that $(1-x)^2 - y^2 \geq (1-x-y)^2$ when $y(1-x-y) \geq 0$, that

$$\begin{aligned}
&\|\theta_g - \widehat{\theta}_h^{(s)}\|^2 - \|\theta_g - \widehat{\theta}_g^{(s)}\|^2 \\
&\geq (1-\Lambda_s)^2\|\theta_g - \theta_h\|^2 - \Lambda_s^2\|\theta_g - \theta_h\|^2 \geq (1-2\Lambda_s)^2\|\theta_g - \theta_h\|^2. \qquad (31)
\end{aligned}$$

In view of the last inequality, using the fact

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

we continue (30) to get

$$\mathbf{1}_{\left\{z_i=g,\widehat{z}_i^{(s+1)}=h\right\}} \leq \mathbf{1}_{\left\{\frac{(1-2\Lambda_s)^2}{2(1+2\Lambda_s)}\|\theta_g-\theta_h\|\leq\|w_i\|\right\}}. \tag{32}$$

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

$$\mathbf{1}_{\left\{z_i=g,\widehat{z}_i^{(s+1)}=h\right\}} \leq \mathbf{1}_{\{\epsilon_0^2\|\theta_g-\theta_h\|\leq\|w_i\|\}} \tag{33}$$

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 13, we get on the set $\mathcal{E}_{\epsilon_0}^{\mathsf{con}}$

$$n_{gh}^{(s+1)} \leq \sum_{i\in T_g^*}\mathbf{1}_{\{\epsilon_0^2\Delta\leq\|w_i\|\}} \leq \frac{5n_g^*}{4\log(1/G(\epsilon_0^2\Delta/\sigma))},\ \forall h\in[k], h\neq g. \tag{34}$$

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

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

Next we switch g,h in (34) and sum over $h\in[k], h\neq g$. We get on the event $\mathcal{E}_{\epsilon_0}^{\mathsf{con}}$

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

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

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

Using the above with (35) we get with probability $1-2kn^{-c/4}$

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

Next we present below the proof of Lemma 20(ii). We use Proposition 21 provided below.

PROPOSITION 21. *Suppose given any $\tau\in(0,1)$, there is a 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 $\{i\in S:\|w_i\|_2\leq D_\tau\}$ 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 pick

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

In view of Lemma 19 note that the event $\mathcal{F}_\tau$ has probability at least $1-e^{-0.3n}$ and satisfies the requirement in Proposition 21. 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 21. We prove this 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 $\left|\left\{j \in S : \|w_j\|_2 \leq D_\tau\right\}\right| \geq (1-\tau)|S|$. 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

$$\left|\left\{j \in S : \|w_j\|_2 \leq D_\tau\right\}\right| \geq \frac{1+\gamma}{1+2\gamma}|S| \geq \left(\frac{1+\gamma}{1+2\gamma}\right)\left(\frac{1}{2}+\gamma\right) n_h^{(s)} \geq \left(\frac{1}{2}+\frac{\gamma}{2}\right) n_h^{(s)}.$$

This gives us

$$\left|\left\{j \in T_h^{(s)} : \|Y_j - \theta_h\| \leq D_\tau\right\}\right| \geq \left(\frac{1}{2}+\frac{\gamma}{2}\right) n_h^{(s)} \tag{36}$$

Next we will show

$$\left|\left\{j \in T_h^{(s)} : \|Y_j - \widehat{\theta}_h^{(s)}\| \leq 4D_\tau\right\}\right| \geq (1-\delta) n_h^{(s)} \geq \left(\frac{1}{2}+\frac{\gamma}{4}\right) n_h^{(s)}. \tag{37}$$

To prove the above, 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

$$\left|\left\{j \in T_h^{(s)} : \|Y_j - Y_{j_0}\| \leq 2D_\tau\right\}\right| \geq \left(\frac{1}{2}+\frac{\gamma}{2}\right) n_h^{(s)}. \tag{38}$$

Now, remember the computation of $\mathsf{TM}_\delta(\left\{Y_j : j \in T_h^{(s)}\right\})$ in Algorithm 2 according to Algorithm 1. In view of (38) we then have $R_{i^*} \leq 2D_\tau$ and hence for $\delta = \frac{1}{2}-\frac{\gamma}{4}$

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

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

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

This completes the proof of (37).

Finally, combining (36) and (37) we get a contradiction as

- $\left|\left\{j \in T_h^{(s)} : \|Y_j - \widehat{\theta}_h^{(s)}\| \le 4D_\tau\right\} \cup \left\{j \in T_h^{(s)} : \|Y_j - \theta_h\| \le D_\tau\right\}\right| \le \left|\{j \in T_h^{(s)}\}\right| = n_h^{(s)}$
- $\left|\left\{j \in T_h^{(s)} : \|Y_j - \widehat{\theta}_h^{(s)}\| \le 4D_\tau\right\} \cap \left\{j \in T_h^{(s)} : \|Y_j - \theta_h\| \le D_\tau\right\}\right| = 0$
- $\left|\left\{j \in T_h^{(s)} : \|Y_j - \widehat{\theta}_h^{(s)}\| \le 4D_\tau\right\}\right| + \left|\left\{j \in T_h^{(s)} : \|Y_j - \theta_h\| \le D_\tau\right\}\right| \ge \left(1 + \frac{3\gamma}{4}\right) n_h^{(s)}$.

□

A.2. Proof of Theorem 1. In view of the above results we provide the proof of Theorem 1 below. For an ease of notations denote

$$A_s = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\left\{z_i^{(s)} \neq z_i\right\}} = \frac{1}{n}\sum_{h \neq g \in [k]} n_{hg}^{(s)}. \tag{39}$$

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

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

Then from Lemma 20 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 20 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 $H_{s+1} \ge 1 - \frac{5}{2\alpha \log(1/G(\epsilon_0^2 \Delta/\sigma))} \ge 0.8$,

(P2) If $H_s \ge \frac{1}{2} + \gamma_1$ then $\Lambda_s \le \frac{8\sigma}{\Delta} G^{-1}\left(\exp\left\{-\frac{(1+2/\alpha)(1+2\gamma_1)}{\gamma_1}\right\}\right) \le \frac{G^{-1}(e^{-c_4/\alpha})}{\mathsf{SNR}} \le 0.2$,

where c_4 is an absolute constant. Note that from Lemma 20 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

$$\mathcal{E} = \mathcal{E}_{\epsilon_0,\gamma_0} \cap \mathcal{E}_{\epsilon_1,\gamma_1} \text{ with } \mathbb{P}[\mathcal{E}] \ge 1 - 8ke^{-\frac{n\alpha}{4}}. \tag{40}$$

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

$$\Lambda_s \le \frac{G^{-1}(e^{-c_4/\alpha})}{\mathsf{SNR}} \le 0.2, \ H_s \ge 0.8 \text{ for all } s \ge 1. \tag{41}$$

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 (32)

$$\mathbf{1}_{\left\{z_i = g, \widehat{z}_i^{(s+1)} = h\right\}} \le \mathbf{1}_{\left\{\frac{(1-2\Lambda_s)^2}{2(1+2\Lambda_s)}\|\theta_g - \theta_h\| \le \|w_i\|\right\}}. \tag{42}$$

Then in view of the inequalities

- $(1-x)^2 \ge 1 - 2x, (1+x)^{-1} \ge 1 - x$ with $x = 2\Lambda_s < 1$
- $(1-2x)(1-x) \ge 1 - 3x$ with the above choices of x,

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

$$\mathbf{1}_{\left\{z_i=g,\widehat{z}_i^{(s+1)}=h\right\}} \leq \mathbf{1}_{\left\{\frac{1}{2}(1-6\Lambda_s)\Delta\leq\|w_i\|\right\}} \leq \mathbf{1}_{\{\sigma(\mathsf{SNR}-6G^{-1}(e^{-c_4/\alpha}))\leq\|w_i\|\}},$$

where the last inequality followed using the bound on $\Lambda_s \leq \frac{G^{-1}(e^{-c_4/\alpha})}{\mathsf{SNR}}$ in (41). Taking expectation conditioned on the event $\mathcal{E}$ in (40) and using the inequality

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

we get

$$\mathbb{P}\left[z_i \neq \widehat{z}_i^{(s+1)}\middle|\,\mathcal{E}\right] \leq k^2 \max_{\substack{g,h\in[k]\\ g\neq h}} \mathbb{P}\left[z_i = g, \widehat{z}_i^{(s+1)} = h\middle|\,\mathcal{E}\right] \leq k^2 G\left(\mathsf{SNR} - 6G^{-1}(e^{-c_4/\alpha})\right)$$

This implies

$$\mathbb{E}\left[A_{s+1}|\mathcal{E}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}\left[z_i \neq z_i^{(s+1)}\middle|\,\mathcal{E}\right] \leq k^2 G\left(\mathsf{SNR} - 6G^{-1}(e^{-c_4/\alpha})\right).$$

Combining the above with (40) we get

$$\mathbb{E}\left[A_{s+1}\right] \leq 8ke^{-\frac{n\alpha}{4}} + k^2 G\left(\mathsf{SNR} - 6G^{-1}(e^{-c_4/\alpha})\right).$$

APPENDIX B: PROOF OF RESULTS WITH OUTLIER (THEOREM 2)

B.1. 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 22. *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 $\frac{5}{2\alpha\log(1/G(\epsilon_0^2\Delta/\sigma))} < \frac{1}{2}$ 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 $s \geq 0$, 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 $H_{s+1} \geq \frac{1}{2} + \frac{\psi-2\xi}{2(2-\psi)}$ where $\xi = \frac{5}{4\alpha\log(1/G(\epsilon_0^2\Delta/\sigma))}$,*
(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 22. Repeating the argument in (32) in the proof of Lemma 20 we have

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

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 13, we get on the set $\mathcal{E}_{\epsilon_0}^{\mathsf{con}}$

$$n_{gh}^{(s+1)} \leq \sum_{i\in T_g^*}\mathbf{1}_{\{\epsilon_0^2\Delta\leq\|w_i\|\}} \leq \frac{5n_g^*}{4\log(1/G(\epsilon_0^2\Delta/\sigma))},\ \forall h \in [k], h \neq g. \tag{43}$$

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

$$\frac{\sum_{\substack{h\in[k]\\ h\neq g}} n_{gh}^{(s+1)}}{n_g^*} \leq \frac{5k}{4\log(1/G(\epsilon_0^2\Delta/\sigma))} \leq \frac{5}{4\alpha\log(1/G(\epsilon_0^2\Delta/\sigma))}. \tag{44}$$

Next we switch g, h in (43) and sum over $h \in [k], h \neq g$. We get with probability similar to $\mathbb{P}\left[\mathcal{E}_{\epsilon_0}^{\mathsf{con}}\right]$

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

We define $\xi = \frac{5}{4\alpha\log(1/G(\epsilon_0^2\Delta/\sigma))}$. In view of (44) this implies

$$n_{gg}^{(s+1)} = n_g^* - \sum_{\substack{h\in[k]\\h\neq g}} n_{gh}^{(s+1)} \geq n_g^* - n\alpha\xi \geq n_g^*(1-\xi). \tag{45}$$

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 (45) 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 22(ii) is similar to the proof of Lemma 20(ii). □

B.2. Proof of Theorem 2. For $c_1 > 0$ to be chosen later, we define

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

Then from Lemma 22 it follows that we can choose $c_2 > 0$ such that if

$$\Delta \geq c_2\sigma\max\left\{G^{-1}\left(\exp\left\{-\frac{1+2/\alpha}{\tau}\right\}\right), G^{-1}\left(e^{-\frac{c_1}{\alpha\psi}}\right)\right\}, \quad \tau = \frac{\gamma_0}{1+2\gamma_0},$$

then on the set $\mathcal{E}_{\epsilon_0,\gamma_0}$, 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 22, 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 set

$$\mathcal{E} = \mathcal{E}_{\epsilon_0,\gamma_0} \cap \mathcal{E}_{\epsilon_1,\gamma_1} \text{ with } \mathbb{P}[\mathcal{E}] \geq 1 - 8ke^{-\frac{n\alpha}{4}} \tag{46}$$

we have

$$\Lambda_s \le \frac{G^{-1}(e^{-c_4/(\alpha\psi)})}{\mathsf{SNR}} \le 0.2,\ H_s \ge \frac{1}{2} + \frac{\psi}{6} \quad \text{for all } s \ge 2. \tag{47}$$

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 set $\mathcal{E}$, $\ell(\widehat{z}^{(s+1)}, z) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\left\{z_i \neq \widehat{z}_i^{(s+1)}\right\}}$ is also small in probability. From this, using a Markov inequality we will conclude the result. From this point onward the proof is again similar to the proof in Theorem 1 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. Define

$$\begin{gathered}\mathcal{Z}^* = \bar{z} \times \{1,2\}^{n-m} \subseteq [k]^n, \\ m = k\lceil n\alpha/k \rceil, \quad \bar{z} \in [k]^m, \quad \bar{z}_i = u, u \in \{1,\dots,k\},\ i \in (u-1)\frac{m}{k} + 1, \dots, u\frac{m}{k}.\end{gathered} \tag{48}$$

In other words, for each $z \in \mathcal{Z}^*$, we already know the labels corresponding to the first $k\lceil n\alpha/k \rceil$ entries. For the rest of the entries the labels can either be 1 or 2. Note that for each label vector $z \in \mathcal{Z}^*$ we have $|\{i : z_i = g\}| \ge \lceil n\alpha/k \rceil$ for each $g = 1, \dots, n$. With this parameter (label) set $\mathcal{Z}^*$ we will apply Assouad's Lemma (Yu, 1997):

LEMMA 23 (Assouad). *Let $r \ge 1$ be an integer and let $\mathcal{F}_r = \{\mathcal{Q}_z : z \in \mathcal{Z}\}$ contains 2^r probability measures. 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 m pseudo-distances 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

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

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

$$\begin{gathered}z \in \mathcal{Z}^*, \quad \mathcal{Q}_{z_i} = \text{Distribution of } \theta_{z_i} + w_i,\ i = 1, \dots, n. \\ \overline{\mathcal{Q}} = \mathcal{Q}_{\bar{z}_1} \times \cdots \times \mathcal{Q}_{\bar{z}_m}, \quad Q_z = \bar{\mathcal{Q}} \times \mathcal{Q}_{z_{m+1}} \times \cdots \times \mathcal{Q}_{z_n}.\end{gathered}$$

In view of the above definition, to apply Lemma 23, we choose

$$\mathcal{Z} = \mathcal{Z}^*, \quad r = n - m, \quad f(Q_z) = z, \quad d_j(z, z') = \mathbf{1}_{\left\{z_{n-m+j} \neq z'_{n-m+j}\right\}}, \quad \delta = 1.$$

Hence, using Lemma 23 we get that given any estimator $\widehat{z}$ (which we can choose to be in $\mathcal{Z}^*$) it satisfies

$$\begin{aligned}\max_{\mathcal{Q}_z \in \mathcal{F}_r} \mathbb{E}_z\left[\sum_{i=1}^n \mathbf{1}_{\{\widehat{z}_i \neq z\}}\right] &\ge \frac{n-m}{2} \min\{1 - \mathrm{TV}(\mathcal{Q}_z, \mathcal{Q}_{z'}) : z \sim z' \in \mathcal{Z}^*\} \\ &= \frac{n-m}{2}(1 - \mathrm{TV}(P_1, P_2)),\end{aligned} \tag{49}$$

We further specify the error distributions corresponding to the labels $\{z_{m+1},\dots,z_n\}$ 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

$$\begin{aligned} &\{w_i\}_{i=m+1}^n \overset{\text{iid}}{\sim} R\cdot V,\\ &\mathbb{P}[R\ge x]=1-G(x/\sigma),\quad \mathbb{P}\left[V=\frac{\theta_1-\theta_2}{\|\theta_1-\theta_2\|}\right]=\frac12=\mathbb{P}\left[V=\frac{\theta_2-\theta_1}{\|\theta_1-\theta_2\|}\right]. \end{aligned} \tag{50}$$

In view of the above we can simplify (49) as

$$\max_{\mathcal{Q}_z\in\mathcal{F}_r}\mathbb{E}_z\left[\sum_{i=1}^n \mathbf{1}_{\{\widehat{z}_i\neq z\}}\right]=\frac{n-m}{2}(1-\mathrm{TV}(P_1,P_2)), \tag{51}$$

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 shift and rotational transformation. Then the distributions in (50) can be simplified to in terms of the density of w_i-s given by

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

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

$$dP_1(y)=-\frac{1}{2\sigma}G'\left(\frac{|y+\Delta/2|}{\sigma}\right)dy,\quad dP_2(y)=-\frac{1}{2\sigma}G'\left(\frac{|y-\Delta/2|}{\sigma}\right)dy. \tag{52}$$

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

$$\begin{aligned}
\mathrm{TV}(P_1,P_2)&=\frac12\int_{-\infty}^{\infty}|dP_1(y)-dP_2(y)|\\
&=\frac{1}{4\sigma}\int_{-\infty}^{\infty}\left|G'\left(\frac{|y+\frac{\Delta}{2}|}{\sigma}\right)-G'\left(\frac{|y-\frac{\Delta}{2}|}{\sigma}\right)\right|dy\\
&\overset{\text{(a)}}{=}\frac{1}{2\sigma}\int_{-\infty}^{0}\left|G'\left(\frac{|y+\frac{\Delta}{2}|}{\sigma}\right)-G'\left(\frac{|y-\frac{\Delta}{2}|}{\sigma}\right)\right|dy\\
&=\frac{1}{2\sigma}\left(\int_{-\infty}^{-\frac{\Delta}{2}-\sigma c_G}+\int_{-\frac{\Delta}{2}-\sigma c_G}^{-\frac{\Delta}{2}+\sigma c_G}+\int_{-\frac{\Delta}{2}+\sigma c_G}^{0}\right)\left|-G'\left(\frac{|y+\frac{\Delta}{2}|}{\sigma}\right)-\left(-G'\left(\frac{|y-\frac{\Delta}{2}|}{\sigma}\right)\right)\right|dy\\
&\overset{\text{(b)}}{\le}-\frac{1}{2\sigma}\left(\int_{-\infty}^{-\frac{\Delta}{2}-\sigma c_G}+\int_{-\frac{\Delta}{2}+\sigma c_G}^{0}\right)G'\left(\frac{|y+\frac{\Delta}{2}|}{\sigma}\right)dy\\
&\quad-\frac{1}{2\sigma}\int_{-\frac{\Delta}{2}-\sigma c_G}^{-\frac{\Delta}{2}+\sigma c_G}\left(G'\left(\frac{|y+\frac{\Delta}{2}|}{\sigma}\right)+G'\left(\frac{|y-\frac{\Delta}{2}|}{\sigma}\right)\right)dy
\end{aligned}$$

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

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

$$-G'\left(\frac{|y+\frac{\Delta}{2}|}{\sigma}\right)\ge -G'\left(\frac{|y-\frac{\Delta}{2}|}{\sigma}\right)\ge 0,\quad y\in(-\infty,-\frac{\Delta}{2}-\sigma c_G)\cup(-\frac{\Delta}{2}+\sigma c_G,0).$$

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

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

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 (52)
- (b) followed by change of variables
- (c) followed as $-G'(y) = |G'(y)|$ is decreasing over (c_G, ∞) and $\Delta \ge 2\sigma c_G$.

Plugging the last display in (51) completes the proof.

APPENDIX D: TECHNICAL RESULTS

PROOF OF LEMMA 19. It suffices to show that for any set $S \subseteq [n]$

$$\mathbb{P}\left[\sum_{i\in S} \mathbf{1}_{\left\{\|w_i\|_2 > \sigma G^{-1}\left(\exp\left\{-\frac{1+1/\beta}{\tau}\right\}\right)\right\}} \ge \tau |S|\right] \le e^{-n},$$

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

$$V_i = \mathbf{1}_{\left\{\|w_i\|_2 > \sigma G^{-1}\left(\exp\left\{-\frac{1+1/\beta}{\tau}\right\}\right)\right\}}.$$

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

$$\mathbb{P}\left[\sum_{i\in S} V_i \ge \tau |S|\right] \le e^{-n} \text{ for all } S \subseteq [n]. \tag{53}$$

Note that

$$\mathbb{P}\left[\|w_i\|_2 > \sigma G^{-1}\left(\exp\left\{-\frac{1+1/\beta}{\tau}\right\}\right)\right] \le \exp\left\{-\frac{1+1/\beta}{\tau}\right\} \le \frac{\tau}{1+1/\beta} < \tau. \tag{54}$$

This implies $\sum_{i\in S} V_i$ is stochastically smaller than $\mathrm{Binom}(|S|, \exp\left\{-\frac{1+1/\beta}{\tau}\right\})$. We continue to analyze (53) using the Chernoff's inequality for the Binomial random variable:

LEMMA 24. *(Boucheron, Lugosi and Massart, 2013, Section 2.2) For a Binom(m, q) random variable we have*

$$\mathbb{P}\left[\textit{Binom}(m,q) \ge ma\right] \le \exp\left(-mh_q(a)\right); \quad q < a < 1,\ h_q(a) = a\log\frac{a}{q} + (1-a)\log\frac{1-a}{1-q}.$$

We use the above result for the $\mathrm{Binom}(|S|, \exp\left\{-\frac{1+1/\beta}{\tau}\right\})$ distribution. Using $q = \exp\left\{-\frac{1+1/\beta}{\tau}\right\}, a = \tau$, $m = |S|$ in the above lemma and $x \log x \ge -0.5$ for $x \in (0,1)$ we

get

$$\mathbb{P}\left[\sum_{i\in S} V_i \geq \tau|S|\right] \leq \mathbb{P}\left[\text{Binom}(|S|, \exp\left\{-\frac{1+1/\beta}{\tau}\right\}) \geq \tau|S|\right] \tag{55}$$

$$\leq \exp\left(-m\left(\tau\log\frac{\tau}{q} + (1-\tau)\log\frac{1-\tau}{1-q}\right)\right)$$

$$\leq \exp\left(-m\left\{(1+1/\beta) + \tau\log\tau + (1-\tau)\log(1-\tau)\right\}\right) \leq e^{-n}. \tag{56}$$

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

APPENDIX E: PROOF OF INITIALIZATION RESULTS (THEOREM 6 AND THEOREM 9)

E.1. Continuing the proof of Theorem 6 from Section 7.2: Lemma 17 and Lemma 18.

PROOF OF LEMMA 17. In view of Lemma 15, without a loss of generality we assume that θ_1 is the closest centroid to $\mu_1^{(1)}$ and (29). We first prove the following claims:

$$\left|\mathcal{P}_1^{(\ell)} \cap \{Y_i : i \in T_1^*\}\right| \geq n_1^* - m - \frac{5n}{4\log(1/(G(\Delta/32\sigma)))} \tag{57}$$

$$\left|\mathcal{P}_1^{(\ell)} \cap \{Y_i : i \in T_2^*\}\right| \leq n\beta + \frac{5n}{4\log(1/(G(\Delta/32\sigma)))}. \tag{58}$$

The first claim (57) 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 distance from $\mu_1^{(\ell)} = \mu_1^{(\ell+1)}$ we get

$$\mathcal{P}_1^{(\ell+1)} \supseteq \{Y_i : i \in [n]\} \cap \mathcal{B}\left(\mu_1^{(\ell)}, \frac{\Delta}{16}\right). \tag{59}$$

  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

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

  In view of (59) the last display implies

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

  and hence

$$\left|\mathcal{P}_1^{(\ell)} \cap \{Y_i : i \in T_1^*\}\right| \geq \left|\mathcal{B}\left(\mu_1^{(\ell)}, \frac{\Delta}{16}\right) \cap \{Y_i : i \in T_1^*\}\right| - |A|$$

$$\geq \left|\mathcal{B}\left(\mu_1^{(\ell)}, \frac{\Delta}{16}\right) \cap \{Y_i : i \in T_1^*\}\right| - m. \tag{60}$$

- As we have from (29), with $\Delta \geq 96c_1$:

$$\mathcal{B}\left(\mu_1^{(\ell)}, \frac{\Delta}{16}\right) \supseteq \mathcal{B}\left(\theta_1, \frac{\Delta}{16} - 3c_1\right) \supseteq \mathcal{B}\left(\theta_1, \frac{\Delta}{32}\right), \tag{61}$$

in view of Lemma 14(ii) we get

$$\left|\mathcal{B}\left(\mu_1^{(\ell)}, \frac{\Delta}{16}\right) \cap \{Y_i : i \in T_1^*\}\right| \geq \left|\mathcal{B}\left(\theta_1, \frac{\Delta}{32}\right) \cap \{Y_i : i \in T_1^*\}\right|$$

$$\geq n_1^* \left(1 - \frac{5}{4\log\left(1/G\left(\frac{\Delta}{32\sigma}\right)\right)}\right) \geq n_1^* - \frac{5n}{4\log(1/(G(\Delta/32\sigma)))}. \tag{62}$$

Combining (60) and (62) we get (57). Next, to prove the claim (58), we note that:

- In view of Lemma 14 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 (29) 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

$$\left|\mathcal{P}_1^{(\ell)} \cap \{Y_i : i \in T_2^*\} \cap \mathcal{B}\left(\mu_1^{(\ell)}, \frac{\Delta}{16}\right)\right| \leq \left|\{Y_i : i \in T_2^*\}/\mathcal{B}\left(\theta_2, \frac{\Delta}{32}\right)\right| \leq \frac{5n}{4\log(1/(G(\Delta/32\sigma)))}. \tag{63}$$

- 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

$$\left|\mathcal{P}_1^{(\ell)} \cap \{Y_i : i \in T_2^*\}/\mathcal{B}\left(\mu_1^{(\ell)}, \frac{\Delta}{16}\right)\right| \leq n\beta. \tag{64}$$

Combining (63) and (64) we get (58).

Hence, we have proven the inequalities (57) and (58). These inequalities together imply

$$\left|\overline{\mathcal{P}_1^{(\ell)}} \cap \{Y_i : i \in T_2^*\}\right| \geq n_2^* - n\beta - \frac{5n}{4\log(1/(G(\Delta/32\sigma)))} \tag{65}$$

$$\left|\overline{\mathcal{P}_1^{(\ell)}} \cap \{Y_i : i \in T_1^*\}\right| \leq m + \frac{5n}{4\log(1/(G(\Delta/32\sigma)))}. \tag{66}$$

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 14, we have

$$\begin{aligned}
&\left|\overline{\mathcal{P}_1^{(\ell)}}/\mathcal{B}(\theta_2, \Delta/32)\right| \\
&\leq \left|\left\{\overline{\mathcal{P}_1^{(\ell)}} \cap \{Y_i : i \in T_2^*\}/\mathcal{B}(\theta_2, \Delta/32)\right\} \cup \left\{\overline{\mathcal{P}_1^{(\ell)}} \cap \{Y_i : i \in T_1^*\}/\mathcal{B}(\theta_2, \Delta/32)\right\}\right| + n^{\mathsf{out}} \\
&\leq \left|\{Y_i : i \in T_2^*\}/\left\{\mathcal{B}(\theta_2, \Delta/32)\right\}\right| + \left|\overline{\mathcal{P}_1^{(\ell)}} \cap \{Y_i : i \in T_1^*\}\right| + n^{\mathsf{out}} \\
&\leq \frac{5n}{4\log(1/(G(\Delta/32\sigma)))} + m + \frac{5n}{4\log(1/(G(\Delta/32\sigma)))} + \frac{n\alpha^2}{32} \\
&\leq \frac{3n\alpha^2}{32} + \frac{5n}{2\log(1/(G(\Delta/32\sigma)))} \leq \frac{5n\alpha^2}{32},
\end{aligned} \tag{67}$$

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 (65), 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 (67) 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 $(1-\beta)\,|\overline{\mathcal{P}_1^{(\ell)}}|$.
- Hence, the diameter of the tightest subset of $\overline{\mathcal{P}_1^{(\ell)}}$ with size $(1-\beta)\,|\overline{\mathcal{P}_1^{(\ell)}}|$ is at most $\frac{\Delta}{16}$.

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

PROOF OF LEMMA 18. 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

$$\mathcal{B}\left(\mu_2^{(\ell)}, \frac{\Delta}{8}\right) \cap \left\{\cup_{i\in\{1,2\}}\mathcal{B}\left(\theta_i, \frac{\Delta}{8}\right)\right\} = \phi, . \tag{68}$$

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

$$\left|\overline{\mathcal{P}_1^{(\ell)}} \cap \mathcal{B}\left(\mu_2^{(\ell)}, \frac{\Delta}{8}\right)\right| \geq (1-\beta)\left|\overline{\mathcal{P}_1^{(\ell)}}\right|. \tag{69}$$

Using Lemma 14 we get that

$$\left|\{Y_i : i \in [n]\}/\left\{\cup_{i\in\{1,2\}}\mathcal{B}\left(\theta_i, \frac{\Delta}{8}\right)\right\}\right| \leq \frac{5n}{4\log(1/(G(\Delta/32\sigma)))} + n^{\mathsf{out}}. \tag{70}$$

In view of the last display, using (68) and(69) we get

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

The last display implies

$$\left|\mathcal{P}_1^{(\ell)}\right| \geq n - \frac{5n}{2\log(1/(G(\Delta/32\sigma)))} - 2n^{\mathsf{out}} \geq n - \frac{n\alpha^2}{8}, \tag{71}$$

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

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

In view of (70) with $\cap_{i\in\{1,2\}}\mathcal{B}\left(\theta_i, \frac{\Delta}{8}\right) = \phi$, $n^{\mathsf{out}} \leq \frac{n\alpha^2}{32}$ and $n_1^*, n_2^* \geq n\alpha$, 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 (71) we get

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

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

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

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

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

- In view of the last argument we have

$$\mathcal{B}\left(\theta_2, \frac{\Delta}{8}\right) \cap \mathcal{B}\left(\mu_2^{(\ell)}, \mathsf{dist}_2^{(\ell)}\right) \subseteq \mathcal{B}\left(\theta_2, \frac{\Delta}{8}\right) \cap \mathcal{B}\left(\theta_1, \frac{11\Delta}{24}\right) = \phi, \tag{72}$$

- From Lemma 14, whenever $\frac{5n}{8\log(1/(G(\Delta/32\sigma)))} \leq \frac{n\alpha}{6}$, i.e., $\Delta \geq 32\sigma G^{-1}\left(e^{-\frac{3}{4\alpha}}\right)$, we get

$$\frac{1}{2}\left|\{Y_i : i \in T_2^*\} \cap \mathcal{B}\left(\theta_2, \frac{\Delta}{8}\right)\right| \geq \frac{n\alpha}{2} - \frac{5n}{8\log(1/(G(\Delta/32\sigma)))} \geq \frac{n\alpha}{3}. \tag{73}$$

However, this leads to a contradiction, as in view of (72) we have

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

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 (73)

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

□

E.2. Proof of Theorem 9. We will use the following high probability guarantee for proving our initialization result. The proof is identical to that of Lemma 14 and is omitted.

LEMMA 25. *The following statements hold for the β in Theorem 9. 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) $\left|\mathcal{B}(\theta_i, \frac{\Delta}{16k}) \cap \{Y_i : i \in T_i^*\}\right| \geq n_i^*\left(1 - \frac{5}{4\log(1/G(\Delta/(16\sigma k)))}\right)$ *for each* $i \in [k]$.

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 25 holds, which has a high probability. In view of our notations we also have $n^{\mathsf{out}} \leq \frac{n\alpha\beta}{16k}$. 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 26. *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 27. *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 28. *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 $(1-\beta)|\mathcal{P}_2^{(\ell_2)}|$-th closest point from $\mu_1^{(1,1)}$, within the set $\mathcal{P}_2^{(\ell_2)}$ will be smaller than $\frac{\Delta}{8k}$.*

LEMMA 29. *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 26, Lemma 27 and Lemma 28 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 29 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 26, Lemma 27, Lemma 28 and Lemma 29.

PROOF OF LEMMA 26. In view of Lemma 25, there is a constant $c_1 > 0$ such that

$$|\{j \in [n] : Y_j \in \mathcal{B}(\theta_i, c_1)\}| \geq n_i^*\left(1-\beta^2\right), \quad i \in [k]. \tag{74}$$

As we have $n_i^* > \frac{n\alpha}{k}$ 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 (27)

$$D(x, S) = \{\|x - Y_i\| : i \in S\}, \quad x \in \mathbb{R}^d, S \subseteq [n], \tag{75}$$

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

$$\{D(Y_{i^*}, [n])\}^{\{1-\frac{m_1}{n}\}} = \min_{j\in[n]} \{D(Y_j, [n])\}^{\{1-\frac{m_1}{n}\}}. \tag{76}$$

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 (74) it will imply that their union will contain more than n points from $Y_1, \ldots, Y_n$

$$\begin{aligned} &\left|\{i \in [n] : Y_i \in \mathcal{B}(Y_{i^*}, 2c_1)\} \cup \left[\cup_{j\in[k]} \{i \in [n] : Y_i \in \mathcal{B}(\theta_j, c_1)\}\right]\right| \\ &\geq \frac{n\alpha}{4k} + \sum_{j\in[k]} n_j^*\left(1-\beta^2\right) \geq (n - n^{\mathsf{out}})(1-\beta^2) + \frac{n\alpha}{4k} > n + \frac{n\alpha}{8k}, \end{aligned} \tag{77}$$

where we use the fact that $\{i \in [n] : Y_i \in \mathcal{B}(\theta_j, c_1)\}, j \in [k]$ are disjoint sets as $\min_{g \neq h \in [k]} \|\theta_g - \theta_h\| \geq \Delta$ and $n^{\mathsf{out}} \leq \frac{n\alpha}{16k}$. 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 27. In view of Lemma 26 we have for $c_1 = \sigma G^{-1}\left(e^{-\frac{5}{4\beta^2}}\right)$

$$\mu_k^{(k,1)} \in \mathcal{B}(\theta_k, 3c_1), \quad \left|\{Y_i : i \in T_j^*\} \cap \mathcal{B}(\theta_j, c_1)\right| \geq n_j^*\left(1-\beta^2\right), j \in [k]. \tag{78}$$

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. 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) (78) implies that $\cup_{j \in [k-1]} \left\{\{Y_i : i \in T_j^*\} \cap \mathcal{B}(\theta_j, c_1)\right\}$ will contain at least $\frac{n\alpha}{2k}$ points. Hence we get that this neighborhood N will contain at least one point y from the set $\cup_{j \in [k-1]} \{Y_i : i \in [n], Y_i \in \mathcal{B}(\theta_j, c_1)\}$. Let that $y \in \{Y_i : i \in [n], Y_i \in \mathcal{B}(\theta_j, c_1)\}$ for some $j \in [k-1]$. Then the distance of y from $\mu_k^{(\widetilde{\ell}_k)}$ is at least $\Delta - 4c_1$,

$$\|\mu_k^{(\widetilde{\ell}_k)} - y\| \geq \|\theta_k - \theta_j\| - \|\mu_k^{(\widetilde{\ell}_k)} - \theta_k\| - \|\theta_j - y\| \geq \Delta - 4c_1. \tag{79}$$

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 28. 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 $h = 1, \ldots, i-1$

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

(Q4) $\left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \left\{\cup_{g=i}^k \{Y_j : j \in T_g^*\}\right\}\right| \leq (k-i+1)m + \frac{5\sum_{g=i}^k n_g^*}{4\log(1/G(\Delta/(16k\sigma)))}$.

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 26 we get that the approximate $\mu_k^{(k,1)}$ lies within a radius $3c_1$, with $c_1 = \sigma G^{-1}\left(e^{-\frac{5}{4\beta^2}}\right)$, of θ_k, and hence (Q1) is satisfied. In view of Lemma 27, when we run the k-th for loop at the iteration $\bar{\ell}_k$, we get $\mathsf{dist}_k^{(\bar{\ell}_k)} \le \frac{\Delta}{8k}$, and hence (Q2) is satisfied.

Let $\bar{\ell}_k$ is as in Lemma 27. Without a loss of generality we assume that θ_k is the closest centroid to $\mu_k^{(k,1)}$ and (78) holds. We first prove the following claims:

$$\left|\mathcal{P}_k^{(\bar{\ell}_k)} \cap \{Y_i : i \in T_k^*\}\right| \ge n_k^* - m - \frac{5n_k^*}{4\log(1/(G(\Delta/16\sigma k)))} \tag{80}$$

$$\left|\mathcal{P}_k^{(\bar{\ell}_k)} \cap \{Y_i : i \in T_j^*\}\right| \le n\beta + \frac{5n_j^*}{4\log(1/(G(\Delta/16\sigma k)))}, \quad j \in [k-1]. \tag{81}$$

The first claim (80) follows from the following sequence of arguments (note the definition in (27))

- Using $\mu_k^{(k,\bar{\ell}_k)} = \mu_k^{(k,\bar{\ell}_k+1)}$ and from Lemma 27 we get

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

  which implies

$$\mathcal{P}_k^{(\bar{\ell}_k+1)} \supseteq \{Y_i : i \in [n]\} \cap \mathcal{B}\left(\mu_k^{(k,\bar{\ell}_k)}, \frac{\Delta}{8k}\right). \tag{82}$$

  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 (82) the last display implies

$$\begin{aligned}\left\{\mathcal{P}_k^{(\bar{\ell}_k)} \cap \{Y_i : i \in T_k^*\}\right\} &\supseteq \left\{\mathcal{P}_k^{(\bar{\ell}_k+1)} \cap \{Y_i : i \in T_k^*\}/A\right\} \\ &\supseteq \left\{\mathcal{B}\left(\mu_k^{(k,\bar{\ell}_k)}, \frac{\Delta}{8k}\right) \cap \{Y_i : i \in T_k^*\}/A\right\}, \end{aligned} \tag{83}$$

  and hence

$$\begin{aligned}\left|\mathcal{P}_k^{(\bar{\ell}_k)} \cap \{Y_i : i \in T_k^*\}\right| &\ge \left|\mathcal{B}\left(\mu_k^{(k,\bar{\ell}_k)}, \frac{\Delta}{8k}\right) \cap \{Y_i : i \in T_k^*\}\right| - |A| \\ &\ge \left|\mathcal{B}\left(\mu_k^{(k,\bar{\ell}_k)}, \frac{\Delta}{8k}\right) \cap \{Y_i : i \in T_k^*\}\right| - \frac{n\beta^2}{2}. \end{aligned} \tag{84}$$

- Note that we have from (78) and $\Delta \ge 48\sigma k$:

$$\mathcal{B}\left(\mu_k^{(k,\bar{\ell}_k)}, \frac{\Delta}{8k}\right) \supseteq \mathcal{B}\left(\theta_k, \frac{\Delta}{8k} - 3c_1\right) \supseteq \mathcal{B}\left(\theta_k, \frac{\Delta}{16k}\right). \tag{85}$$

  In view of Lemma 25(ii) this implies

$$\begin{aligned}\left|\mathcal{B}\left(\mu_k^{(k,\bar{\ell}_k)}, \frac{\Delta}{8k}\right) \cap \{Y_i : i \in T_k^*\}\right| &\ge \left|\mathcal{B}\left(\theta_k, \frac{\Delta}{16k}\right) \cap \{Y_i : i \in T_k^*\}\right| \\ &\ge n_k^*\left(1 - \frac{5}{4\log\left(1/G\left(\frac{\Delta}{16\sigma k}\right)\right)}\right) \ge n_k^* - \frac{5n_k^*}{4\log(1/(G(\Delta/16\sigma k)))}. \end{aligned} \tag{86}$$

Combining (84) and (86) we get (80).

Next, to prove the claim (81), we note that:

- In view of Lemma 25, for each $j=1,\dots,k-1$, there are at most $\frac{5n_j^*}{4\log(1/(G(\Delta/16\sigma k)))}$ many points from $\left\{Y_i : i\in T_j^*\right\}$ outside $\mathcal{B}(\theta_j,\frac{\Delta}{16k})$. In view of (78) 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 $\Delta\geq 24c_1k$. 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 $j=1,\dots,k-1$

$$\begin{aligned}&\left|\mathcal{P}_k^{(\bar\ell_k)}\cap\{Y_i:i\in T_j^*\}\cap\mathcal{B}\left(\mu_k^{(k,\bar\ell_k)},\frac{\Delta}{8k}\right)\right|\\ &\leq\left|\{Y_i:i\in T_j^*\}/\cup_{j=1}^{k-1}\mathcal{B}\left(\theta_j,\frac{\Delta}{16k}\right)\right|\leq\frac{5n_j^*}{4\log(1/(G(\Delta/16\sigma k)))}.\end{aligned}\tag{87}$$

- On the other hand, from Lemma 27 we have $\mathsf{dist}_k^{(\bar\ell_k)}=\left\{D(\mu_k^{(k,\bar\ell_k)},\mathcal{P}_k^{(\bar\ell_k)})\right\}^{\{1-\beta\}}\leq\frac{\Delta}{8k}$. This implies that for each $j\in[k-1]$

$$\left|\mathcal{P}_k^{(\bar\ell_k)}\cap\{Y_i:i\in T_j^*\}/\mathcal{B}\left(\mu_k^{(k,\bar\ell_k)},\frac{\Delta}{8k}\right)\right|\leq n\beta.\tag{88}$$

Combining (87) and (88) we get (81).

Hence, we have proven the inequalities (80) and (81). These inequalities together imply

$$\left|\overline{\mathcal{P}_k^{(\bar\ell_k)}}\cap\{Y_i:i\in T_j^*\}\right|\geq n_j^*-n\beta-\frac{5n_j^*}{4\log(1/(G(\Delta/16\sigma k)))},\quad j\in[k-1]\tag{89}$$

$$\left|\overline{\mathcal{P}_k^{(\bar\ell_k)}}\cap\{Y_i:i\in T_k^*\}\right|\leq m+\frac{5n_k^*}{4\log(1/(G(\Delta/16\sigma k)))}.\tag{90}$$

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 26, and is presented at the end of this section.

LEMMA 30. *Suppose that we have for* $i\geq 3$ *and* $h=1,\dots,i-1$

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

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* $\Delta\geq 16\sigma kG^{-1}\left(e^{-\frac{5k}{\alpha}}\right)$.

Next we prove (Q2) for the loop indexed by ℓ_{i-1}. In view of the Lemma 30 and Lemma 25 we note that

$$\mu_{i-1}^{(i-1,1)}\in\mathcal{B}\left(\theta_{i-1},3c_1\right),\quad\left|\{Y_j:j\in T_h^*\}\cap\mathcal{B}\left(\theta_h,c_1\right)\right|\geq n_h^*\left(1-\beta^2\right),h\in[i-1].\tag{91}$$

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 27 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 $\ell_{i-1}=2,\dots,\lfloor\frac{n'-m_1}{m}\rfloor$, 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 (91) we have $\mathcal{B}(\mu_{i-1}^{(i-1,1)}, 4c_1) \supseteq \mathcal{B}(\theta_{i-1}, c_1)$, which implies

$$\left|\{Y_j : j \in T_{i-1}^*\}/\mathcal{B}(\mu_{i-1}^{(i-1,1)}, 4c_1)\right| \le \left|\{Y_j : j \in T_{i-1}^*\}/\mathcal{B}(\theta_{i-1}, c_1)\right| \le n_{i-1}^*\beta^2. \tag{92}$$

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

$$\begin{aligned}
&\left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_{i-1}^*\} \cap \mathcal{B}(\mu_{i-1}^{(i-1,1)}, 4c_1)\right| \\
&= \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_{i-1}^*\}\right| - \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_{i-1}^*\}/\mathcal{B}(\mu_{i-1}^{(i-1,1)}, 4c_1)\right| \\
&\ge \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_{i-1}^*\}\right| - \left|\{Y_j : j \in T_{i-1}^*\}/\mathcal{B}(\mu_{i-1}^{(i-1,1)}, 4c_1)\right| \\
&\overset{(a)}{\ge} n_{i-1}^* - (k-i+1)n\beta - n_{i-1}^*\beta^2 - \frac{5(k-i+1)n_{i-1}^*}{4\log(1/(G\left(\frac{\Delta}{16\sigma k}\right)))} \overset{(b)}{\ge} \frac{n\alpha}{2k},
\end{aligned} \tag{93}$$

where (a) used the inequality (92) and (b) holds whenever $\Delta \ge 16\sigma k G^{-1}\left(e^{-\frac{10k}{\alpha}}\right)$ as $(k-i+1)n\beta \le \frac{n\alpha}{4k}, n\beta^2 \le \frac{n\alpha}{16k^4}$. 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 $\mathsf{dist}_{i-1}^{(1)} \le \frac{\Delta}{8k}$.

- 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

$$\left|\overline{\mathcal{P}_i^{(\widetilde{\ell}_i)}}\right| = \overline{\mathcal{P}_{i-1}^{(\widetilde{\ell}_{i-1})}} \cup \mathcal{P}_{i-1}^{(\widetilde{\ell}_{i-1})}, \quad \left|\overline{\mathcal{P}_{i-1}^{(\widetilde{\ell}_{i-1})}}\right| \le m. \tag{94}$$

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

$$\begin{aligned}
&\left|\mathcal{P}_{i-1}^{(\widetilde{\ell}_{i-1})} \cap \{Y_j : j \in T_1^*\}\right| \\
&\ge \left|\overline{\mathcal{P}_i^{(\widetilde{\ell}_i)}} \cap \{Y_j : j \in T_1^*\}\right| - m \overset{(a)}{\ge} n_1^* - kn\beta - \frac{n\beta^2}{2} - \frac{5kn_1^*}{4\log(1/G(\frac{\Delta}{16\sigma k}))} \overset{(b)}{\ge} 2n\beta,
\end{aligned} \tag{95}$$

where (a) followed from (94) and (b) followed as $\Delta \ge 16\sigma k G^{-1}\left(e^{-\frac{5k^2}{\alpha}}\right)$ and $kn\beta \le \frac{n\alpha}{4k}, n\beta^2 \le \frac{n\alpha}{16k^4}$. 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 (95) we get that the neighborhood N will include at least $n\beta$ points from $\{Y_j : j \in T_1^*\}$. Now, (78) 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 $\Delta - 4c_1$,

$$\|\mu_{i-1}^{(i-1,\widetilde{\ell}_{i-1})} - y\| \ge \|\theta_1 - \theta_{i-1}\| - \|\mu_{i-1}^{(i-1,\widetilde{\ell}_{i-1})} - \theta_{i-1}\| - \|\theta_1 - y\| \ge \Delta - 4c_1. \tag{96}$$

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

$$\bar{\ell}_{i-1} = \min\left\{r \ge 1 : \mathsf{dist}_{i-1}^{(r+1)} > \frac{\Delta}{8k}\right\}. \tag{97}$$

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 $h=1,\dots,i-2$

$$\left|\overline{\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})}}\cap\{Y_j: j\in T_h^*\}\right|\geq n_h^*-(k-i+2)n\beta-\frac{5(k-i+2)n_h^*}{4\log(1/(G(\Delta/16\sigma k)))}, \tag{98}$$

$$\left|\overline{\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})}}\cap\{Y_j: j\in T_{i-1}^*\}\right|\leq m+\frac{5n_{i-1}^*}{4\log(1/(G(\Delta/16\sigma k)))}. \tag{99}$$

To prove the claim (98), we note that:

- In view of Lemma 25, for each $h=1,\dots,i-2$, 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 (91) 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 $h\in[i-2]$
$$\begin{aligned}&\left|\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})}\cap\{Y_j: j\in T_h^*\}\cap\mathcal{B}\left(\mu_{i-1}^{(i-1,\bar{\ell}_{i-1})},\frac{\Delta}{8k}\right)\right|\\ &\leq\left|\{Y_j: j\in T_h^*\}/\cup_{h=1}^{i-2}\mathcal{B}\left(\theta_h,\frac{\Delta}{16k}\right)\right|\leq\frac{5n_h^*}{4\log(1/(G(\Delta/16\sigma k)))},\end{aligned} \tag{100}$$
where the last inequality followed from Lemma 25.
- On the other hand, in view of already proven (Q2) at the induction step $i-1$ we get $\mathsf{dist}_{i-1}^{(\bar{\ell}_{i-1})}=\left\{D(\mu_{i-1}^{(i-1,\bar{\ell}_{i-1})},\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})})\right\}^{\{1-\beta\}}\leq\frac{\Delta}{8k}$, which implies that for each $h\in[i-2]$
$$\left|\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})}\cap\{Y_j: j\in T_h^*\}/\mathcal{B}\left(\mu_{i-1}^{(i-1,\bar{\ell}_{i-1})},\frac{\Delta}{8k}\right)\right|\leq n\beta. \tag{101}$$

Combining (100) and (101) we get

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

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

$$\begin{aligned}\left|\overline{\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})}}\cap\{Y_j: j\in T_h^*\}\right|&=\left|\overline{\mathcal{P}_{i}^{(\bar{\ell}_{i})}}\cap\{Y_j: j\in T_h^*\}\right|-\left|\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})}\cap\{Y_j: j\in T_h^*\}\right|\\ &\geq n_h^*-(k-i+2)n\beta-\frac{5(k-i+2)n_h^*}{4\log(1/(G(\Delta/16\sigma k)))}.\end{aligned} \tag{102}$$

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

The claim (99) follows from the following sequence of arguments (note the definition in (27))

- Using $\mu_{i-1}^{(i-1,\bar{\ell}_{i-1})} = \mu_{i-1}^{(i-1,\bar{\ell}_{i-1}+1)}$ and (97) we get

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

which implies

$$\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1}+1)} \supseteq \overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \mathcal{B}\left(\mu_{i-1}^{(i-1,\bar{\ell}_{i-1})}, \frac{\Delta}{8k}\right). \tag{103}$$

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 (103) the last display implies

$$\begin{aligned}\left\{\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})} \cap \{Y_j : j \in T_{i-1}^*\}\right\} &\supseteq \left\{\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1}+1)} \cap \{Y_j : j \in T_{i-1}^*\}/A\right\} \\ &\supseteq \left\{\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \mathcal{B}\left(\mu_{i-1}^{(i-1,\bar{\ell}_{i-1})}, \frac{\Delta}{8k}\right) \cap \{Y_j : j \in T_{i-1}^*\}/A\right\}, \qquad (104)\end{aligned}$$

and hence

$$\begin{aligned}\left|\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})} \cap \{Y_j : j \in T_{i-1}^*\}\right| &\ge \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \mathcal{B}\left(\mu_{i-1}^{(i-1,\bar{\ell}_{i-1})}, \frac{\Delta}{8k}\right) \cap \{Y_j : j \in T_{i-1}^*\}\right| - |A| \\ &\ge \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \mathcal{B}\left(\mu_{i-1}^{(i-1,\bar{\ell}_{i-1})}, \frac{\Delta}{8k}\right) \cap \{Y_j : j \in T_{i-1}^*\}\right| - m. \qquad (105)\end{aligned}$$

- As we have from (91) and $\Delta \ge 48c_1k$:

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

in view of Lemma 25(ii) we get

$$\begin{aligned}&\left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \mathcal{B}\left(\mu_{i-1}^{(i-1,\bar{\ell}_{i-1})}, \frac{\Delta}{8k}\right) \cap \{Y_j : j \in T_{i-1}^*\}\right| \\ &\ge \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \mathcal{B}\left(\theta_{i-1}, \frac{\Delta}{16k}\right) \cap \{Y_j : j \in T_{i-1}^*\}\right| \\ &\ge \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_{i-1}^*\}\right| - \left|\{Y_j : j \in T_{i-1}^*\}/\mathcal{B}\left(\theta_{i-1}, \frac{\Delta}{16k}\right)\right| \\ &\ge \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_{i-1}^*\}\right| - \frac{5n_{i-1}^*}{4\log(1/(G(\Delta/16\sigma k)))}. \qquad (106)\end{aligned}$$

where the last inequality followed from Lemma 25.

Combining (105), (106) 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

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

In view of (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

$$\left|\overline{\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})}} \cap \left\{\cup_{g=i-1}^{k}\{Y_j : j \in T_g^*\}\right\}\right| \leq \left|\overline{\mathcal{P}_{i}^{(\bar{\ell}_{i})}} \cap \left\{\cup_{g=i}^{k}\{Y_j : j \in T_g^*\}\right\}\right| + \left|\overline{\mathcal{P}_{i-1}^{(\bar{\ell}_{i-1})}} \cap \{Y_j : j \in T_{i-1}^*\}\right|$$

$$\leq (k-i+2)m + \frac{5\sum_{g=i-1}^{k} n_g^*}{4\log(1/G(\Delta/(16k\sigma)))}. \tag{107}$$

This concludes the checking 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

$$\mathsf{dist}_i^{(\bar{\ell}_i)} \leq \frac{\Delta}{8k}, \quad i = k, k-1, \ldots, 2. \tag{108}$$

Finally, to complete the proof of Lemma 28 it remains to show that $\mathsf{dist}_1^{(\ell_2)} \leq \frac{\Delta}{8k}$. In view of the induction statement we have that

1. $\left|\overline{\mathcal{P}_2^{(\bar{\ell}_2)}} \cap \{Y_j : j \in T_1^*\}\right| \geq n_1^* - (k-1)n\beta - \frac{5(k-1)n_1^*}{4\log(1/(G(\Delta/16\sigma k)))}$.
2. $\left|\overline{\mathcal{P}_2^{(\bar{\ell}_2)}} \cap \left\{\cup_{g=2}^{k}\{Y_j : j \in T_g^*\}\right\}\right| \leq (k-1)m + \frac{5\sum_{g=2}^{k} n_g^*}{4\log(1/G(\Delta/(16\sigma k)))} \leq \frac{n\alpha\beta}{8k} + \frac{5\sum_{g=2}^{k} n_g^*}{4\log(1/G(\Delta/(16\sigma k)))}$.

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

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

from Lemma 25, we have

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

where the last inequality followed from

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

As we have

$$\left|\overline{\mathcal{P}_2^{(\bar{\ell}_2)}}\right| \geq \left|\overline{\mathcal{P}_2^{(\bar{\ell}_2)}} \cap \{Y_j : j \in T_1^*\}\right| \geq \frac{n_1^*}{2}, \tag{109}$$

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 (108) 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 28. 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 31. *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 $i \in 1, \dots, k$.*

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

$$\begin{aligned} &\left|\{Y_i : i \in [n]\} \cap \mathcal{B}\left(\mu_1^{(1,\ell_1)}, \frac{\Delta}{8}\right)\right| \\ &\le \left|\{Y_i : i \in [n]\} / \cup_{i=1}^k \mathcal{B}(\theta_i, \frac{\Delta}{8})\right| \le \frac{5n}{4\log(1/(G(\Delta/16\sigma k)))} + n^{\mathsf{out}}, \end{aligned} \tag{110}$$

where the last inequality followed from Lemma 25. For an ease of notation, throughout the proof we define

$$\ell_1 \triangleq \ell_2, \quad \mathcal{P}_1^{(\ell_1)} \triangleq \overline{\mathcal{P}_1^{(\ell_2)}}, \quad \mu_1^{(1,\ell_1)} \triangleq \mu_1^{(1,\ell_2)}. \tag{111}$$

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

$$\{Y_i : i \in [n]\} = \cup_{g=1}^k \mathcal{P}_g^{(\ell_g)}, \quad \mathcal{P}_g^{(\ell_g)} \cap \mathcal{P}_h^{(\ell_h)} = \phi, g \ne h \in [k]. \tag{112}$$

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

$$\left|\mathcal{P}_1^{(\ell_1)} / \mathcal{B}\left(\mu_1^{(1,\ell_1)}, \frac{\Delta}{8}\right)\right| \le \left|\mathcal{P}_1^{(\ell_1)} / \mathcal{B}\left(\mu_1^{(1,\ell_1)}, \mathsf{dist}_1^{(\ell_1)}\right)\right| \le n\beta. \tag{113}$$

In view of (110) the last display implies

$$\begin{aligned} \left|\mathcal{P}_1^{(\ell_1)}\right| &\le \left|\mathcal{P}_1^{(\ell_1)} / \mathcal{B}\left(\mu_1^{(1,\ell_1)}, \frac{\Delta}{8}\right)\right| + \left|\{Y_i : i \in [n]\} \cap \mathcal{B}\left(\mu_1^{(1,\ell_1)}, \frac{\Delta}{8}\right)\right| \\ &\le n\beta + \frac{3n}{4\log(1/(G(\Delta/16\sigma k)))} + n^{\mathsf{out}}. \end{aligned} \tag{114}$$

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 $j = 2, \dots, k$

$$\begin{aligned} \left|\{Y_i : i \in T_j^*\} \cap \left\{\cup_{g=2}^k \mathcal{P}_g^{(\ell_g)}\right\}\right| &= \left|\{Y_i : i \in T_j^*\} / \mathcal{P}_1^{(\ell_1)}\right| \\ &\ge n_j^* - n\beta - \frac{5n}{4\log(1/(G(\Delta/16\sigma k)))} - n^{\mathsf{out}} \ge \frac{7n\alpha}{8k}, \end{aligned} \tag{115}$$

where the last inequality follows from $\beta \leq \frac{\alpha}{12k}$ as $k \geq 3$, $\Delta \geq$ and $n^{\mathsf{out}} \leq \frac{n\beta}{2}$. Then we have for $j = 1, \ldots, k$

$$\begin{aligned}
&\left|\{Y_i : i \in T_j^*\} \cap \left[\cup_{g=2}^k \left\{\mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\}\right]\right| \\
&= \left|\{Y_i : i \in T_j^*\} \cap \left\{\cup_{g=2}^k \mathcal{P}_g^{(\ell_g)}\right\} \cap \left[\cup_{g=2}^k \left\{\mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\}\right]\right| \\
&= \left|\{Y_i : i \in T_j^*\} \cap \left\{\cup_{g=2}^k \mathcal{P}_g^{(\ell_g)}\right\}\right| \\
&\quad - \left|\{Y_i : i \in T_j^*\} \cap \left\{\cup_{g=2}^k \mathcal{P}_g^{(\ell_g)}\right\} / \left[\cup_{g=2}^k \left\{\mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\}\right]\right| \\
&\geq \left|\{Y_i : i \in T_j^*\} \cap \left\{\cup_{g=2}^k \mathcal{P}_g^{(\ell_g)}\right\}\right| \\
&\quad - \left|\left\{\cup_{g=2}^k \mathcal{P}_g^{(\ell_g)}\right\} / \left[\cup_{g=2}^k \left\{\mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\}\right]\right| \\
&\overset{(a)}{\geq} \frac{7n\alpha}{8k} - \left|\cup_{g=2}^k \left\{\mathcal{P}_g^{(\ell_g)} / \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\}\right| \\
&= \frac{7n\alpha}{8k} - \sum_{g=2}^k \left|\mathcal{P}_g^{(\ell_g)} / \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right| \overset{(b)}{\geq} \frac{7n\alpha}{8k} - n\beta \geq \frac{3n\alpha}{4k}, \qquad (116)
\end{aligned}$$

where (a) followed (115) 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

$$\begin{aligned}
&\left|\left\{\mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\} \cap \{Y_i : i \in T_j^*\}\right| \\
&\geq \frac{\min_{j=1}^k \left|\{Y_i : i \in T_j^*\} \cap \left[\cup_{g=2}^k \left\{\mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\}\right]\right|}{k} \geq \frac{3n\alpha}{4k^2}, \quad j = j_1, j_2,
\end{aligned}$$

where the last display followed using (116). However, as $\Delta \geq 16\sigma k G^{-1}\left(e^{-\frac{5k^2}{\alpha}}\right)$ implies

$$\left|\{Y_i : i \in T_j^*\} / \mathcal{B}(\theta_j, \Delta/3)\right| \leq \frac{5n_j^*}{4\log(1/G(\Delta/16\sigma k))} \leq \frac{n\alpha}{4k^2}, \quad j = j_1, j_2,$$

we get that for $j = j_1, j_2$

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

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

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

$$y \in \left\{\mathcal{P}_g^{(\ell_g)} \cap \mathcal{B}(\mu_g^{(g,\ell_g)}, \mathsf{dist}_g^{(\ell_g)})\right\} \cap \{Y_i : i \in T_{j_2}^*\} \cap \mathcal{B}(\theta_{j_2}, \Delta/3). \tag{117}$$

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

$$\|x - y\| \geq \|\theta_{j_1} - \theta_{j_2}\| - \|x - \theta_{j_1}\| - \|y - \theta_{j_2}\| \geq \frac{\Delta}{3}.$$

$$\|x - y\| \leq \|x - \mu_g^{(g,\ell_g)}\| + \|y - \mu_g^{(g,\ell_g)}\| \leq 2\mathsf{dist}_g^{(\ell_g)} \leq \frac{\Delta}{4}. \tag{118}$$

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

$$\mathcal{B}\left(\theta_g, \frac{\Delta}{8}\right) \cap \mathcal{B}\left(\mu_j^{(j,\ell_j)}, \mathsf{dist}_j^{(\ell_j)}\right) = \phi, \quad j = 1, \ldots, k. \tag{119}$$

Note that by Lemma 25

$$\left|\{Y_i : i \in [n]\} \cap \mathcal{B}\left(\theta_g, \frac{\Delta}{8}\right)\right| \geq \frac{n\alpha}{2k}. \tag{120}$$

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

$$\{Y_i : i \in [n]\} \cap \mathcal{B}\left(\theta_g, \frac{\Delta}{8}\right) = \left\{\cup_{j=1}^k \mathcal{P}_j^{(\ell_j)}\right\} \cap \mathcal{B}\left(\theta_g, \frac{\Delta}{8}\right)$$

$$= \cup_{j=1}^k \left\{\mathcal{P}_j^{(\ell_j)} \cap \mathcal{B}\left(\theta_g, \frac{\Delta}{8}\right)\right\} \subseteq \cup_{j=1}^k \left\{\mathcal{P}_j^{(\ell_j)} / \mathcal{B}\left(\mu_j^{(j,\ell_j)}, \mathsf{dist}_j^{(\ell_j)}\right)\right\}, \tag{121}$$

which implies

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

$$\leq \sum_{j=1}^k \beta \left|\mathcal{P}_j^{(\ell_j)}\right| \leq n\beta = \frac{n\alpha}{4k^2}. \tag{122}$$

This provides a contradiction to (120). Hence, all the centroids must lie in different $\mathcal{B}\left(\theta_j, \frac{\Delta}{3}\right)$ sets. □

□

PROOF OF LEMMA 30. In view of Lemma 25, there is a constant $c_1 = \sigma G^{-1}\left(e^{-\frac{5}{4\beta^2}}\right)$ such that

$$|\{Y_j : j \in T_h^*\}/\mathcal{B}(\theta_h, c_1)| \leq n_h^*\beta^2, \quad h \in [k]. \tag{123}$$

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

$$\begin{aligned}
&\left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_h^*\} \cap \mathcal{B}(\theta_h, c_1)\right| \\
&= \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_h^*\}\right| - \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_h^*\}/\mathcal{B}(\theta_h, c_1)\right| \\
&\geq \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_h^*\}\right| - |\{Y_j : j \in T_h^*\}/\mathcal{B}(\theta_h, c_1)| \\
&\geq \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_h^*\}\right| - n_h^*\beta^2 \qquad (124) \\
&\geq n_h^* - (k-i+1)n\beta - \frac{5(k-i+1)n_h^*}{4\log(1/(G(\Delta/16\sigma k)))} - n_h^*\beta^2 \geq m_1.
\end{aligned}$$

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

$$\left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \left\{\cup_{g=i}^k \{Y_j : j \in T_g^*\}\right\}\right| \leq \frac{2n\alpha\beta}{5k} \leq \frac{m_1}{2}, \tag{125}$$

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 (124) it will imply that

$$\begin{aligned}
\left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}}\right| &\geq \left|\left\{\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \mathcal{B}(Y_{i^*}, 2c_1)\right\} \cup \left\{\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \left\{\cup_{h\in[i-1]}\mathcal{B}(\theta_h, c_1)\right\}\right\}\right| \\
&\overset{(a)}{\geq} m_1 + \sum_{h\in[i-1]} \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_h^*\} \cap \mathcal{B}(\theta_h, c_1)\right| \\
&\overset{(b)}{\geq} m_1 - n\beta^2 + \sum_{h\in[i-1]} \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \{Y_j : j \in T_h^*\}\right| \\
&= m_1 - n\beta^2 + \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \left[\cup_{h\in[i-1]}\{Y_j : j \in T_h^*\}\right]\right| \\
&\geq m_1 - n\beta^2 + \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}}\right| - \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}} \cap \left\{\cup_{g=i}^k \{Y_j : j \in T_g^*\}\right\}\right| - n^{\mathsf{out}} \\
&\overset{(c)}{=} \frac{m_1}{2} - n\beta^2 + \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}}\right| - n^{\mathsf{out}} \overset{(d)}{\geq} \left|\overline{\mathcal{P}_i^{(\bar{\ell}_i)}}\right| + \frac{m_1}{8}. \qquad (126)
\end{aligned}$$

where (a) follows from the fact that $\{i \in [n] : Y_i \in \mathcal{B}(\theta_j, c_1)\}, j \in [k]$ are disjoint sets as $\min_{g\neq h\in[k]} \|\theta_g - \theta_h\| \geq \Delta$, (b) follows from (125), (c) followed from (124) and (d) follows

from $n^{\mathsf{out}} \leq \frac{m_1}{8}$. Hence, Y_{i^*} is at a distance at most $3c_1$ from one of the centroids. Without loss of we can the closest centroid to be θ_{i-1}.

□