Structures of Spurious Local Minima in k-means

Wei Qian, Yuqian Zhang, Yudong Chen

***Abstract*—The k-means clustering problem concerns finding a partition of the data points into k clusters such that the total within-cluster squared distance is minimized. This optimization objective is non-convex, and not everywhere differentiable. In general, there exist spurious local solutions other than the global optimum. Moreover, the simplest and most popular algorithm for k-means, namely Lloyd's algorithm, generally converges to such spurious local solutions both in theory and in practice. In this paper, we investigate the *structures* of these spurious local solutions under a probabilistic generative model with k ground truth clusters. As soon as $k = 3$, spurious local minima provably exist, even for well-separated clusters. One such local minimum puts two centers at one true cluster, and the third center in the middle of the other two true clusters. We prove that this is essentially the *only* type of spurious local minima under a separation condition. In particular, any local minimum solution only involves a configuration that puts multiple centers at a true cluster, and one center in the middle of multiple true clusters. Our results pertain to the k-means formulation for mixtures of Gaussians or bounded distributions, and hold in the over- and under-parametrization regimes where the number of centers in k-means may not equal to the number of true clusters. Our theoretical results corroborate existing empirical observations and provide justification for popular heuristics for k-means clustering.**

***Index Terms*—k-means, Gaussian mixture, non-convex optimization, spurious local minima, structured minima**

I. Introduction

k-means clustering is one of the most fundamental problems in unsupervised learning, with a wide range of applications in multiple fields including machine learning, image analysis, computer graphics and beyond; see the survey [1] and the references therein. The k-means problem can be formulated as follows: given n data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, find k centers $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) \in \mathbb{R}^{d \times k}$ such that the following sum of squared distances is minimized:[1]

$$G_n(\boldsymbol{\beta}) := \sum_{i=1}^{n} \min_{j \in [k]} \|\boldsymbol{x}_i - \boldsymbol{\beta}_j\|^2, \tag{1}$$

W. Qian was with the School of Operations Research and Information Engineering, Cornell University, Ithaca, NY, 14853 (email: wq34@cornell.edu).

Y. Chen is with the School of Operations Research and Information Engineering, Cornell University, Ithaca, NY, 14853 (email: yudong.chen@cornell.edu.

Y. Zhang is with the Department of Electrical and Computer Engineering, Rutgers University, New Brunswick, NJ, 08854 (email: yqz.zhang@rutgers.edu).

W. Qian and Y. Chen are partially supported by NSF grants 1657420 and 1704828 and CAREER Award CCF-2047910.

[1] Another common way of formulating the k-means problem involves finding a partition of the data points into k clusters such that the within-cluster sum of squared distance is minimized. This partition-based formulation is equivalent to the center-based formulation (1) used in this paper, as we show in Appendix A.

where $\|\cdot\|$ denotes the ℓ_2 norm. The k-means objective function (1) is non-convex: it involves the minimization of quadratic functions and is symmetric with respect to permutation of the indices of components of $\boldsymbol{\beta}$. This optimization problem is NP-hard in general [2, 3, 4]. It has been observed that standard algorithms for k-means often converge to spurious local solutions of (1) that are not globally optimal [5, 6]. Moreover, these local minima of k-means are prevalent in practice [7, 8].

Recent theoretical work has made progress in understanding the k-means and related clustering problems with *two* clusters. In particular, if the data is generated from a balanced mixture of two identical and spherical Gaussians, the work in [9, 10, 11] effectively shows that there is no spurious local minima, and that greedy algorithms such as the Lloyd's algorithm and Expectation-Maximization (EM) are guaranteed to converge to a global minimizer from a random initialization. However, as soon as there are more than two clusters, non-trivial spurious local solutions do exist, even when the ground truth clusters are well-separated and balanced. Worst yet, these spurious local solutions may have objective values arbitrarily worse than the global optimum, and randomly-initialized greedy algorithms may provably converge to these local solutions with high probability [12].

Despite the above negative results, not all hope is lost. In this paper, we show that even with a general number of clusters, a lot can be said about the structural properties of these spurious local minima. In particular, under certain mixture models, we prove that *all* spurious local minima of k-means are well-behaved, in the sense that they possess the same type of structure that partially recovers the global minimum. We elaborate below.

A. Main Contributions

Consider the k-means problem under the following probabilistic generative model. Let $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_k^* \in \mathbb{R}^d$ be k^* distinct unknown true cluster centers. For each $s \in [k^*]$, let f_s be the density of a distribution with mean $\boldsymbol{\beta}_s^*$. Each data point $\boldsymbol{x} \in \mathbb{R}^d$ is sampled independently from a mixture f of these k^* distributions $\{f_s\}_{s \in [k^*]}$, with the density

$$f(\boldsymbol{x}) = \frac{1}{k^*} \sum_{s=1}^{k^*} f_s(\boldsymbol{x}). \tag{2}$$

If each f_s is a Gaussian distribution centered at $\boldsymbol{\beta}_s^*$, the above distribution becomes the (balanced/equally-weighted) Gaussian Mixture Model (GMM). Under the generative model (2), we consider the following population version of the k-means objective function:

$$G(\boldsymbol{\beta}) = \int \min_{j \in [k]} \|\boldsymbol{x} - \boldsymbol{\beta}_j\|^2 f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

$$= \frac{1}{k^*} \sum_{s=1}^{k^*} \int \min_{j \in [k]} \|\boldsymbol{x} - \boldsymbol{\beta}_j\|^2 f_s(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \quad (3)$$

The objective function above can be viewed as the infinite-sample ($n \to \infty$) limit of the empirical objective function in equation (1). Note that this population objective is also non-convex.

Below for simplicity we assume $k = k^*$, that is, the number of fitted centers in the k-means formulation is the same as that of the true mixture, though our theoretical results hold more generally.

a) Existence of spurious local minima.: Under general conditions, the ground truth centers $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_k^*) \in \mathbb{R}^{d \times k}$ and any permutation thereof are (close to) a global minimum of G; see Proposition 1. However, there exist additional spurious local minima, even in the simple one-dimensional setting with $k = 3$ clusters and when the densities $\{f_s\}_{s \in [k]}$ have bounded and disjoint supports. In particular, we show that one spurious local minimum $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ has the following configuration:

$$\boldsymbol{\beta}_1 \approx \boldsymbol{\beta}_2 \approx \boldsymbol{\beta}_1^*, \quad \boldsymbol{\beta}_3 \approx \frac{\boldsymbol{\beta}_2^* + \boldsymbol{\beta}_3^*}{2}. \quad (4)$$

In words, this local solution uses two centers to fit one true cluster, and the third center to fit the other two true clusters. See Proposition 2 for details. A similar observation was made in [12] for the log-likelihood objective function of Gaussian mixtures.

b) Structures of spurious local minima.: The local solution in equation (4) involves disjoint many-fit-one and one-fit-many associations. As our main result, we show that this is essentially the *only* type of spurious local minima for k-means under a separation condition:

Theorem (Informal)**.** *For well-separated mixture models, any non-degenerate local minima $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ of G only involves the following configurations: (i) multiple centers $\{\boldsymbol{\beta}_j\}$ lie near a true cluster $\boldsymbol{\beta}_s^*$, and (ii) one center $\boldsymbol{\beta}_j$ lies near the mean of multiple true clusters $\{\boldsymbol{\beta}_s^*\}$. Moreover, the configurations (i) and (ii) involve disjoint sets of $\boldsymbol{\beta}_j$'s and $\boldsymbol{\beta}_s^*$'s.*

See Theorems 1 and 2 for the precise statement of this result. In words, viewing a solution $\boldsymbol{\beta}$ as an assignment of the centers $\{\boldsymbol{\beta}_j\}$ for fitting the ground truth clusters, we show that a local minimum $\boldsymbol{\beta}$ can only involve many-fit-one associations (case (i) above) and one-fit-many associations (case (ii) above), and each fitted center $\boldsymbol{\beta}_j$ or true center $\boldsymbol{\beta}_s^*$ only participates in one of these associations. Any other solution $\boldsymbol{\beta}$ with many-fit-many associations (or other configurations) *cannot* be a local minimum.

We illustrate the above results in Figure 1 under a two-dimensional GMM with 4 components. The top panels show different candidate solutions of k-means. The ground-truth centers are the only global minimum, as in Panel 1a. Panel 1b shows a spurious local minimum, where the orange center fits two clusters, and the blue and purple centers fit one cluster. In Panel 1c, the blue and green centers together fit 3 clusters; in Panel 1d, the blue and purple centers together fit 2 clusters. These two solutions contain many-fit-many associations and are *not* local minima.

For further verification, we run the Lloyd's algorithm [13] with the above four solutions as the initial solution. The Lloyd's algorithm is an iterative greedy method that alternates between assigning each data point to its closest center and updating the centers to be the means of the new clusters. It can be viewed as a quasi-Newton algorithm applied to the objective function (1) with a specific choice of step size [14]. The bottom panels in Figure 1 show the trajectories of intermediate solutions of Lloyd's algorithm and the final solutions they converge to. When initialized at a global or local minimum, the algorithm stays at the initial solutions as expected (Panels 2a and 2b). In Panel 2c, the algorithm escapes from the initial solution, which is not a local minimum, and then converges to the spurious local minimum plotted in Panel 1b. In Panel 2d, the algorithm again escapes from the initial solution and converges to the globally optimal ground-truth solution plotted in Panel 1a.

To put the above results in context, we note that even the existence (or the lack thereof) of spurious population local minima in GMM, posted as an open problem in the Conference on Learning Theory (COLT) 2007 [15], was not rigorously resolved until recently [10, 11, 12]. As mentioned, in general spurious local minima do exist [12], as demonstrated by an example similar to that in equation (4). Our results above provide a positive message in the context of the k-means objective: all local minima *partially* recover the global minimum, in the sense that they identify some of true cluster centers and the means of the other true cluster centers. Again see Panel 2b in Figure 1 for an illustration.

On the technical side, note that the k-means objective function G is complicated due to the $\min$ operator in its definition (3), and in particular is not everywhere differentiable. Our analysis makes uses of a new technique based on analyzing the Voronoi sets and boundaries associated with the $\min$ operator as well as the directional derivatives of certain smooth upper bounds of G. Doing so allows us to leverage both the first and second order optimality conditions to deduce the structures of a local minimum. We hope that this technique is useful for studying other mixture and clustering problems.

B. Related Work

With a history of more than 50 years [13, 5], the k-means problem has found broad applications in computer science, astronomy, biology, social science and beyond. We refer to the papers [16, 1] for a comprehensive survey of the work on this problem.

Without additional assumptions on the data points, optimizing the k-means objective in (1) is NP-hard when the number of components k is fixed [2] or when the dimension d is fixed [3]. Even finding a $(1+\epsilon)$ approximation with varying (k, d) is hard [4]. Progress has been made on designing constant-ratio approximation algorithms; see, e.g., the results in [17, 18] among many others.

Lloyd's algorithm [13], often called *the* k-means algorithm, is arguably the most popular method for the k-means problem.

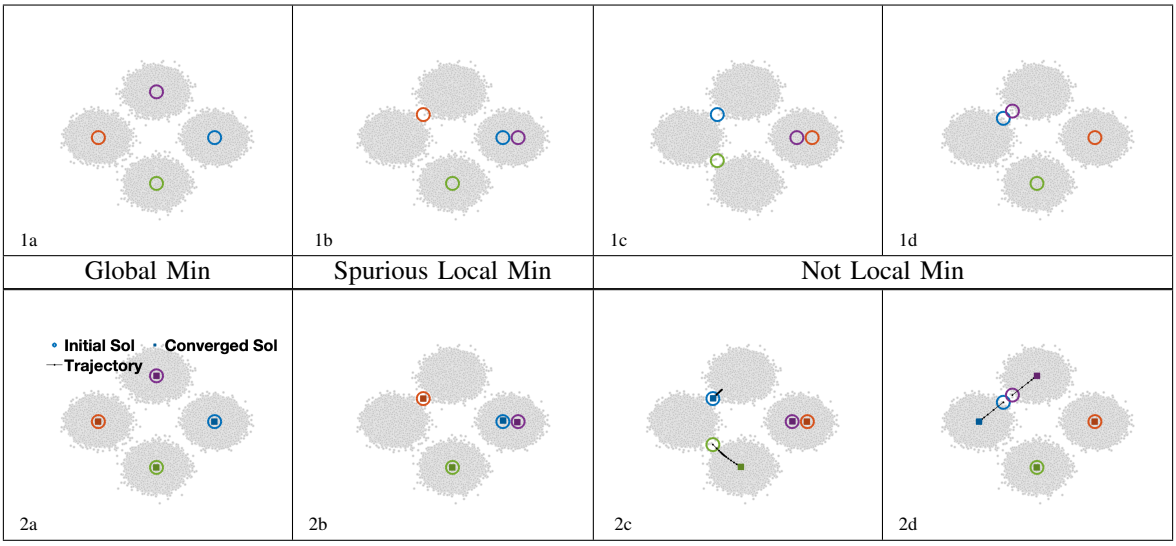

Figure 1. *Top panels:* Local minima and non-minima in GMM with 4 components. Solutions with many-fit-many configurations are not local minima. *Bottom panels:* Trajectories of greedy algorithm when initialized at different solutions. The colored circles correspond to an initial configuration of $\boldsymbol{\beta}$. Running the Lloyd's k-means algorithm from this initialization converges to a solution denoted by colored squares. The black lines correspond to the trajectory of the intermediate iterates. The algorithm escapes from non-minima and converges to a global or local minimum.

In general, Lloyd's algorithm is only guaranteed to converge to a local minimum and is sensitive to initialization [19]. Moreover, it may take exponentially many steps to converge in the worst case [20, 21]. Under certain probabilistic assumptions of the data, several theoretical guarantees have been established for the Lloyd's algorithm [9, 22, 23]. In particular, the work in [22] shows that under a separation condition of the clusters, Lloyd's algorithm can learn a mixture of Gaussians as well as a mixture of bounded distributions, after (i) a dimension reduction step via PCA and (ii) an initialization that has a constant-factor approximation guarantee. There is also substantial work on designing provably efficient initialization schemes for Lloyd's algorithm [24, 25]. Particularly relevant to us is the work in [26], which considers over-parametrized k-means/EM (which fits k^* clusters using $k > k^*$ centers) equipped with extra pruning steps. Interestingly, the fitted centers they try to prune correspond to, in our language, many-fit-one associations (and sometimes almost-empty associations as well; see our main theorems). As Lloyd's algorithm finds local minima of k-means, our results can be used to characterize the structural properties of the output of Lloyd's. We emphasize that our results are in fact more general, applicable to the general k-means objective function (with or without over-parametrization) and hence not tied to a specific algorithm.

A recent line of work also considers convex relaxation methods for the k-means problem based on linear or semidefinite programming [27, 28, 29]. Theoretical guarantees have been established on when the solution of the convex program coincides with (or approximates) the global minimum of k-means [30, 31, 32, 33].

The k-means objective function can be viewed as a "hard" or limit version of the negative log-likelihood function for the Gaussian Mixture Model (GMM); see Section II-B. As such, our results are related to recent theoretical work on the Expectation-Maximization (EM) algorithm [34], which is a local/greedy algorithm for optimizing the likelihood function. Positive results have been obtained on provable convergence of EM under GMM with $k = 2$ components [35, 10, 11, 36, 37]. In particular, these results show that the negative log-likelihood function has no spurious local minima for a balanced mixture of two Gaussians with the same covariance matrix. However, in more general mixture models, it has been proved that spurious local minima do exist with high probability. Examples include a mixture of $k \geq 3$ equally weighted components [12], and a mixture of $k = 2$ unequally weighted components with known mixing weights [38].

II. Problem Setup

In this section, we introduce the statistical models for our main results. Recall the mixture model in equation (2) with k^* true clusters. We consider two concrete instantiations of this model.

A. Ball Mixture

The first instantiation is a mixture of uniform distributions on k^* disjoint balls. For each $\boldsymbol{u} \in \mathbb{R}^d$, let $\mathbb{B}_{\boldsymbol{u}}(r)$ denote the Euclidean ball centered at $\boldsymbol{u}$ with radius r. As the true centers $\{\boldsymbol{\beta}_s^*\}_{s\in[k^*]}$ and the radius r are fixed throughout this paper, we use the shorthand $\mathbb{B}_s \equiv \mathbb{B}_{\boldsymbol{\beta}_s^*}(r)$ for brevity. We assume that each data point $\boldsymbol{x}$ is sampled independently and uniformly from one of k^* disjoint balls centered at the true centers $\boldsymbol{\beta}_s^*$; that is, $\boldsymbol{x} \sim \text{unif}(\mathbb{B}_s)$ with probability $\frac{1}{k^*}$.

This model, sometimes called the Stochastic Ball Model [30], is formally described below.

Definition 1 (Stochastic Ball Model)**.** *The Stochastic Ball Model is the mixture (2) where each component has density*

$$f_s(\boldsymbol{x}) = \frac{1}{\text{Vol}(\mathbb{B}_s)} \mathbb{1}_{\mathbb{B}_s}(\boldsymbol{x}), \quad s \in [k^*].$$

Here $\text{Vol}(T)$ denotes the volume of a set $T \subseteq \mathbb{R}^d$ with respect to the Lebesgue measure, and $\mathbb{1}_T$ is the indicator function for the set T.

B. Gaussian Mixture

The second instantiation is the (spherical) Gaussian mixture model, where each data point $\boldsymbol{x}$ is sampled independently from one of k^* Gaussian distributions whose means are the true centers $\{\boldsymbol{\beta}_s^*\}$; that is, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\beta}_s^*, \sigma^2 \boldsymbol{I})$ with probability $\frac{1}{k^*}$. A formal description of GMM is given below.

Definition 2 (Gaussian Mixture Model)**.** *The (spherical) Gaussian Mixture Model is the mixture (2) where each component has density*

$$f_s(\boldsymbol{x}) = \frac{1}{(\sqrt{2\pi}\sigma)^d} \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{\beta}_s^*\|^2}{2\sigma^2}\right), \quad s \in [k^*].$$

We point out that the population negative likelihood function of GMM (with a posited variance parameter τ^2), namely[2]

$$L_\tau(\boldsymbol{\beta}) := -\int \log\left[\sum_{j\in[k]} \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{\beta}_j\|^2}{2\tau^2}\right)\right] f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

is closely related to the population k-means objective function G defined in equation (3). As the log-sum-exp function above is a form of soft maximum, L_τ can be viewed as a smooth approximation of G. Moreover, as $\tau \to 0$, we have $2\tau^2 L_\tau(\boldsymbol{\beta}) \to G(\boldsymbol{\beta})$ for each $\boldsymbol{\beta}$. In other words, the k-means objective function corresponds to the limit case of the GMM log-likelihood function, and hence results for one have immediate bearing upon the other.

C. Model Parameters

For both of the above models, we define the quantities

$$\Delta_{\max} := \max_{s\neq s'} \|\boldsymbol{\beta}_s^* - \boldsymbol{\beta}_{s'}^*\|,$$
$$\Delta_{\min} := \min_{s\neq s'} \|\boldsymbol{\beta}_s^* - \boldsymbol{\beta}_{s'}^*\|,$$

which are the maximum and minimum pairwise separations between the true centers. Accordingly, we introduce two quantities measuring the Signal-to-Noise Ratios (SNR) of the models. In particular, for the Stochastic Ball Model we define

$$\eta_{\max} := \frac{\Delta_{\max}}{r},$$
$$\eta_{\min} := \frac{\Delta_{\min}}{r},$$

which are the maximum and minimum separations normalized by the radius of the balls. For the Gaussian Mixture Model, we similarly define

$$\eta_{\max} := \frac{\Delta_{\max}}{\sigma\sqrt{\min(k+k^*, d)}},$$
$$\eta_{\min} := \frac{\Delta_{\min}}{\sigma\sqrt{\min(k+k^*, d)}}.$$

Note the $\sqrt{\min(k+k^*, d)}$ factor in the denominators above. This factor is the typical value of the norm of a random vector from a d-dimensional standard Gaussian distribution when projected to the $(k+k^*)$-dimensional subspace spanned by the true and fitted centers $\{\boldsymbol{\beta}_s^*\}_{s\in[k^*]}$ and $\{\boldsymbol{\beta}_i\}_{i\in[k]}$.

[2]Here we ignore a constant additive term that is independent of the variable $\boldsymbol{\beta}$.

The above models are sometimes said to be *well-separated* if $\eta_{\min} = \Omega(1)$ [12]. Also note that the ratio $\frac{\eta_{\max}}{\eta_{\min}} \in [1, \infty)$ measures how evenly-spaced the true centers are. This ratio is close to 1 when the true centers are approximately equidistant to each other.

D. Voronoi sets

Each candidate solution $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ of the k-means problem induces a *Voronoi diagram*, namely, a partition of the space $\mathbb{R}^d$ based on proximity to the $\boldsymbol{\beta}_s$'s. The Voronoi diagram plays a crucial role in understanding the k-means objective (3), which is defined by the quantity $\min_{j\in[k]} \|\boldsymbol{x} - \boldsymbol{\beta}_j\|$, the distance of a point $\boldsymbol{x}$ to its closest center. Here we review some basic concepts related to Voronoi diagrams, which are useful for future development.

Given a set of k centers $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) \in \mathbb{R}^{d\times k}$ in $\mathbb{R}^d$, let $\mathcal{V}_i(\boldsymbol{\beta})$ be the region consisting of points that are closer to $\boldsymbol{\beta}_i$ than to any other center $\boldsymbol{\beta}_j$, $j \neq i$. Formally, for each $i \in [k]$ we define

$$\mathcal{V}_i(\boldsymbol{\beta}) := \left\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{\beta}_i\| \leq \|\boldsymbol{x} - \boldsymbol{\beta}_j\|, \forall j \neq i\right\}. \quad (5)$$

We call each $\mathcal{V}_i(\boldsymbol{\beta})$ the *Voronoi set* associated with $\boldsymbol{\beta}_i$. The *Voronoi diagram* of $\boldsymbol{\beta}$ is the collection of the Voronoi sets, that is, $\mathcal{V}(\boldsymbol{\beta}) := \{\mathcal{V}_i(\boldsymbol{\beta}) : i \in [k]\}$. Note that each Voronoi set is a polyhedron in $\mathbb{R}^d$ with at most $k-1$ facets,[3] as we can rewrite the definition in equation (5) as

$$\mathcal{V}_i(\boldsymbol{\beta}) = \left\{\boldsymbol{x} \in \mathbb{R}^d : 2\langle \boldsymbol{\beta}_j - \boldsymbol{\beta}_i, \boldsymbol{x}\rangle \leq \|\boldsymbol{\beta}_j\|^2 - \|\boldsymbol{\beta}_i\|^2, \forall j \neq i, j \in [k]\right\}.$$

In addition, for each index pair (i, j) with $i \neq j$, we define the *Voronoi boundary* $\partial_{i,j}(\boldsymbol{\beta})$ as the intersection of the Voronoi sets associated with $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$; that is,

$$\partial_{i,j}(\boldsymbol{\beta}) := \mathcal{V}_i(\boldsymbol{\beta}) \cap \mathcal{V}_i(\boldsymbol{\beta}) = \left\{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{\beta}_i\| = \|\boldsymbol{x} - \boldsymbol{\beta}_j\|\right\}.$$

Note that $\partial_{i,j}(\boldsymbol{\beta})$ is the set of points with equal distance to $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$. If $\partial_{i,j}(\boldsymbol{\beta})$ has dimension $d-1$, we say that $\mathcal{V}_i(\boldsymbol{\beta})$ is *adjacent* to $\mathcal{V}_j(\boldsymbol{\beta})$, written as $\mathcal{V}_i(\boldsymbol{\beta}) \sim \mathcal{V}_j(\boldsymbol{\beta})$. In this case, the two Voronoi sets $\mathcal{V}_i(\boldsymbol{\beta})$ and $\mathcal{V}_j(\boldsymbol{\beta})$ intersect at a common (full dimensional) facet of the two polyhedra. We use the notation $\partial(\boldsymbol{\beta}) := \{\partial_{i,j}(\boldsymbol{\beta}) : \mathcal{V}_i(\boldsymbol{\beta}) \sim \mathcal{V}_j(\boldsymbol{\beta})\}$ to denote the collection of the Voronoi boundaries of adjacent Voronoi sets.

III. Main Results

In this section, we present our main theoretical results on the structures of the local minima of the population k-means objective G defined in equation (3). In what follows, we use $\mathbb{P}$ to denote the probability measure with respect to f, the density of the ground truth mixture. Similarly, for each $s \in [k^*]$, we use $\mathbb{P}_s$ to denote the probability measure with respect to f_s, the density of the s-th component of the ground truth mixture.

[3]A facet is a $(d-1)$ dimensional face of a d-dimensional polyhedron.

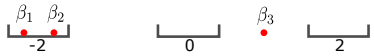

Figure 2. One-dimensional Stochastic Ball Model with radius $r < 0.4$ and ground truth cluster centers $\boldsymbol{\beta}^* = (-2, 0, 2)$. The solution $\boldsymbol{\beta} = (-2 - \frac{r}{2}, -2 + \frac{r}{2}, 1)$ is a spurious local minimum.

A. Stochastic Ball Model

Consider the Stochastic Ball Model in Definition 1. We first state two simple results concerning the global and local minima of the k-means objective G. The first proposition, proved in Appendix A, states that the ground truth centers is the only global minimum of G.

Proposition 1 (Ground truth is global minimum)**.** *Under the Stochastic Ball Model with $k^* = k \geq 1$ and $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \ldots, \boldsymbol{\beta}_k^*) \in \mathbb{R}^{d \times k}$. if $\eta_{\min} \geq 6\sqrt{k}$, then the true centers $\boldsymbol{\beta}^*$ (up to permutation of its k components) is the unique global minimum of G.*

The next proposition, proved in Appendix A, states that in general G has a spurious local minimum that is not a global minimum. An illustration is given in Figure 2.

Proposition 2 (Existence of spurious local minima)**.** *Consider the Stochastic Ball Model in one dimension with $k^* = k = 3$ and $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*, \beta_3^*) \in \mathbb{R}^{1\times 3}$ with $\beta_1^* = -2$, $\beta_2^* = 0$, $\beta_3^* = 2$, where each ground truth ball/interval has radius r. When $r < 0.4$ or equivalently $\eta_{\min} > 5$, the solution $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3) \in \mathbb{R}^{1\times 3}$ with $\beta_1 = -2 - \frac{r}{2}$, $\beta_2 = -2 + \frac{r}{2}$ and $\beta_3 = 1$ is a local minimum of G.*

Proposition 2 is similar in spirit to the results in the work [12], which proves the existence of spurious local minima for the log-likelihood function of a Gaussian mixture. We note that their results are established using a limiting argument with $\eta_{\min} \to \infty$, whereas our result gives explicit values of $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}$.

Conceptually, Proposition 1 shows that when $k^* = k$, the k-means objective G is a *statistically* sensible objective function for clustering, as its global minimum recovers the ground truth clustering. On the other hand, Proposition 2 highlights the *computational* difficulty of this optimization task, due to the existence of spurious local minima in the form of the configuration plotted in Figure 2.

As the main result of this paper, we show that the above configuration is essentially the *only* local minimum. The following theorem, proved in Section VI, holds in the general setting where the number of centers k in the k-means objective G is not necessarily equal to the number of clusters k^* in the true mixture.

Theorem 1 (Local minimum structures, Stochastic Ball Model)**.** *Under the Stochastic Ball Model, for each constant integer $c_0 \geq 1$ there exists a universal constant $c > 0$ for which the following holds. Assume that $\max(k^*, k) \leq c_0$, $\eta_{\max} > 4c^2$ and $\eta_{\min} \geq 14c\sqrt{\eta_{\max}}$. If $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) \in \mathbb{R}^{d\times k}$ is a local minimum of G, then the ground truth centers and fitted centers can be partitioned as $[k^*] = \bigcup_{a=1}^m S_a^*$ and $[k] = \bigcup_{a=0}^m S_a$ respectively, such that for each $a \in [m]$, exactly one of the following holds:*

- ***(many/one-fit-one association)*** *$|S_a| \geq 1$ and $S_a^* = \{s\}$ for some $s \in [k]$; moreover,*
$$\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_s^*\| \leq \frac{10c}{\sqrt{\eta_{\max}}}\Delta_{\max} = 10c\sqrt{r\Delta_{\max}}, \quad \forall i \in S_a.$$
- ***(one-fit-many association)*** *$S_a = \{i\}$ for some $i \in [k]$ and $|S_a^*| \geq 2$; moreover,*
$$\Big\|\boldsymbol{\beta}_i - \frac{1}{|S_a^*|}\sum_{s\in S_a^*}\boldsymbol{\beta}_s^*\Big\| \leq \frac{14c}{\sqrt{\eta_{\max}}}\Delta_{\max} = 14c\sqrt{r\Delta_{\max}}.$$

In addition, for each $i \in S_0$, we have $\mathbb{P}\big(\mathcal{V}_i(\boldsymbol{\beta})\big) \leq \frac{c}{\sqrt{\eta_{\max}}}$ ***(almost-empty association)****.*

Theorem 1 states that *all local minima have the same type of structure.* In particular, if we view a candidate solution $\boldsymbol{\beta} = (\boldsymbol{\beta}_i)_{i=1}^k$ as configuring the centers $\boldsymbol{\beta}_i$'s to fit the true clusters, then any local minimum $\boldsymbol{\beta}$ must be composed of only the following configurations:

(i) many-fit-one: multiple $\boldsymbol{\beta}_i$'s are close to the same ground truth center;
(ii) one-fit-many: one $\boldsymbol{\beta}_i$ is close to the mean of several ground truth centers;
(iii) almost-empty: a $\boldsymbol{\beta}_i$ is far (relatively to other $\boldsymbol{\beta}_j$'s) from any ground truth center, in the sense that the Voronoi set of $\boldsymbol{\beta}_i$ is almost empty with a small measure.

In configurations (i) and (ii) above, being close means that the distance is order-wise smaller than the separation $\Delta_{\max}$, by a factor of $1/\sqrt{\eta_{\max}}$; similarly, in configuration (iii), the measure being small means that it is on the order of $1/\sqrt{\eta_{\max}}$. Moreover, the configurations (i), (ii) and (iii) must involve disjoint sets of $\boldsymbol{\beta}_j$s' and $\boldsymbol{\beta}_s^*$s'. For concrete examples, recall Figure 1: the ground truth solution in Panel 1a has 4 one-fit-one associations, whereas the spurious local minimum in Panel 1b consists of a two-fit-one, a one-fit-two and a one-fit-one association.

Put differently, Theorem 1 implies that if a solution $\boldsymbol{\beta}$ involves any configuration other than the three above, then $\boldsymbol{\beta}$ can be perturbed locally to strictly decreases its objective value. For example, the solutions in Panels 1c and 1d in Figure 1 use two centers to fit three and two true clusters, respectively. The objective value can be decreased by moving these two centers *away* from each other and towards different true clusters, as shown in Panels 2c and 2d. Our proof of Theorem 1 in fact makes use of this geometric idea, by testing the optimality conditions of the local minima along certain judiciously chosen directions.

It is instructive to specialize Theorem 1 to the limit case of a "point model", where $r \to 0$ or equivalently $\eta_{\max} \to \infty$; that is, each ground truth cluster s has a *point mass* at $\boldsymbol{\beta}_s^*$. In this case, the three possibilities guaranteed in the theorem reduce to: (i) several $\boldsymbol{\beta}_i$'s are located exactly at one true cluster $\boldsymbol{\beta}_s^*$ (many-fit-one); (ii) one center $\boldsymbol{\beta}_i$ is located at the mean of several true $\boldsymbol{\beta}_s^*$'s (one-fit-many); (iii) for all the other $\boldsymbol{\beta}_i$'s, their Voronoi sets do not contain any true clusters.

In the general setting with $r > 0$, Theorem 1 guarantees that the above result for the point model still holds approximately,

with an approximation error due to each true cluster having a mass spread around the true center. The three bounds in Theorem 1 control the approximation errors with respect to cases (i)–(iii) in the point model above. These error bounds all scale with $1/\sqrt{\eta_{\max}}$, which becomes smaller if the SNR $\eta_{\max}$ increases.

a) Tightness of the error bounds: The approximation errors above are unavoidable in general. We have already shown in Proposition 2 that there exists a local minimum $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are close but not exactly equal to $\boldsymbol{\beta}_1^*$; see Figure 2. Here the mass of the first true cluster $\mathbb{B}_1$ is equally split between the Voronoi sets of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, each of which lies at the corresponding center of mass (cf. Lemma 2), leading to a nonzero approximation error in the many-fit-one association. In addition, in Example 2 in Appendix 0c, we demonstrate another local minimum with a non-zero approximation error in the one-fit-many association.

Theorem 1 gives upper bounds for these approximation errors, where the bounds take the form $1/\sqrt{\eta_{\max}}$. The proof of the theorem in fact establishes a family of bounds that provide a trade-off between the errors for the three types of associations (see Theorem 4). In particular, when $k^* = k$, for each number $\lambda \in (0, \frac{1}{2k^2 r})$, one can derive the bounds

- many-fit-one – if $S_a^* = \{s\}$ and $|S_a| \geq 1$,
$$\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_s^*\| \leq O\left(\max\left(\frac{k}{\lambda}, kr + k^2\lambda r \Delta_{\max}\right)\right) \quad \forall i \in S_a;$$
- one-fit-many – if $S_a = \{i\}$ and $|S_a^*| > 1$,
$$\left\|\boldsymbol{\beta}_i - \frac{1}{|S_a^*|}\sum_{s\in S_a^*}\boldsymbol{\beta}_s^*\right\| \leq O\left(kr + k^2\lambda r\Delta_{\max} + \frac{k}{\lambda}\right);$$
- almost-empty – $\mathbb{P}\big(\mathcal{V}_i(\boldsymbol{\beta})\big) \leq \lambda k r, \ \forall i \in S_0$.

where the partitions $[k^*] = \bigcup_{a=1}^m S_a^*$ and $[k] = \bigcup_{a=0}^m S_a$ may depend on λ. Taking $\lambda = \frac{c}{\sqrt{r\Delta_{\max}}} = \frac{c}{r\sqrt{\eta_{\max}}}$ gives the bounds in Theorem 1. We are currently not sure whether these bounds are tight in general.

b) Separation condition: The result in Theorem 1 holds under the separation/SNR condition $\eta_{\min} = \Omega(1)$. Note that this kind of separation condition is needed even in recent work on more restrictive settings of mixture problems and EM/k-means, including those on $k^* = 2$ clusters [35, 39, 40, 31], on the local behavior of EM/k-means near the global optimum $\boldsymbol{\beta}^*$ [41, 42], and on the existence of spurious local minima [12] (the last work assumes $\eta_{\min} = \Omega(\sqrt{d})$). In our setting, we believe such a separation condition is in general necessary. Indeed, Example 1 in Appendix 0c shows that if $\eta_{\min}$ is too small, then there exists a local minimum that involves a two-fit-three configuration and hence qualitatively violates the structural properties in Theorem 1. Under our separation condition and additional assumptions, it may be possible to solve the mixture estimation problem using specialized algorithms with a good initialization (e.g., spectral initialization [43]). Nevertheless, as is the case in a recent line of work on the landscape of non-convex problems [44, 45, 46, 47], our results are not tied to a specific algorithm, but rather concern the general geometry and structure of the k-means problem itself. We have not attempted to optimize the separation condition (particularly its scaling with k^*) and leave it as an important future problem whether this condition can be improved. Indeed, we empirically observe that even with a much milder separation condition, such as the four clusters in Figure 1 , the one-fit-many and many-fit-one configurations still hold for local minimum solution.

c) Over/under-parametrization: Theorem 1 holds even when the number of centers k in the population k-means objective is different from the number of clusters k^* in the true mixture, including the over-parametrization regime $k > k^*$ and the under-parametrization regime $k < k^*$. As we discuss in greater details in Section IV, such flexibility has important algorithmic implications.

B. Gaussian Mixture Model

We next consider the Gaussian Mixture Model in Definition 2. The main difference between this model and the Stochastic Ball Model is that the Gaussian distribution has an unbounded support and thus the tails of the mixture components overlap with each other. Nevertheless, much of the results for the Ball Model can be extended to the Gaussian case. For example, one can establish results that are analogous to Propositions 1 and 2 regarding the global minima and the existence of spurious local minima. Here we focus on establishing an analogue of Theorem 1, which characterizes the structures of all local minima of the population k-means objective G.

Our main result is given in the following theorem, whose proof is given in Appendix A.

Theorem 2 (Local minimum structures, Gaussian Mixture Model)**.** *Under the Gaussian Mixture Model, for each constant integer $c_0 \geq 1$ there exists a universal constant $c > 0$ for which the following holds. Assume that $\widetilde{k} := \max(k, k^*) \leq c_0$ and let $t \geq 1$ be any number satisfying $\varphi(t) := 2\exp(-t^2 \min(d, \widetilde{k})/8) < \frac{1}{4}$ and $7\widetilde{k}\varphi(t) < 1$. Further assume that $\eta_{\max} \geq 16tc^2$ and $\eta_{\min} \geq 7c\sqrt{t\eta_{\max}} + 7\varphi(t)\eta_{\max}$. If $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) \in \mathbb{R}^{d\times k}$ is a local minimum of G, then the ground truth centers and fitted centers can be partitioned as $[k^*] = \bigcup_{a=1}^m S_a^*$ and $[k] = \bigcup_{a=0}^m S_a$, respectively, such that for each $a \in [m]$, exactly one of the following holds:*

- ***(many/one-fit-one association)*** *$|S_a| \geq 1$ and $S_a^* = \{s\}$ for some $s \in [k]$; moreover,*
$$\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_s^*\| \leq \Delta_{\max}\left\{\frac{5c\sqrt{t}}{\sqrt{\eta_{\max}}} + 7c\varphi(t)\right\}, \quad \forall i \in S_a. \tag{6}$$
- ***(one-fit-many association)*** *$S_a = \{i\}$ for some $i \in [k]$ and $|S_a^*| \geq 2$; moreover,*
$$\left\|\boldsymbol{\beta}_i - \frac{1}{|S_a^*|}\sum_{s\in S_a^*}\boldsymbol{\beta}_s^*\right\| \leq \Delta_{\max}\left\{\frac{7c\sqrt{t}}{\sqrt{\eta_{\max}}} + 7c\varphi(t)\right\}. \tag{7}$$

In addition, for each $i \in S_0$, we have $\mathbb{P}\big(\mathcal{V}_i(\boldsymbol{\beta})\big) \leq \frac{c\sqrt{t}}{\sqrt{\eta_{\max}}} + \varphi(t)$ ***(almost-empty association)****.*

Theorem 2 is qualitatively similar to Theorem 1 and shows that the local minima in GMM have a similar type of structure. The only differences are that the separation condition in Theorem 2 has an additional t factor, and that the bounds

for the three possibilities have an additional error term $\varphi(t)$ that decays exponentially in t^2. The $\varphi(t)$ term reflects the influence of the exponential tail of a Gaussian distribution outside a ball of radius $t\sigma\sqrt{\min(d, k+k^*)}$. In fact, the proof of Theorem 2 proceeds by effectively reducing GMM to the Stochastic Ball Model, treating the bulk of the Gaussian as a bounded distribution and the tail as additional errors. The choice of t here determines what is viewed as the tail and hence controls the trade-off between the separation condition and the two terms in the error bounds. For a rough interpretation of the theorem, one could simply think of t as a numerical constant large enough so that $\varphi(t)$ is dominated by the other terms in the error bounds.

IV. Implications and Connections

The theorems in the last section provide *structural* results for the k-means objective. In this section, we discuss some *algorithmic* implications of these results and remark on their connections to the literature.

a) Algorithmic Implications: When $k^* = k$, our result implies that one can find the global minimum of k-means as long as the characteristic many-fit-one association for local minima can be avoided — in this case, one-fit-many association will also disappear as the number of true clusters and that of fitted centers are equal. This observation suggests that to avoid many-fit-one, one should initialize a greedy clustering algorithm without putting the fitted centers close to each other. Several popular heuristics for k-means implement precisely this idea. For instance, the celebrated k-means++ algorithm [24] is a version of the Lloyd's algorithm in which the initial centers are generated iteratively as follows: the first center is selected uniformly from the data points; after selecting j centers, one computes the minimal distance of each data point to these j centers, and select a data point randomly as the $(j+1)$-th center with probability proportional to the above distance. By design, this procedure tends to pick k initial centers that are far away from each other. Many other heuristics for k-means follow a similar spirit; see, e.g., the work in [48, 49, 50, 51].

On the other hand, our structural results also highlight the inherent combinatorial difficulty of the problem. In particular, when the number of clusters grows, there is a growing number of possible configurations with many-fit-one and one-fit-many associations. It then becomes easier to get trapped in one of the corresponding local minima. This is consistent with the systematic empirical study in [52], which observes that algorithms for k-means perform worse when there are more clusters.

b) Connection to Over-Parametrization: Our structural results in Theorem 1 and 2 hold in the over-parametrization setting where $k > k^*$ centers are used to fit k^* ground truth clusters. Over-parametrization appears to be a promising approach for avoiding local minima. In particular, when k is sufficiently bigger than k^*, a random initial solution is likely to assign *at least* one center to each true cluster. In this case, one-fit-many association would be avoided. Running a greedy algorithm from this initial solution, one would expect that it converges to a solution with only many-fit-one and almost-empty associations, which can then be pruned by inspecting the pairwise distances of the fitted centers and the sizes of their Voronoi sets. The work in [26] implements this idea in the context of over-parametrized EM. In particular, after EM converges, they remove fitted centers with low mixing weights (corresponding to almost-empty association) and combine fitted centers that are close to each other (corresponding to many-fit-one association).

In fact, the extensive empirical study in [53] shows that the above idea can be applied to other latent variable models, as these models often have a similar solution structure, i.e., some estimated latent variables having duplicated values or low prior probabilities.

V. Preliminary Properties for the k-means Objective

In this section, we derive several preliminary results on the analytical properties of the population k-means objective function G defined in equation (3), focusing on the Stochastic Ball Model. These properties are later used in the proofs of our main theorems.

When $\boldsymbol{\beta}$ has pairwise distinct components (i.e., $\boldsymbol{\beta}_i \neq \boldsymbol{\beta}_j, \forall i \neq j \in [k]$), it is often convenient to rewrite the function G using the notation of Voronoi sets:

$$G(\boldsymbol{\beta}) = \sum_{i=1}^{k} \int_{\mathcal{V}_i(\boldsymbol{\beta})} \|\boldsymbol{x} - \boldsymbol{\beta}_i\|^2 f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{8}$$

We can see that G depends on $\boldsymbol{\beta}$ in a complicated way through both $\|\boldsymbol{x} - \boldsymbol{\beta}_i\|^2$ and $\mathcal{V}_i(\boldsymbol{\beta})$. As shall become clear later, the dependence through the squared distance $\|\boldsymbol{x} - \boldsymbol{\beta}_i\|^2$ determines the first-order condition for local optimality for G; on the other hand, understanding second-order conditions requires us to study the behaviors of the Voronoi sets $\mathcal{V}_i(\boldsymbol{\beta})$ under small perturbation of $\boldsymbol{\beta}$. To deal with this complication, our main strategy is to understand the directional behaviors of G along certain (judiciously chosen) directions, and to construct upper bounds on G that are easier to work with.

A. Directional Behaviors of G

Throughout the remainder of this section, we fix a candidate solution $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) \in \mathbb{R}^{d \times k}$. For a given direction $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2 \ldots, \boldsymbol{v}_k) \in \mathbb{R}^{d \times k}$, we are interested in how the objective $G(\boldsymbol{\beta})$ changes after we perturb $\boldsymbol{\beta}$ to $\boldsymbol{\beta} + t\boldsymbol{v}$. Restricting the function G to the direction $\boldsymbol{v}$, we define the directional objective function

$$H^{\boldsymbol{v}}(t) := G(\boldsymbol{\beta} + t\boldsymbol{v}), \quad t \in \mathbb{R}.$$

Note that $\boldsymbol{\beta}$ is a local minimum of G if and only if $\mathbf{0}$ is local minimum of $H^{\boldsymbol{v}}$ for all $\boldsymbol{v}$.

The functions G and $H^{\boldsymbol{v}}$ are not everywhere differentiable, as they involve the minimum of quadratic functions. However, they are differentiable almost everywhere. In particular, whenever $\boldsymbol{\beta}$ has pairwise distinct components, the directional derivative $\frac{\mathrm{d}}{\mathrm{d}t} H^{\boldsymbol{v}}(0)$ is guaranteed to exist and admits a simple expression, as shown in the following lemma.

Lemma 1 (Directional derivative)**.** *Suppose that $\boldsymbol{\beta}$ satisfies $\boldsymbol{\beta}_i \neq \boldsymbol{\beta}_j$ whenever $i \neq j$. For any choice of direction $\boldsymbol{v}$, the*

directional derivative $\frac{\mathrm{d}}{\mathrm{d}t}H^{\boldsymbol{v}}(0)$ *exists and has the following analytic formula:*

$$\frac{\mathrm{d}}{\mathrm{d}t}H^{\boldsymbol{v}}(0) = -\sum_{i=1}^{k}\int_{\mathcal{V}_i(\boldsymbol{\beta})} 2\langle \boldsymbol{v}_i, \boldsymbol{x}-\boldsymbol{\beta}_i\rangle f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

The lemma follows from the Leibniz integral rule; we defer the proof to Appendix A. Note that the above expression only involves differentiating the integrand in the expression (8); the Voronoi sets $\mathcal{V}_i(\boldsymbol{\beta})$ remain unchanged when only the first-order derivative is concerned.

B. First-Order Necessary Condition for Local Optimality

Using the first-order derivative expression in Lemma 1, we can derive a necessary condition for $\boldsymbol{\beta}$ being a local minimum. In particular, the following lemma states that any local minimum (satisfying a non-degeneracy condition) must have pairwise distinct components, each of which must be the center of its Voronoi set.

Lemma 2 (Local minimum must be Voronoi centers)**.** *Suppose that $\boldsymbol{\beta}$ is a local minimum of G. Then for each pair $i \neq j$, we must have $\boldsymbol{\beta}_i \neq \boldsymbol{\beta}_j$ whenever $\mathcal{V}_i(\boldsymbol{\beta}) \cup \mathcal{V}_j(\boldsymbol{\beta})$ having a positive measure (with respect to f). Moreover, for each $\boldsymbol{\beta}_i$ whose Voronoi set $\mathcal{V}_i(\boldsymbol{\beta})$ has a positive measure, $\boldsymbol{\beta}_i$ must be at the center of probability mass of the Voronoi set $\mathcal{V}_i(\boldsymbol{\beta})$; that is,*

$$\boldsymbol{\beta}_i = \frac{\int_{\mathcal{V}_i(\boldsymbol{\beta})} \boldsymbol{x} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\int_{\mathcal{V}_i(\boldsymbol{\beta})} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}. \tag{9}$$

We prove Lemma 2 in Appendix A. The conclusion of the above lemma can be written equivalently in a more explicit way. In particular, for each pair $(i,s) \in [k]\times[k^*]$, let $m_{i,s}(\boldsymbol{\beta})$ and $\boldsymbol{c}_{i,s}(\boldsymbol{\beta})$ denote the probability mass and the center of mass of the set $\mathcal{V}_i(\boldsymbol{\beta})$ with respect to f_s respectively:

$$m_{i,s}(\boldsymbol{\beta}) := \int_{\mathcal{V}_i(\boldsymbol{\beta})} f_s(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$
$$\boldsymbol{c}_{i,s}(\boldsymbol{\beta}) := \frac{\int_{\mathcal{V}_i(\boldsymbol{\beta})} \boldsymbol{x} f_s(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{m_{i,s}(\boldsymbol{\beta})}.$$

Then equation (9) can be rewritten as

$$\boldsymbol{\beta}_i = \frac{\sum_{s=1}^{k^*} m_{i,s}(\boldsymbol{\beta})\boldsymbol{c}_{i,s}(\boldsymbol{\beta})}{\sum_{s=1}^{k^*} m_{i,s}(\boldsymbol{\beta})}.$$

C. Decomposition of $H^{\boldsymbol{v}}$

Lemmas 1 and 2 provide a first-order characterization of the local minima of G. For a more precise characterization, we need to account for the change in the Voronoi sets $\mathcal{V}(\boldsymbol{\beta})$ and its boundaries $\partial(\boldsymbol{\beta})$ when perturbing $\boldsymbol{\beta}$ to $\boldsymbol{\beta}+t\boldsymbol{v}$. With $t>0$ considered arbitrarily small, we make two observations:

1) The Voronoi set boundaries $\partial(\boldsymbol{\beta}+t\boldsymbol{v})$ change continuously with respect to t.
2) When $\boldsymbol{\beta}$ is perturbed by $t\boldsymbol{v}$, the points swept by the boundaries $\partial(\boldsymbol{\beta}+t\boldsymbol{v})$ change their association from one Voronoi set to another.

Formally, for each pair $(i,j)\in[k]\times[k]$ we define the set

$$\Delta^{\boldsymbol{v}}_{i\to j}(t) := \mathcal{V}_i(\boldsymbol{\beta}) \cap \mathcal{V}_j(\boldsymbol{\beta}+t\boldsymbol{v}),$$

which is the set of points that change association from the i-th fitted center to the j-th fitted center due to the perturbation $t\boldsymbol{v}$. Being the intersection of two polyhedra, the set $\Delta^{\boldsymbol{v}}_{i\to j}(t)$ is a also polyhedron. An illustration of $\Delta^{\boldsymbol{v}}_{i\to j}(t)$ is provided in Figure 3.

As previously shown in Lemma 2, any non-degenerate local minimum $\boldsymbol{\beta}$ must have distinct components, so the corresponding Voronoi sets are also distinct. The same holds for the perturbed solution $\boldsymbol{\beta}+t\boldsymbol{v}$ when t is sufficiently small. In this case, we can decompose the directional objective function $H^{\boldsymbol{v}}$ as follows:

$$\begin{aligned}
&H^{\boldsymbol{v}}(t)\\
=&\sum_{i=1}^{k}\int_{\mathcal{V}_i(\boldsymbol{\beta}+t\boldsymbol{v})} \|\boldsymbol{x}-\boldsymbol{\beta}_i - t\boldsymbol{v}_i\|^2 f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
=&\underbrace{\sum_{i=1}^{k}\int_{\mathcal{V}_i(\boldsymbol{\beta})} \|\boldsymbol{x}-\boldsymbol{\beta}_i - t\boldsymbol{v}_i\|^2 f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}_{U^{\boldsymbol{v}}(t)}\\
&+\underbrace{\sum_{(i,j):i\neq j}\int_{\Delta^{\boldsymbol{v}}_{i\to j}(t)} \left(\|\boldsymbol{x}-\boldsymbol{\beta}_j - t\boldsymbol{v}_j\|^2 - \|\boldsymbol{x}-\boldsymbol{\beta}_i - t\boldsymbol{v}_i\|^2\right) f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}_{W^{\boldsymbol{v}}(t)}.
\end{aligned} \tag{10}$$

Here $U^{\boldsymbol{v}}(t)$ and $W^{\boldsymbol{v}}(t)$ correspond to the change in the objective value from two different sources. In particular, $U^{\boldsymbol{v}}(t)$ is due to the change in the distance between the data points and the centers, and $W^{\boldsymbol{v}}(t)$ is due to the data points changing association with the Voronoi sets.

Remark 1. *By definition of $\Delta^{\boldsymbol{v}}_{i\to j}(t)$, for each $\boldsymbol{x}\in\Delta^{\boldsymbol{v}}_{i\to j}(t)$, the integrand $\|\boldsymbol{x}-\boldsymbol{\beta}_j-t\boldsymbol{v}_j\|^2-\|\boldsymbol{x}-\boldsymbol{\beta}_i-t\boldsymbol{v}_i\|^2$ in the definition of $W^{\boldsymbol{v}}(t)$ is non-positive.*

Proof of equation (10). When the Voronoi sets $\mathcal{V}(\boldsymbol{\beta})$ are perturbed to $\mathcal{V}(\boldsymbol{\beta}+t\boldsymbol{v})$, each point $\boldsymbol{x}$ in $\mathbb{R}^d$ either remains associated with the i-th center for some i, or changes its association from the i-th center to the j-th center for some $j\neq i$. In the first case, since $\boldsymbol{x}\in\mathcal{V}_i(\boldsymbol{\beta})\cap\mathcal{V}_i(\boldsymbol{\beta}+t\boldsymbol{v})$, we see that the contribution from $\boldsymbol{x}$ to $H^{\boldsymbol{v}}(t)$ appears in $U^{\boldsymbol{v}}$. In the second case, since $\boldsymbol{x}\in\Delta^{\boldsymbol{v}}_{i\to j}(t)=\mathcal{V}_i(\boldsymbol{\beta})\cap\mathcal{V}_j(\boldsymbol{\beta}+t\boldsymbol{v})$, we can write the contribution from $\boldsymbol{x}$ as

$$\begin{aligned}
&\|\boldsymbol{x}-\boldsymbol{\beta}_j-t\boldsymbol{v}_j\|^2\\
=&\|\boldsymbol{x}-\boldsymbol{\beta}_i-t\boldsymbol{v}_i\|^2+\left(\|\boldsymbol{x}-\boldsymbol{\beta}_j-t\boldsymbol{v}_j\|^2-\|\boldsymbol{x}-\boldsymbol{\beta}_i-t\boldsymbol{v}_i\|^2\right),
\end{aligned}$$

which appears in both $U^{\boldsymbol{v}}$ and $W^{\boldsymbol{v}}$. □

D. Smooth Upper Bounds of $H^{\boldsymbol{v}}$

The expression (10) for $H^{\boldsymbol{v}}$ is quite complicated. To understand the local minima of $H^{\boldsymbol{v}}$, we instead study a simpler, better-behaved upper bound function of $H^{\boldsymbol{v}}$ that preserves the local minima and is amenable to calculus tools. In particular, we make use of the following elementary lemma.

Lemma 3 (Smooth upper bound)**.** *Suppose that $h,\widetilde{h}:\mathbb{R}\to\mathbb{R}$ are two continuous functions that satisfy $h\leq\widetilde{h}$ and $h(0)=\widetilde{h}(0)$. If 0 is a local minimum of h, then 0 is also a local*

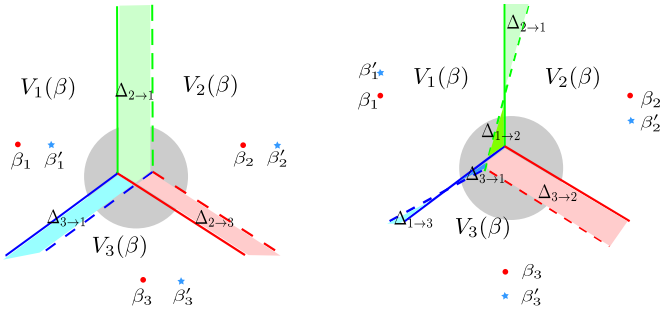

Figure 3. Illustration of the set $\Delta^{\boldsymbol{v}}_{i\to j}(t)$. The red dots represent the original centers $\{\boldsymbol{\beta}_i\}$ and the blue stars represent the perturbed centers $\{\boldsymbol{\beta}'_i = \boldsymbol{\beta}_i + t\boldsymbol{v}_i\}$. The blue solid lines are the original Voronoi boundaries $\partial(\boldsymbol{\beta})$, and the red dashed lines are the perturbed Voronoi boundaries $\partial(\boldsymbol{\beta}')$. Top panel: the moving direction $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3)$ satisfies $\boldsymbol{v}_1 = \boldsymbol{v}_2 = \boldsymbol{v}_3$, in which case the Voronoi boundaries are shifted parallelly by $t\boldsymbol{v}_1$. Bottom panel: the moving direction satisfies $\boldsymbol{v}_1 = -\boldsymbol{v}_2 = -\boldsymbol{v}_3$, in which case the boundary $\partial_{1,2}(\boldsymbol{\beta}')$ rotates around the mid point $\frac{\boldsymbol{\beta}_1+\boldsymbol{\beta}_2}{2}$, $\partial_{1,3}(\boldsymbol{\beta}')$ rotates around the mid point $\frac{\boldsymbol{\beta}_1+\boldsymbol{\beta}_3}{2}$, and $\partial_{2,3}(\boldsymbol{\beta}')$ shifts parallelly in the direction of $\boldsymbol{v}_2$. Each colored region represents $\Delta^{\boldsymbol{v}}_{i\to j}(t)$, the set of points that change the association from the i-th center to the j-th center.

minimum of $\widetilde{h}$; moreover, we have $\lim_{t\to0}\frac{\widetilde{h}(t)-\widetilde{h}(0)}{t} = 0$ and $\lim_{t\to0}\frac{\widetilde{h}(t)-\widetilde{h}(0)}{t^2} \geq 0$ whenever the limits exist.

Proof. Since 0 is a local minimum of h, we have $\widetilde{h}(0) = h(0) \leq h(t) \leq \widetilde{h}(t)$ for all t in a neighborhood of 0, so 0 is also a local minimum of $\widetilde{h}$. The first-order optimality condition for 0 gives $\lim_{t\to0}\frac{\widetilde{h}(t)-\widetilde{h}(0)}{t} = \widetilde{h}'(0) = 0$. Moreover, we have $\widetilde{h}(t) - \widetilde{h}(0) \geq 0 \implies \frac{\widetilde{h}(t)-\widetilde{h}(0)}{t^2} \geq 0$ for all $t \neq 0$ in a neighborhood of 0, which implies that $\lim_{t\to0}\frac{\widetilde{h}(t)-\widetilde{h}(0)}{t^2} \geq 0$. □

With the above lemma, we can study the structure of each local minimum of $H^{\boldsymbol{v}}$ (and hence that of G) by exploiting the optimality conditions of a smooth upper bound of $H^{\boldsymbol{v}}$ that is tight at the minimum. Let us take a first step in constructing such an upper bound. In view of the non-positivity result in Remark 1, we can obtain an upper bound of the function $W^{\boldsymbol{v}}$ defined in (10) by only considering those pairs (i,j) for which $\mathcal{V}_i(\boldsymbol{\beta}) \sim \mathcal{V}_j(\boldsymbol{\beta})$ are adjacent:

$$\begin{aligned} &W^{\boldsymbol{v}}(t) \\ &\leq \sum_{\substack{(i,j):\\ \mathcal{V}_i(\boldsymbol{\beta})\sim\mathcal{V}_j(\boldsymbol{\beta})}} \int_{\Delta^{\boldsymbol{v}}_{i\to j}(t)} \Big(\|\boldsymbol{x}-\boldsymbol{\beta}_j-t\boldsymbol{v}_j\|^2 - \|\boldsymbol{x}-\boldsymbol{\beta}_i-t\boldsymbol{v}_i\|^2\Big) f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \sum_{\substack{(i,j):\\ \mathcal{V}_i(\boldsymbol{\beta})\sim\mathcal{V}_j(\boldsymbol{\beta})}} \frac{1}{k^*}\sum_{s=1}^{k^*} W^{\boldsymbol{v}}_{i\to j,s}(t) \end{aligned}$$

where the quantity $W^{\boldsymbol{v}}_{i\to j,s}$, a shorthand for

$$\int_{\Delta^{\boldsymbol{v}}_{i\to j}(t)} (\|\boldsymbol{x}-\boldsymbol{\beta}_j-t\boldsymbol{v}_j\|^2 - \|\boldsymbol{x}-\boldsymbol{\beta}_i-t\boldsymbol{v}_i\|^2) f_s(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

represents the contribution from the points in the s-th true cluster that change association from the i-th center to the j-th center. Combining the above inequality with equation (10), we obtain the following upper bound

$$H^{\boldsymbol{v}}(t) \leq U^{\boldsymbol{v}}(t) + \sum_{\substack{(i,j):\\ \mathcal{V}_i(\boldsymbol{\beta})\sim\mathcal{V}_j(\boldsymbol{\beta})}} \frac{1}{k^*}\sum_{s=1}^{k^*} W^{\boldsymbol{v}}_{i\to j,s}(t). \tag{11}$$

In the proofs of our main theorems, we build upon equation (11) to derive further smooth upper bounds of $H^{\boldsymbol{v}}$.

VI. Proof of Theorem 1

In this section, we prove our main result under the Stochastic Ball Model (Definition 1). We in fact establish a more general and quantitative version of Theorem 1 that holds for any k^* and k.

Theorem 3 (General version of Theorem 1)**.** *Under the Stochastic Ball Model, assume that $\eta_{\max} > 4c^2k^4$ and $\eta_{\min} \geq \sqrt{\eta_{\max}}\left(2ck^2\left(1+\frac{2k^*}{k}\right) + 4k^*\left(3+\frac{2k^*}{k}\right)\right)$ for some universal constant $c \geq 3$. If $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) \in \mathbb{R}^{d\times k}$ is a local minimum of G, then the ground truth centers and fitted centers can be partitioned as $[k^*] = \bigcup_{a=1}^m S^*_a$ and $[k] = \bigcup_{a=0}^m S_a$ respectively, such that for each $a \in [m]$, exactly one of the following holds:*

- ***(many/one-fit-one association)*** *$|S_a| \geq 1$ and $S^*_a = \{s\}$ for some $s \in [k^*]$; moreover,*

$$\|\boldsymbol{\beta}_i - \boldsymbol{\beta}^*_s\| \leq \Delta_{\max}\frac{(4k^* + 2ck^2)\left(1+\frac{2k^*}{k}\right)}{\sqrt{\eta_{\max}}}, \quad \forall i \in S_a.$$

- ***(one-fit-many association)*** *$S_a = \{i\}$ for some $i \in [k]$ and $|S^*_a| \geq 2$; moreover,*

$$\left\|\boldsymbol{\beta}_i - \frac{1}{|S^*_a|}\sum_{s\in S^*_a}\boldsymbol{\beta}^*_s\right\| \leq \Delta_{\max}\frac{2ck^2\left(1+\frac{2k^*}{k}\right) + 4k^*\left(3+\frac{2k^*}{k}\right)}{\sqrt{\eta_{\max}}}.$$

In addition, for each $i \in S_0$, we have $\mathbb{P}\big(\mathcal{V}_i(\boldsymbol{\beta})\big) \leq \frac{ck}{\sqrt{\eta_{\max}}}$ ***(almost-empty association)****.*

Given Theorem 3, Theorem 1 follows immediately. The rest of the section is devoted to proving Theorem 3.

Throughout the proof, let $\boldsymbol{\beta}$ be a fixed local minimum of the k-means objective G. Note that G is invariant under translation of the space and permutation of the true centers. Consequently, we may assume without loss of generality that $\boldsymbol{\beta}^*_1 = \mathbf{0}$ and $\max_{s\in[k]}\|\boldsymbol{\beta}^*_s\| \leq \Delta_{\max}$.

A. Notations

We use V_d to denote the volume of a unit ball in $\mathbb{R}^d$ with respect to the Lebesgue measure. For a set $T \subset \mathbb{R}^d$, let $\mathrm{int}(T)$ denote its interior, and $\mathrm{ReVol}(T)$ denote the relative volume of T with respect to the Lebesgue measure on the affine hull of T, with the convention that $\mathrm{ReVol}(\emptyset) = 0$. For two vector $\boldsymbol{u}, \boldsymbol{u}' \in \mathbb{R}^d$, $\angle(\boldsymbol{u}, \boldsymbol{u}') := \arccos(\frac{\boldsymbol{u}^\top \boldsymbol{u}'}{\|\boldsymbol{u}\|\|\boldsymbol{u}'\|}) \in [0, \pi]$ is the angle between $\boldsymbol{u}$ and $\boldsymbol{u}'$. For each tuple $(i, j, s) \in [k] \times [k] \times [k^*]$ with $i \neq j$, we use $\mathcal{L}_{i,j,s}(\boldsymbol{\beta})$ to denote the two-dimensional plane that contains $\boldsymbol{\beta}_i$, $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}^*_s$ (if such a plane is not unique, we fix an arbitrary one). Since we are concerned with a fixed local minimum $\boldsymbol{\beta}$, we sometimes suppress the dependency on $\boldsymbol{\beta}$ and write, for example, $\mathcal{V}_i \equiv \mathcal{V}_i(\boldsymbol{\beta})$, $\partial_{i,j} \equiv \partial_{i,j}(\boldsymbol{\beta})$ and $\mathcal{L}_{i,j,s} \equiv \mathcal{L}_{i,j,s}(\boldsymbol{\beta})$.

B. Proof of Theorem 3

To prove Theorem 3, we establish an intermediate result as given in Theorem 4, which provides a family of bounds parametrized by $\lambda > 0$.

Theorem 4 (Family of bounds for ball model)**.** *Under the Stochastic Ball Model, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ be a local minimum of the k-means objective function G defined in (3) and $\lambda > 0$ be an arbitrary fixed number. For each $i, j \in [k]$ and $s \in [k^*]$, let $\rho_s(\partial_{i,j}) := \frac{1}{V_d r^d}\mathrm{ReVol}(\partial_{i,j} \cap \mathbb{B}_s)$. For each $i \in [k]$, define the sets*

$$T_i := \big\{s \in [k^*] : \mathcal{V}_i \cap \mathbb{B}_s \neq \emptyset\big\} \quad \text{and}$$
$$A_i := \big\{s \in [k^*] : \boldsymbol{\beta}^*_s \in \mathrm{int}(\mathcal{V}_i)\big\} \subseteq T_i.$$

Then the following is true for each $i \in [k]$.

1) *If $\rho_s(\partial_{j,\ell}) > \lambda$ for some $s \in T_i$ and some pair (j, ℓ), then*

$$\|\boldsymbol{\beta}_i - \boldsymbol{\beta}^*_s\| \leq \frac{k^*}{\lambda} + 3r.$$

2) *For each $s \in T_i$, if $\rho_s(\partial_{j,\ell}) \leq \lambda$ for all pair (j, ℓ), then the following bounds hold:*

$$\mathbb{P}_s\big(\mathcal{V}_i\big) \geq 1 - k^2\lambda r, \quad \text{if } s \in A_i,$$
$$\mathbb{P}_s\big(\mathcal{V}_i\big) \leq k\lambda r, \quad \text{if } s \in T_i \setminus A_i.$$

Furthermore, if $\rho_s(\partial_{j,\ell}) \leq \lambda$ for all $s \in T_i$ and all pair (j, ℓ), then:

a) When $|A_i| = 0$, we have

$$\mathbb{P}(\mathcal{V}_i) \leq k\lambda r.$$

b) When $|A_i| > 0$, we have

$$\|\boldsymbol{\beta}_i - \boldsymbol{b}_i\| \leq \frac{k^* r}{1 - k^2\lambda r} + \frac{k^*(k^2\lambda r^2)(1 + (k^* k - k^2)\lambda r)}{(1 - k^2\lambda r)^2} + \frac{(k + 2k^*)k\lambda r}{1 - k^2\lambda r}\Delta_{\max},$$

where $\boldsymbol{b}_i := \frac{1}{|A_i|}\sum_{s \in A_i} \boldsymbol{\beta}^*_s$.

The proof of Theorem 4, which lies at the core of our analysis, is given in Section VI-C.

We now derive Theorem 3 from Theorem 4. Doing so involves several elementary though somewhat tedious steps of a combinatorial flavor. To this end, we fix $\lambda = \frac{c}{\sqrt{r\Delta_{\max}}} = \frac{c}{r\sqrt{\eta_{\max}}}$. Recall the assumption in the main theorem that $\eta_{\max} > 4c^2k^4$ for $c > 3$. This assumption implies that $k^2\lambda r < 0.5$. If $\rho_s(\partial_{i,j}) > \lambda$, we say that a true cluster $\mathbb{B}_s$ encloses the Voronoi boundary $\partial_{i,j}$ with a large relative volume; otherwise, we say that $\mathbb{B}_s$ encloses the Voronoi boundary $\partial_{i,j}$ with a small relative volume.

We first state two simple implications of Theorem 4 used frequently in the subsequent proof.

Observation 1. For each $i \in [k]$, there exists at most one $s \in T_i$ such that $\rho_s(\partial_{j,\ell}) > \lambda$ for some pair (j, ℓ). In words, each Voronoi set $\mathcal{V}_i$ can intersect at most one true cluster $\mathbb{B}_s$ that encloses some Voronoi boundary with a large relative volume.

Observation 2. For each $s \in [k]$, if $\rho_s(\partial_{j,\ell}) \leq \lambda$ for all pair (j, ℓ), then $\boldsymbol{\beta}^*_s \in \mathcal{V}_i$ implies that $s \in A_i$. In words, if all Voronoi boundaries enclosed by a true cluster $\mathbb{B}_s$ have small relative volumes, then the center $\boldsymbol{\beta}^*_s$ cannot itself lie on a Voronoi boundary.

Proof of observations. We prove these observations by contradiction. For Observation 1, suppose otherwise that there exists $s \neq s' \in T_i$ for which the statement holds. Part 1 of Theorem 4 ensures that $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}^*_s\| \leq \frac{k^*}{\lambda} + 3r$ and $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{s'^*}\| \leq \frac{k^*}{\lambda} + 3r$. Using the triangle inequality and the value for λ, we obtain that $\|\boldsymbol{\beta}^*_s - \boldsymbol{\beta}^*_{s'}\| \leq \frac{2k^*}{\lambda} + 6r \leq \Delta_{\max}\frac{8k^*}{c\sqrt{\eta_{\max}}}$, which contradicts the assumption on $\eta_{\min}$.

For Observation 2, suppose otherwise that $\boldsymbol{\beta}^*_s \in \mathcal{V}_i$ and $s \notin A_i$ for some $i \in [k]$, which implies that $\boldsymbol{\beta}^*_s$ lies on a Voronoi boundary and hence $\boldsymbol{\beta}^*_s \notin A_j, \forall j \in [k]$. If $s \in T_j$, then Part 2 of Theorem 4 ensures that $\mathbb{P}_s(\mathcal{V}_j) \leq k\lambda r, \forall j \in [k]$; if $s \notin T_j$, then $\mathbb{P}_s(\mathcal{V}_j) = 0$ by definition of T_j. Summing over $j \in [k]$, we obtain that $1 = \mathbb{P}_s(\mathbb{B}_s) = \sum_{j \in [k]} \mathbb{P}_s(\mathcal{V}_j) \leq k^2\lambda r < 0.5$, which is a contradiction. □

We now construct a partition $\bigcup_{a=0}^m S_a = [k]$ of the fitted centers and a partition $\bigcup_{a=1}^m S^*_a = [k^*]$ of the true centers that satisfy the conclusion of Theorem 3. These partitions induce an association between the fitted centers in S_a and the true centers in S^*_a, for each $a = 1, \ldots, m$. The construction proceeds in three steps.

a) Step 1 (almost-empty association): First consider the fitted centers indexed by the set

$$S_0 := \big\{i \in [k] : \rho_s(\partial_{j,\ell}) \leq \lambda, \; \forall (s, j, \ell) \in T_i \times [k] \times [k]; |A_i| = 0\big\}.$$

Part 2(a) of Theorem 4 ensures that for all $i \in S_0$, we have $\mathbb{P}(\mathcal{V}_i) \leq k\lambda r = \frac{ck}{\sqrt{\eta_{\max}}}$ as claimed in Theorem 3.

b) Step 2 (many/one-fit-one association):: We next consider the fitted centers indexed by the set

$$\mathcal{J} := \{i \in [k] : |A_i| \leq 1\} \setminus S_0.$$

For each $i \in \mathcal{J}$, there are two complementary cases:

- $\rho_s(\partial_{j,\ell}) \le \lambda$ for all $(s,j,\ell) \in T_i \times [k] \times [k]$; that is, all true clusters that intersect $\mathcal{V}_i$ only enclose Voronoi boundaries with a small relative volume. Since $i \notin S_0$, by definition of S_0 and $\mathcal{J}$ we must have $|A_i| = 1$; say $A_i = \{s\}$. Applying Part 2(b) of Theorem 4, we have

$$\begin{aligned}
&\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_s^*\| = \|\boldsymbol{\beta}_i - \boldsymbol{b}_i\| \\
\le& \frac{k^* r}{1-k^2\lambda r} + \frac{k^*(k^2\lambda r^2)(1+(k^*k-k^2)\lambda r)}{(1-k^2\lambda r)^2} \\
&+ \frac{(k+2k^*)k\lambda r}{1-k^2\lambda r}\Delta_{\max} \\
\overset{\text{(i)}}{\le}& 2k^* r + 2k^* r\left[1 + (k^*k-k^2)\frac{c}{\sqrt{\eta_{\max}}}\right] \\
&+ 2k(k+2k^*)\frac{c}{\sqrt{\eta_{\max}}}\Delta_{\max} \\
\overset{\text{(ii)}}{\le}& \Delta_{\max}\frac{(4k^*+2ck^2)(1+\frac{2k^*}{k})}{\sqrt{\eta_{\max}}},
\end{aligned} \tag{12}$$

  where in step (i) we plug in $\lambda = \frac{c}{\sqrt{\Delta_{\max} r}}$ and use the assumption that $k^2\lambda r < \frac{1}{2}$; in step (ii), we use the assumption that $\eta_{\max} \ge 4c^2k^4$ with $c > 3$ to further simplify the bound.
- $\rho_s(\partial_{j,\ell}) > \lambda$ for some $(s,j,\ell) \in T_i \times [k] \times [k]$; that is, there exists some ground truth cluster $\mathbb{B}_s$ that encloses a Voronoi boundary with a large relative volume. Applying Part 1 of Theorem 4 and plugging the value of λ, we obtain that $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_s^*\| \le \frac{k^*}{\lambda} + 3r \le \Delta_{\max}\frac{4k^*}{c\sqrt{\eta_{\max}}}$.

In both cases, we have $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_s^*\| \le \Delta_{\max}\frac{(4k^*+2ck^2)(1+\frac{2k^*}{k})}{\sqrt{\eta_{\max}}}$ as claimed in Theorem 3. For each distinct index $s \in [k^*]$ that appears in the above arguments, set $S_a^* = \{s\}$ and let the corresponding S_a index those $\boldsymbol{\beta}_i$'s for which either of the two cases holds. The sets $\{S_a\}$ constructed in this way are disjoint. Indeed, for each i the above two cases are exclusive, where in the first case A_i contains s and only s, and in the second case above the index s is unique by Observation 1.

c) Step 3 (one-fit-many association):: We are left with the fitted centers indexed by the set

$$\mathcal{K} := \{i \in [k] : |A_i| \ge 2\} = [k] \setminus (S_0 \cup \mathcal{J}).$$

Similarly to before, for each $i \in \mathcal{K}$, there are two complementary cases:

- $\rho_s(\partial_{j,\ell}) \le \lambda$ for all $(s,j,\ell) \in T_i \times [k] \times [k]$. Applying Part 2(b) of Theorem 4 and following the same steps as in equation (12), we obtain that $\|\boldsymbol{\beta}_i - \boldsymbol{b}_i\| \le \Delta_{\max}\frac{(4k^*+2ck^2)(1+\frac{2k^*}{k})}{\sqrt{\eta_{\max}}}$. In this case, we let $S_a = \{i\}$ and $S_a^* = A_i$. Note that $|S_a^*| = |A_i| \ge 2$ by definition of $\mathcal{K}$.
- $\rho_s(\partial_{j,\ell}) > \lambda$ for some $(s,j,\ell) \in T_i \times [k] \times [k]$. In this case, applying Part 1 of Theorem 4 would show that $\boldsymbol{\beta}_i$ is close to $\boldsymbol{\beta}_s^*$. In fact, we can establish a stronger result showing that $\boldsymbol{\beta}_i$ is close to the mean of all the true centers contained in its Voronoi set, regardless of whether we include or exclude $\boldsymbol{\beta}_s^*$. This is the content of the following lemma, which is proved in Section A.

  Lemma 4 (Proximity to mean of true centers)**.** *Under the assumption of Theorem 3, let $\boldsymbol{\beta}$ be a local minimum of G. The following is true for each $i \in [k]$. If $\rho_s(\partial_{j,\ell}) > \lambda = \frac{c}{\sqrt{r\Delta_{\max}}}$ for some $(s,j,\ell) \in T_i \times [k] \times [k]$ and $|A_i \setminus \{s\}| \ge 1$, then we have the bounds*

$$\|\boldsymbol{\beta}_i - \boldsymbol{b}_i^-\| \le \Delta_{\max}\frac{2ck^2\left(1+\frac{2k^*}{k}\right) + 4k^*\left(3+\frac{2k^*}{k}\right)}{\sqrt{\eta_{\max}}},$$

$$\|\boldsymbol{\beta}_i - \boldsymbol{b}_i^+\| \le \Delta_{\max}\frac{2ck^2\left(1+\frac{2k^*}{k}\right) + 4k^*\left(3+\frac{2k^*}{k}\right)}{\sqrt{\eta_{\max}}},$$

  where $\boldsymbol{b}_i^- := \frac{1}{|A_i\setminus\{s\}|}\sum_{s' \in A_i\setminus\{s\}}\boldsymbol{\beta}_{s'}^*$ and $\boldsymbol{b}_i^+ := \frac{1}{|A_i\cup\{s\}|}\sum_{s' \in A_i\cup\{s\}}\boldsymbol{\beta}_{s'}^*$; moreover, we have $|A_i \setminus \{s\}| \ge 2$.

  In this case, we set $S_a = \{i\}$. Also set $S_a^* = A_i \setminus \{s\}$ if the index s has appeared in the sets $\{S_{a'}^*\}$ constructed previously in Step 2 or in this step, and set $S_a^* = A_i \cup \{s\}$ otherwise. Note that $|S_a^*| \ge 2$ by Lemma 4. As shall become clear momentarily, the flexibility allowed by Lemma 4 is important for ensuring that $\{S_a^*\}$ indeed partitions $[k^*]$.

In both cases above, we have the bound $\left\|\boldsymbol{\beta}_i - \frac{1}{|S_a^*|}\sum_{s' \in S_a^*}\boldsymbol{\beta}_{s'}^*\right\| \le \Delta_{\max}\frac{2ck^2\left(1+\frac{2k^*}{k}\right)+4k^*\left(3+\frac{2k^*}{k}\right)}{\sqrt{\eta_{\max}}}$ as claimed in Theorem 3. It is clear that the sets $\{S_a^*\}$ constructed in this step are disjoint from each other and from those constructed in Step 2, because the sets $\{A_i\}$ are disjoint as each true center can be in the interior of only one Voronoi set.

d) Summary: The above procedure constructs a collection of sets $\{S_a\}_{a=0}^m$ and $\{S_a^*\}_{a=1}^m$, which index the fitted and true centers, respectively, and satisfy the bounds in Theorem 3. The sets $\{S_a\}$ indeed form a partition of $[k]$, as we have $\bigcup_{a=0}^m S_a = S_0 \cup \mathcal{J} \cup \mathcal{K} = [k]$ by definition, and $S_a \cap S_b = \emptyset, \forall a \ne b$ by construction and the fact that the three sets $S_0, \mathcal{J}, \mathcal{K}$ are disjoint. For the sets $\{S_a^*\}$, we have argued in the construction above that they are disjoint. On the other hand, each true center $\boldsymbol{\beta}_s^*$ must belong to at least one Voronoi set $\mathcal{V}_i$, in which case we have $s \in T_i$. Consider two complementary cases: (i) If $\exists(j,\ell) : \rho_s(\partial_{j,\ell}) > \lambda$, then s must be covered in the second case of either Step 2 or Step 3. (ii) If $\forall(j,\ell) : \rho_s(\partial_{j,\ell}) \le \lambda$, then Observation 2 ensures that $s \in A_i \subseteq T_i$. In this case, if all other $s' \in T_i$ satisfies $\forall(j,\ell) : \rho_{s'}(\partial_{j,\ell}) \le \lambda$ as well, then s is covered in the first case of either Step 2 or Step 3. Otherwise, if there exists another $s' \in T_i$ satisfying $\exists(j,\ell) : \rho_{s'}(\partial_{j,\ell}) > \lambda$, then s is covered in the second case of Step 3 (with the role of s and s' exchanged therein) as $s \in A_i \subseteq A_i \setminus \{s'\}$. We conclude that the collection of sets $\{S_a^*\}$ covers all $s \in [k]$ and hence is indeed a partition of $[k]$. This completes the proof of Theorem 3.

C. Proof of Theorem 4

In this section, we prove Theorem 4, which shows that a local minimum $\boldsymbol{\beta}$ satisfies a family of bounds that imply our main Theorem 3.

a) Proof strategy: To derive structural properties of the local minimum $\boldsymbol{\beta}$, we exploit the fact that $t = 0$ is a local minimum of the directional objective $H^{\boldsymbol{v}}(t)$ (or a smooth upper bound thereof) for any perturbation direction $\boldsymbol{v} \in \mathbb{R}^{d\times k}$; see

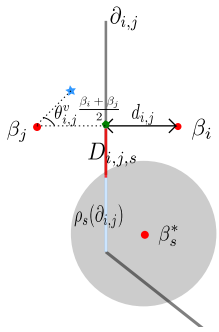

Figure 4. Illustration of the quantities $d_{i,j}$, $\theta^{\boldsymbol{v}}_{i,j}$, $D_{i,j,s}$ and $\rho_s(\partial_{i,j})$. The red dots represent $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$, and the blue star represents the perturbed solution $\boldsymbol{\beta}_j + t\boldsymbol{v}_j$. Here $d_{i,j}$ is the distance between $\boldsymbol{\beta}_i$ and the mid point $\frac{\boldsymbol{\beta}_i+\boldsymbol{\beta}_j}{2}$ (represented by a green dot); $D_{i,j,s}$, which is represented by the red line segment, is the distance between $\frac{\boldsymbol{\beta}_i+\boldsymbol{\beta}_j}{2}$ and the ball $\mathbb{B}_s$ when computed within the hyperplane containing the Voronoi boundary $\partial_{i,j}$; $\theta^{\boldsymbol{v}}_{i,j}$ is the angle between the perturbation direction $\boldsymbol{v}_j$ and the direction of $\boldsymbol{\beta}_i - \boldsymbol{\beta}_j$; $\rho_s(\partial_{i,j})$ is the normalized relative volume of the set $\partial_{i,j} \cap \mathbb{B}_s$, which is represented by the blue line segment inside the ball.

Lemma 3 and the discussion in Section V. The expression (10) of $H^{\boldsymbol{v}}$ involves the set $\Delta^{\boldsymbol{v}}_{i\to j}$, which are points that switch from one Voronoi set from another when $\boldsymbol{\beta}$ is perturbed to $\boldsymbol{\beta} + t\boldsymbol{v}$. For a general direction $\boldsymbol{v}$, these sets are quite complicated. Our main idea is to focus on a special class of directions satisfying

$$\|\boldsymbol{v}_i\| = 1, \forall i \in [k]; \qquad \boldsymbol{v}_i = \boldsymbol{v}_j \text{ or } \boldsymbol{v}_i = -\boldsymbol{v}_j, \forall i \neq j. \quad (13)$$

That is, we perturb the $\boldsymbol{\beta}_i$'s along the same or opposite directions. For these choices of $\boldsymbol{v}$, the Voronoi boundary $\partial_{i,j}(\boldsymbol{\beta} + t\boldsymbol{v})$ behaves in a simple way. In particular, when $\boldsymbol{v}_i = \boldsymbol{v}_j$, the boundary $\partial_{i,j}(\boldsymbol{\beta} + t\boldsymbol{v})$ translates along the direction of $\boldsymbol{v}_i$; when $\boldsymbol{v}_i = -\boldsymbol{v}_j$, the boundary rotates around the mid point $\frac{\boldsymbol{\beta}_i+\boldsymbol{\beta}_j}{2}$. See Figure 3 for an illustration. Using this fact, we can construct simple, tractable upper bounds of $H^{\boldsymbol{v}}$, from which we can deduce the structural properties of the local minimum $\boldsymbol{\beta}$.

b) Key quantities: Our analysis involves several key quantities related to the Voronoi sets of $\boldsymbol{\beta}$ and their boundaries. In particular, for each pair $i \neq j$ whose associated Voronoi sets are adjacent, i.e., $\mathcal{V}_i(\boldsymbol{\beta}) \sim \mathcal{V}_j(\boldsymbol{\beta})$, we introduce the following four quantities.

(a) Denote by $d_{i,j} := \frac{1}{2}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|$ the distance between $\boldsymbol{\beta}_i$ (or $\boldsymbol{\beta}_j$) and the Voronoi boundary $\partial_{i,j} \equiv \partial_{i,j}(\boldsymbol{\beta})$.

(b) Denote by $\theta^{\boldsymbol{v}}_{i,j} = \angle(\boldsymbol{v}_j, \boldsymbol{\beta}_i - \boldsymbol{\beta}_j)$ the angle of the perturbation direction $\boldsymbol{v}_j$ with respect to $\boldsymbol{\beta}_i - \boldsymbol{\beta}_j$.

(c) Define $D_{i,j,s} := \mathrm{dist}\big(\frac{\boldsymbol{\beta}_i+\boldsymbol{\beta}_j}{2}, \mathbb{B}_s \cap \partial_{i,j}\big)$, with the convention that $D_{i,j,s} = 1$ if $\mathbb{B}_s \cap \partial_{i,j} = \emptyset$. Here $D_{i,j,s}$ is the distance between the mid-point $\frac{\boldsymbol{\beta}_i+\boldsymbol{\beta}_j}{2}$ and s-th ground truth cluster $\mathbb{B}_s$, where the distance is computed within the hyperplane containing the Voronoi boundary $\partial_{i,j}$.

(d) Recall the quantity $\rho_s(\partial_{i,j}) := \frac{1}{r^d V_d}\mathrm{ReVol}(\partial_{i,j} \cap \mathbb{B}_s)$ defined in the statement of Theorem 4. Note that $\rho_s(\partial_{i,j})$ is the relative volume of the intersection of the Voronoi boundary $\partial_{i,j}$ and the s-th ground truth cluster $\mathbb{B}_s$, normalized by the volume of $\mathbb{B}_s$.

An illustration of these quantities is given in Figure 4.

We are now ready to prove Theorem 4. We begin with the upper bound of $H^{\boldsymbol{v}}$ given in equation (11), restated in an equivalent way below:

$$H^{\boldsymbol{v}}(t) \leq U^{\boldsymbol{v}}(t) + \frac{1}{2k^*} \sum_{(i,j):\mathcal{V}_i \sim \mathcal{V}_j} \sum_{s=1}^{k^*} \Big(W^{\boldsymbol{v}}_{i\to j,s}(t) + W^{\boldsymbol{v}}_{j\to i,s}(t) \Big). \quad (14)$$

The above inequality holds with equality at $t = 0$, since by definition $U^{\boldsymbol{v}}(0) = H^{\boldsymbol{v}}(0)$ and $W^{\boldsymbol{v}}_{i\to j,s}(0) = 0, \forall\ i, j, s$. Moreover, when $\boldsymbol{\beta}$ is a local minimum of G, a quick calculation using Lemma 2 shows that

$$\lim_{t\to 0} \frac{1}{t}\big(U^{\boldsymbol{v}}(t) - U^{\boldsymbol{v}}(0)\big) = 0,$$
$$\lim_{t\to 0} \frac{1}{t^2}\big(U^{\boldsymbol{v}}(t) - U^{\boldsymbol{v}}(0)\big) = 1. \quad (15)$$

Under the specific choice of the direction $\boldsymbol{v}$ in equation (13), the function $W^{\boldsymbol{v}}_{i\to j} + W^{\boldsymbol{v}}_{j\to i}$ can be further upper bounded, in a small neighborhood of 0, by a smooth function with nice analytical properties. In particular, when the directions $\boldsymbol{v}_i = \boldsymbol{v}_j$ are the same, such an upper bound $\widetilde{W}^{\boldsymbol{v}}_{i,j,s}$ is given in the following proposition, which is proved in Section VI-D.

Proposition 3 (Upper bound, same direction)**.** *Let $\boldsymbol{\beta}$ be a local minimum of G and $\boldsymbol{v}$ satisfy $\|\boldsymbol{v}_i\| = 1, \forall i \in [k]$. If $\boldsymbol{v}_i = \boldsymbol{v}_j$, then $W^{\boldsymbol{v}}_{i\to j,s} + W^{\boldsymbol{v}}_{j\to i,s}$ is upper bounded in a neighborhood of 0 by some smooth function $\widetilde{W}^{\boldsymbol{v}}_{i,j,s}$ satisfying the following properties:*

1) $\widetilde{W}^{\boldsymbol{v}}_{i,j,s}(0) = 0$;

2) $\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{W}^{\boldsymbol{v}}_{i,j,s}(t)\,|_{t=0} = 0$;

3) $\lim_{t\to 0} \frac{1}{t^2}\widetilde{W}^{\boldsymbol{v}}_{i,j,s}(t)\,|_{t=0} = -2\cos^2(\theta^{\boldsymbol{v}}_{i,j}) d_{i,j} \cdot \rho_s(\partial_{i,j})$.

When the directions $\boldsymbol{v}_i = -\boldsymbol{v}_j$ are opposite and $d \geq 2$, an upper bound $\widehat{W}^{\boldsymbol{v}}_{i,j,s}$ is given in the following proposition, which is proved in Section VI-E.

Proposition 4 (Upper bound, opposite direction)**.** *Let $\boldsymbol{\beta}$ be a local minimum of G and $\boldsymbol{v}$ satisfy $\|\boldsymbol{v}_i\| = 1, \forall i \in [k]$. If $\boldsymbol{v}_i = -\boldsymbol{v}_j$ and $\boldsymbol{v}_i, \boldsymbol{v}_j \in \mathcal{L}_{i,j,s}$, then $W^{\boldsymbol{v}}_{i\to j,s} + W^{\boldsymbol{v}}_{j\to i,s}$ is upper bounded in a neighborhood of 0 by some smooth function $\widehat{W}^{\boldsymbol{v}}_{i,j,s}$ satisfying the following properties:*

1) $\widehat{W}^{\boldsymbol{v}}_{i,j,s}(0) = 0$;

2) $\frac{\mathrm{d}}{\mathrm{d}t}\widehat{W}^{\boldsymbol{v}}_{i,j,s}(t)\,|_{t=0} = 0$;

3) $\lim_{t\to 0} \frac{1}{t^2}\widehat{W}^{\boldsymbol{v}}_{i,j,s}(t)\,|_{t=0} = -2\frac{D^2_{i,j,s}}{d_{i,j}} \sin^2(\theta^{\boldsymbol{v}}_{i,j}) \cdot \rho_s(\partial_{i,j})$.

Also note that $W^{\boldsymbol{v}}_{i\to j,s} \leq 0, \forall (i, j, s)$ by Remark 1. For each $s \in [k^*]$ and each un-ordered pair $(i, j) \in [k] \times [k]$ satisfying $\mathcal{V}_i \sim \mathcal{V}_j$, combining equation (14) and the last two propositions give the following smooth upper bound $\widetilde{H}^{\boldsymbol{v}}_{i,j}$ of $H^{\boldsymbol{v}}$:

$$H^{\boldsymbol{v}}(t) \leq \widetilde{H}^{\boldsymbol{v}}_{i,j}(t) := U^{\boldsymbol{v}}(t) + \frac{1}{k^*}\widetilde{W}^{\boldsymbol{v}}_{i,j,s}(t)\mathbb{1}_{\{\boldsymbol{v}_i = \boldsymbol{v}_j\}} + \frac{1}{k^*}\widehat{W}^{\boldsymbol{v}}_{i,j,s}(t)\mathbb{1}_{\{\boldsymbol{v}_i = -\boldsymbol{v}_j \in \mathcal{L}_{i,j,s}\}},$$

which is valid in a neighborhood of 0 and satisfies $\widetilde{H}^{\boldsymbol{v}}_{i,j}(0) = H^{\boldsymbol{v}}(0)$. Since $t = 0$ is a local minimum of $H^{\boldsymbol{v}}$, Lemma 3 ensures that

$$\lim_{t\to 0} \frac{1}{t^2}\Big[\widetilde{H}^{\boldsymbol{v}}_{i,j}(t) - \widetilde{H}^{\boldsymbol{v}}_{i,j}(0)\Big] \geq 0. \quad (16)$$

Moreover, by combining equation (15), Proposition 3 and Proposition 4, we obtain that

$$\begin{aligned}&\lim_{t\to 0}\frac{1}{t^2}\left[\widetilde{H}^{\boldsymbol{v}}_{i,j}(t)-\widetilde{H}^{\boldsymbol{v}}_{i,j}(0)\right]\\ =&1-\frac{2}{k^*}\cos^2(\theta^{\boldsymbol{v}}_{i,j})d_{i,j}\cdot\rho_s(\partial_{i,j})\mathbb{1}_{\{\boldsymbol{v}_i=\boldsymbol{v}_j\}}\\ &-\frac{2}{k^*}\frac{D^2_{i,j,s}}{d_{i,j}}\sin^2(\theta^{\boldsymbol{v}}_{i,j})\rho_s(\partial_{i,j})\mathbb{1}_{\{\boldsymbol{v}_i=-\boldsymbol{v}_j\in\mathcal{L}_{i,j,s}\}}.\end{aligned}\tag{17}$$

Since equation (17) holds for any choice of $\boldsymbol{v}$ satisfying the condition (13), we may choose $\boldsymbol{v}$ judiciously to simplify the right hand side of (17). By doing so we can show that for each $s\in[k^*]$ and each pair $i\neq j\in[k]$ satisfying $\mathcal{V}_i\sim\mathcal{V}_j$, there hold the inequalities

$$d_{i,j}\cdot\rho_s(\partial_{i,j})\le\frac{k^*}{2}\qquad\text{and}\tag{18}$$

$$\frac{D^2_{i,j,s}}{d_{i,j}}\cdot\rho_s(\partial_{i,j})\le\frac{k^*}{2},\tag{19}$$

where the second inequality is valid when $d\ge 2$. To prove the inequality (18), suppose otherwise that $d_{i,j}\cdot\rho_s(\partial_{i,j})>\frac{k^*}{2}$ for some (i,j,s). We can choose the directions $\boldsymbol{v}_i=\boldsymbol{v}_j=\frac{\boldsymbol{\beta}_i-\boldsymbol{\beta}_j}{\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\|}$, which satisfies $\theta^{\boldsymbol{v}}_{i,j}=0$. Combining with equation (17) gives

$$\lim_{t\to 0}\frac{1}{t^2}\left(\widetilde{H}^{\boldsymbol{v}}_{i,j}(t)-\widetilde{H}^{\boldsymbol{v}}(0)\right)<0,$$

which contradicts the inequality (16). Similarly, to prove the inequality (19), suppose otherwise that $\frac{D^2_{i,j,s}}{d_{i,j}}\cdot\rho_s(\partial_{i,j})>\frac{k^*}{2}$ for some (i,j,s) when $d\ge 2$. We can choose $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ to be two unit vectors in the two-dimensional plane $\mathcal{L}_{i,j,s}$ such that $\boldsymbol{v}_i=-\boldsymbol{v}_j$ and $\boldsymbol{v}_i\perp(\boldsymbol{\beta}_j-\boldsymbol{\beta}_i)$, which satisfies $\theta^{\boldsymbol{v}}_{i,j}=\frac{\pi}{2}$. Combining with equation (17) gives

$$\lim_{t\to 0}\frac{1}{t^2}\left(\widetilde{H}^{\boldsymbol{v}}_{i,j}(t)-\widetilde{H}^{\boldsymbol{v}}(0)\right)<0,$$

which again contradicts the inequality (16).

In the remaining of the proof, fix an index $i\in[k]$ and a number $\lambda>0$. We shall use equations (18) and (19) to derive the structural properties of $\boldsymbol{\beta}_i$ and its Voronoi set $\mathcal{V}_i$. To this end, we consider two complementary cases that correspond to Part 1 and Part 2 of Theorem 4. Recall that $T_i:=\{s\in[k^*]:\mathcal{V}_i\cap\mathbb{B}_s\neq\emptyset\}$.

Case 1: there exists some $(s,j,\ell)\in T_i\times[k]\times[k]$ such that $\rho_s(\partial_{j,\ell})>\lambda$.: In this case, the Voronoi set $\mathcal{V}_i$ intersects a true cluster $\mathbb{B}_s$ that encloses some Voronoi boundary $\partial_{j,\ell}$ with a large relative volume. Note that this case corresponds to Part 1 of Theorem 4.

Under the case condition, the inequality (18) implies that $d_{j,\ell}\le\frac{k^*}{2\rho_s(\partial_{j,\ell})}\le\frac{k^*}{2\lambda}$; equivalently,

$$\left\|\frac{\boldsymbol{\beta}_j+\boldsymbol{\beta}_\ell}{2}-\boldsymbol{\beta}_j\right\|=\left\|\frac{\boldsymbol{\beta}_j+\boldsymbol{\beta}_\ell}{2}-\boldsymbol{\beta}_\ell\right\|\le\frac{k^*}{2\lambda}.$$

We consider the one-dimensional and high-dimensional cases separately. When $d=1$, the case condition $\rho_s(\partial_{j,\ell})>\lambda$ further implies that $\frac{\beta_j+\beta_\ell}{2}\in\mathbb{B}_s$ and hence $|\frac{\beta_j+\beta_\ell}{2}-\beta^*_s|\le r$. It follows that $|\beta_j-\beta^*_s|\le\left|\beta_j-\frac{\beta_j+\beta_\ell}{2}\right|+\left|\frac{\beta_j+\beta_\ell}{2}-\beta^*_s\right|\le\frac{k^*}{2\lambda}+r$. When $d\ge 2$, we have $D_{j,\ell,s}\le\frac{k^*}{2\lambda}$ by multiplying the inequalities (18) and (19). Let $\boldsymbol{z}\in\mathbb{B}_s\cap\partial_{j,\ell}$ be the point that attains $\left\|\frac{\boldsymbol{\beta}_j+\boldsymbol{\beta}_\ell}{2}-\boldsymbol{z}\right\|=\operatorname{dist}\left(\frac{\boldsymbol{\beta}_j+\boldsymbol{\beta}_\ell}{2},\mathbb{B}_s\cap\partial_{j,\ell}\right)=D_{j,\ell,s}$. It follows that

$$\begin{aligned}\|\boldsymbol{\beta}_j-\boldsymbol{\beta}^*_s\|&\le\left\|\boldsymbol{\beta}_j-\frac{\boldsymbol{\beta}_j+\boldsymbol{\beta}_\ell}{2}\right\|+\left\|\frac{\boldsymbol{\beta}_j+\boldsymbol{\beta}_\ell}{2}-\boldsymbol{z}\right\|+\|\boldsymbol{z}-\boldsymbol{\beta}^*_s\|\\&\le\frac{k^*}{2\lambda}+\frac{k^*}{2\lambda}+r=\frac{k^*}{\lambda}+r.\end{aligned}$$

In either case of d, we have the bound $\|\boldsymbol{\beta}_j-\boldsymbol{\beta}^*_s\|\le\frac{k^*}{\lambda}+r$. Since $s\in T_i$, there exist a point $\boldsymbol{x}\in\mathbb{B}_s\cap\mathcal{V}_i$. It follows that

$$\begin{aligned}\|\boldsymbol{\beta}_i-\boldsymbol{\beta}^*_s\|\le&\|\boldsymbol{\beta}_i-\boldsymbol{x}\|+\|\boldsymbol{x}-\boldsymbol{\beta}^*_s\|\\\overset{\text{(i)}}{\le}&\|\boldsymbol{\beta}_j-\boldsymbol{x}\|+\|\boldsymbol{x}-\boldsymbol{\beta}^*_s\|\\\le&(\|\boldsymbol{\beta}_j-\boldsymbol{\beta}^*_s\|+\|\boldsymbol{x}-\boldsymbol{\beta}^*_s\|)+\|\boldsymbol{x}-\boldsymbol{\beta}^*_s\|\overset{\text{(ii)}}{\le}\frac{k^*}{\lambda}+3r,\end{aligned}$$

where step (i) follows from $\boldsymbol{x}\in\mathcal{V}_i$, and step (ii) follows from $\boldsymbol{x}\in\mathbb{B}_s$ and the bound on $\|\boldsymbol{\beta}_j-\boldsymbol{\beta}^*_s\|$ proved above. We have established Part 1 of Theorem 4.

Case 2: for all $(s,j,\ell)\in T_i\times[k]\times[k]$ there holds $\rho_s(\partial_{j,\ell})\le\lambda$.: In this case, for all true clusters $\{\mathbb{B}_s\}$ that intersect the Voronoi set $\mathcal{V}_i$, all the Voronoi boundaries enclosed by $\mathbb{B}_s$ have a small relative volume.

Let us partition the set $T_i:=\{s\in[k^*]:\mathcal{V}_i\cap\mathbb{B}_s\neq\emptyset\}$ into two subsets defined as follows:

$$\begin{aligned}A_i&:=\{s\in T_i:\boldsymbol{\beta}^*_s\in\operatorname{int}(\mathcal{V}_i)\}\quad\text{and}\\B_i&:=T_i\setminus A_i=\{s\in T_i:\boldsymbol{\beta}^*_s\notin\operatorname{int}(\mathcal{V}_i)\}.\end{aligned}$$

Here A_i indexes the ground truth clusters whose centers are in the interior of the Voronoi set $\mathcal{V}_i$; B_i indexes the ground truth clusters that intersect $\mathcal{V}_i$ but their centers are outside its interior (i.e., the center either lies on a Voronoi boundary or in some other Voronoi set $\mathcal{V}_j$). Also recall the quantities $m_{i,s}\equiv m_{i,s}(\boldsymbol{\beta})$ and $\boldsymbol{c}_{i,s}\equiv\boldsymbol{c}_{i,s}(\boldsymbol{\beta})$ introduced after Lemma 2; in particular, $m_{i,s}$ is the probability mass of $\mathcal{V}_i\cap\mathbb{B}_s$ with respect to $\mathbb{P}_s$, and $\boldsymbol{c}_{i,s}$ is the corresponding center of mass.

Note the Voronoi set $\mathcal{V}_i$ is a polyhedron with at most k facets. For each $s\in B_i$, if $\rho_s(\partial_{j,\ell})\le\lambda$ for all $\forall(j,\ell)$ (and in particular, for $j=i$), then all facets of $\mathcal{V}_i$ intersect the ball $\mathbb{B}_s$ with a small relative volume. Moreover, $\boldsymbol{\beta}^*_s$ is not in $\operatorname{int}(\mathcal{V}_i)$. With these two facts, an elementary geometric argument (formally given in Lemma 8) shows that the intersection $\mathcal{V}_i\cap\mathbb{B}_s$ must have a small mass; that is,

$$m_{i,s}=\mathbb{P}_s(\mathcal{V}_i)\le k\lambda r,\quad\forall s\in B_i.\tag{20}$$

On the other hand, for each $s\in A_i$, we must have $\boldsymbol{\beta}^*_s\notin\operatorname{int}(\mathcal{V}_j),\forall j\neq i$, since the interiors of Voronoi sets are disjoint. Repeating the same argument above shows that $\mathbb{P}_s(\mathcal{V}_j)\le k\lambda,\forall j\neq i$, whence

$$m_{i,s}=\mathbb{P}_s(\mathcal{V}_i)=1-\sum_{j:j\neq i}\mathbb{P}_s(\mathcal{V}_j)\ge 1-k^2\lambda r,\quad\forall s\in A_i.\tag{21}$$

The inequalities (20) and (21) establish the first two bounds in Part 2 of Theorem 4.

We next turn to Part 2(a) of Theorem 4, which concerns the case with $A_i=\emptyset$. This means that $T_i=B_i$. We therefore have

$$\mathbb{P}(\mathcal{V}_i)=\frac{1}{k^*}\sum_{s\in[k^*]}m_{i,s}\le k\lambda r,$$

where the last step holds due to equation (20) and the fact that $m_{i,s} = 0, \forall s \notin T_i$.

Finally, we consider Part 2(b) of Theorem 4, which concerns the case with $A_i \neq \emptyset$ and hence $\mathbb{P}(\mathcal{V}_i) > 0$. Since $\boldsymbol{\beta}$ is a local minimum, Lemma 2 and the discussion thereafter ensure that

$$\boldsymbol{\beta}_i = \frac{\sum_{s=1}^{k^*} m_{i,s}\boldsymbol{c}_{i,s}}{\sum_{s=1}^{k^*} m_{i,s}} = \frac{\sum_{s\in T_i} m_{i,s}\boldsymbol{c}_{i,s}}{\sum_{s\in T_i} m_{i,s}}.$$

Recalling the definition $\boldsymbol{b}_i := \frac{1}{|A_i|}\sum_{s\in A_i}\boldsymbol{\beta}^*_s$, we can decompose the quantity $(\boldsymbol{\beta}_i - \boldsymbol{b}_i)$ of interest as follows:

$$\begin{aligned}\boldsymbol{\beta}_i - \boldsymbol{b}_i =& \frac{\sum_{s\in A_i} m_{i,s}\boldsymbol{c}_{i,s} + \sum_{s\in B_i} m_{i,s}\boldsymbol{c}_{i,s}}{\sum_{s\in A_i} m_{i,s} + \sum_{s\in B_i} m_{i,s}} - \frac{1}{|A_i|}\sum_{s\in A_i}\boldsymbol{\beta}^*_s \\ =& \underbrace{\frac{\sum_{s\in A_i} m_{i,s}\boldsymbol{c}_{i,s} + \sum_{s\in B_i} m_{i,s}\boldsymbol{c}_{i,s}}{\sum_{s\in A_i} m_{i,s} + \sum_{s\in B_i} m_{i,s}} - \frac{\sum_{s\in A_i} m_{i,s}\boldsymbol{c}_{i,s}}{\sum_{s\in A_i} m_{i,s}}}_{\boldsymbol{\mu}} \\ &+ \underbrace{\frac{\sum_{s\in A_i} m_{i,s}\boldsymbol{c}_{i,s}}{\sum_{s\in A_i} m_{i,s}} - \frac{1}{|A_i|}\sum_{s\in A_i}\boldsymbol{\beta}^*_s}_{\boldsymbol{\nu}}. \end{aligned} \tag{22}$$

The following two lemmas, proved in Section VI-F, control the norms of the vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$.

Lemma 5. *We have* $\|\boldsymbol{\mu}\| \le \frac{k^* r}{1-k^2\lambda r} + \frac{k^*\cdot(k\lambda r)\cdot(k^2\lambda r^2)}{(1-k^2\lambda r)^2} + \frac{2k^* k\lambda r}{1-k^2\lambda r}\Delta_{\max}$.

Lemma 6. *We have* $\|\boldsymbol{\nu}\| \le \frac{k^2\lambda r^2}{1-k^2\lambda r} + \frac{k^2\lambda r}{1-k^2\lambda r}\Delta_{\max}$.

Applying these two lemmas to bound the right hand side of equation (22), we obtain that

$$\begin{aligned}\|\boldsymbol{\beta}_i - \boldsymbol{b}_i\| \le& \frac{k^* r}{1-k^2\lambda r} + \frac{k^2\lambda r^2(1+(k^*k-k^2)\lambda r)}{(1-k^2\lambda r)^2} \\ &+ \frac{(k+2k^*)k\lambda r}{1-k^2\lambda r}\Delta_{\max},\end{aligned}$$

thereby proving Part 2(b) of Theorem 4.

We have completed the proof of Theorem 4.

D. Proof of Proposition 3 (Same Direction)

Under the perturbation direction $\boldsymbol{v}_i = \boldsymbol{v}_j$, the new Voronoi boundary $\partial_{i,j}(\boldsymbol{\beta} + t\boldsymbol{v})$ is a translation of the original boundary $\partial_{i,j}(\boldsymbol{\beta})$ by the amount $t\boldsymbol{v}$; see top panel of Figure 3. When $\partial_{i,j}(\boldsymbol{\beta})$ intersects $\mathbb{B}_s$ trivially, with measure 0 with respect to $\mathbb{P}_s$, setting $\widetilde{W}^{\boldsymbol{v}}_{i,j,s} \equiv 0$ satisfies the conclusions of the proposition as $\rho_s(\partial_{i,j}) = 0$ in this case. Thus we only need to consider the case where $\partial_{i,j}(\boldsymbol{\beta})$ intersects $\mathbb{B}_s$ non-trivially. We assume WLOG that $\mathbb{P}_s(\Delta^{\boldsymbol{v}}_{i\to j}) > 0$ and upper bound $W^{\boldsymbol{v}}_{i\to j,s} + W^{\boldsymbol{v}}_{j\to i,s}$ by $W^{\boldsymbol{v}}_{i\to j,s}$ (as $W^{\boldsymbol{v}}_{j\to i,s}$ is non-positive). It remains to upper bound $W^{\boldsymbol{v}}_{i\to j,s}$.

Recall expression for $W^{\boldsymbol{v}}_{i\to j,s}$:

$$\begin{aligned}&W^{\boldsymbol{v}}_{i\to j,s}(t) \\ :=& \int_{\Delta^{\boldsymbol{v}}_{i\to j}(t)} (\|\boldsymbol{x} - \boldsymbol{\beta}_j - t\boldsymbol{v}_j\|^2 - \|\boldsymbol{x} - \boldsymbol{\beta}_i - t\boldsymbol{v}_i\|^2) f_s(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ =& \int_{\Delta^{\boldsymbol{v}}_{i\to j}(t)} \Big[2\langle \boldsymbol{x}, \boldsymbol{\beta}_i + t\boldsymbol{v}_i - \boldsymbol{\beta}_j - t\boldsymbol{v}_j\rangle \\ &+ (\|\boldsymbol{\beta}_j + t\boldsymbol{v}_j\|^2 - \|\boldsymbol{\beta}_i + t\boldsymbol{v}_i\|^2)\Big] f_s(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned}$$

Since the integrand above only involves the Euclidean norm, we are free to choose any coordinate system. In particular, we choose the origin to be $\frac{1}{2}(\boldsymbol{\beta}_i + \boldsymbol{\beta}_j)$, the principal axis to be the direction of $\boldsymbol{\beta}_i - \boldsymbol{\beta}_j$, and the secondary axis to be the direction orthogonal to $\boldsymbol{\beta}_i - \boldsymbol{\beta}_j$ and in span$\{\boldsymbol{\beta}_i - \boldsymbol{\beta}_j, \boldsymbol{v}_i\}$. Under this coordinate system, we have

$$\begin{aligned}W^{\boldsymbol{v}}_{i\to j,s}(t) =& 2\int_{\Delta^{\boldsymbol{v}}_{i\to j}(t)} (\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\| x_1 - t\langle \boldsymbol{\beta}_i - \boldsymbol{\beta}_j, \boldsymbol{v}_i\rangle) f_s(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ =& 2\int_{x_2,\ldots,x_d:\boldsymbol{x}\in\Delta^{\boldsymbol{v}}_{i\to j}(t)} \int_{x_1=0}^{t\cos(\theta)} \Big[\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\| x_1 \\ &- t\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|\cos(\theta)\Big] f_s(\boldsymbol{x})\mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_d,\end{aligned}$$

where we introduce the shorthand $\theta \equiv \theta^{\boldsymbol{v}}_{i,j} := \angle(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j, \boldsymbol{v}_j) = \angle(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j, \boldsymbol{v}_i)$ and, slightly abusing notation, still use f_s to denote the density function under the new coordinate system. Note that the region

$$S(z,t) := \{(x_2,\ldots,x_d) : \boldsymbol{x} \in \Delta^{\boldsymbol{v}}_{i\to j}(t), x_1 = z\}$$

is a vertical slice under the current coordinate system; in particular, $S(z,t)$ is the intersection of the set $\Delta^{\boldsymbol{v}}_{i\to j,s}(t)$ and the hyperplane that is parallel to $\partial_{i,j}$ and at a distance z from $\partial_{i,j}$. Defining the integral

$$\rho^{\boldsymbol{v}}_{i\to j,s}(z,t) := \int_{S(z,t)} f_s(z, x_2, \ldots, x_d)\mathrm{d}x_2\ldots\mathrm{d}x_d,$$

we can write

$$\begin{aligned}W^{\boldsymbol{v}}_{i\to j,s}(t) =& 2\int_{x_1=0}^{t\cos(\theta)} \Big[\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\| x_1 \\ &- t\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|\cos(\theta)\Big]\rho^{\boldsymbol{v}}_{i\to j,s}(x_1,t)\mathrm{d}x_1.\end{aligned} \tag{23}$$

When t is small, we have the sandwich bound $m(t) \le \rho^{\boldsymbol{v}}_{i\to j,s}(x_1,t) \le M(t)$, where

$$m(t) := \min_{x_1\in[0,t\cos(\theta)]} \rho^{\boldsymbol{v}}_{i\to j,s}(x_1,t),$$

and

$$M(t) := \max_{x_1\in[0,t\cos(\theta)]} \rho^{\boldsymbol{v}}_{i\to j,s}(x_1,t).$$

Here $m(t)$ and $M(t)$ are well-defined as they are the max/min of the bounded function $\rho^{\boldsymbol{v}}_{i\to j,s}$ over the compact interval $[0, t\cos(\theta)]$. Moreover, $m(t)$ and $M(t)$ satisfy

$$\lim_{t\to 0} m(t) = \lim_{t\to 0} M(t) = \frac{\mathrm{ReVol}(\partial_{i,j})}{\mathrm{Vol}(\mathbb{B}_s(r))} = \rho_s(\partial_{i,j}).$$

Bounding the two terms in the bracket in equation (23) separately, we obtain that

$$\begin{aligned}&2\int_{x_1=0}^{t\cos(\theta)} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\| x_1 \rho^{\boldsymbol{v}}_{i\to j,s}(x_1,t)\mathrm{d}x_1 \\ \in& \left[2d_{i,j}\cos^2(\theta)\cdot m(t)t^2, 2d_{i,j}\cos^2(\theta)\cdot M(t)t^2\right]\end{aligned}$$

and

$$\begin{aligned}&2t\int_{x_1=0}^{t\cos(\theta)} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|\cos(\theta)\rho^{\boldsymbol{v}}_{i\to j,s}(x_1,t)\mathrm{d}x_1 \\ \in& \left[4d_{i,j}\cos^2(\theta)\cdot m(t)t^2, 4d_{i,j}\cos^2(\theta)\cdot M(t)t^2\right]\end{aligned}$$

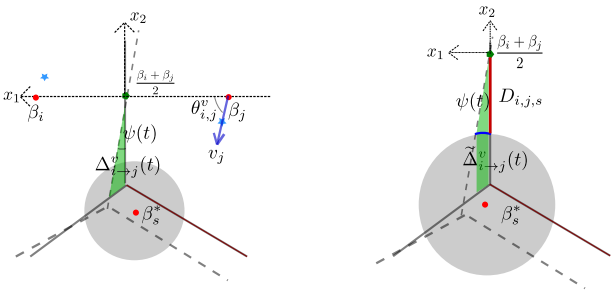

Figure 5. Illustration of the local coordinate system and the upper bound function. The local coordinate system has the origin at $\frac{\boldsymbol{\beta}_i+\boldsymbol{\beta}_j}{2}$, represented by the dark green dot. Its principal axis is in the direction of $\boldsymbol{\beta}_i-\boldsymbol{\beta}_j$, plotted as the x_1 axis, and its secondary axis is plotted as the x_2 axis. In the left panel, the red dots represent $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ respectively; the blue stars represent $\boldsymbol{\beta}_i+t\boldsymbol{v}_i$ and $\boldsymbol{\beta}_j+t\boldsymbol{v}_j$ respectively, with $\boldsymbol{v}_i=-\boldsymbol{v}_j$. The dark blue arrow indicates the direction of $\boldsymbol{v}_j$ and it has an angle $\theta^v_{i,j}$ with the vector $\boldsymbol{\beta}_i-\boldsymbol{\beta}_j$, the x_1 axis. Correspondingly, the Voronoi boundary $\partial_{i,j}(\boldsymbol{\beta}+t\boldsymbol{v})$ rotates around the origin with an angle $\psi(t)$. The boundaries $\partial(\boldsymbol{\beta}+t\boldsymbol{v})$ are plotted using dotted lines. The shaded green region in the left panel corresponds to the set $\Delta^{\boldsymbol{v}}_{i\to j}(t)$, in which the point becomes closer to $\boldsymbol{\beta}_j+t\boldsymbol{v}_j$ than $\boldsymbol{\beta}_i+t\boldsymbol{v}_i$ after $\boldsymbol{\beta}$ is moved to $\boldsymbol{\beta}+t\boldsymbol{v}$. In the right panel, we demonstrate the set $\widetilde{\Delta}^{\boldsymbol{v}}_{i\to j}(t)$ using the shaded green region. It is a subset of $\Delta^{\boldsymbol{v}}_{i\to j}(t)$, enclosed by the hyperplane $\{\boldsymbol{x}: x_1=0\}$ and the translated hyperplane $\{\boldsymbol{x}: x_1=D_{i,j,s}\tan(\psi(t))\}$

whence

$$\begin{aligned}&2(m(t)-2M(t))d_{i,j}\cos^2(\theta)t^2\le W^{\boldsymbol{v}}_{i\to j,s}(t)\\ \le&2(M(t)-2m(t))d_{i,j}\cos^2(\theta)t^2.\end{aligned}$$

It is then easy to see that $W^{\boldsymbol{v}}_{i\to j,s}(0)=0$, $\frac{\mathrm{d}}{\mathrm{d}t}W^{\boldsymbol{v}}_{i\to j,s}(t)\mid_{t=0}=0$ and

$$\lim_{t\to 0}\frac{W^{\boldsymbol{v}}_{i\to j,s}(t)}{t^2}=-2d_{i,j}\cos^2(\theta)\rho_s(\partial_{i,j}).$$

In summary, setting $\widetilde{W}^{\boldsymbol{v}}_{i,j,s}=W^{\boldsymbol{v}}_{i\to j,s}$, we have established that $W^{\boldsymbol{v}}_{i\to j,s}+W^{v}_{j\to i,s}\le\widetilde{W}^{\boldsymbol{v}}_{i,j,s}$ and that $\widetilde{W}^{\boldsymbol{v}}_{i,j,s}$ satisfies the desired analytical properties in Proposition 3.

E. Proof of Proposition 4 (Opposite Direction)

Under the perturbation direction $\boldsymbol{v}_i=-\boldsymbol{v}_j$, the Voronoi boundary $\partial_{i,j}(\boldsymbol{\beta})$ rotates around the mid point $\frac{\boldsymbol{\beta}_i+\boldsymbol{\beta}_j}{2}$; see right panel of Figure 3. When $\partial_{i,j}(\boldsymbol{\beta})$ intersects $\mathbb{B}_s$ trivially, with measure 0 with respect to $\mathbb{P}_s$, setting $\widehat{W}_{i,j,s}\equiv 0$ satisfies the conclusion of the proposition, as $\rho_s(\partial_{i,j})=0$ in this case. When $\frac{1}{2}(\boldsymbol{\beta}_i+\boldsymbol{\beta}_j)\in\mathbb{B}_s$, we can also set $\widehat{W}^{\boldsymbol{v}}_{i\to j,s}\equiv 0$, as $D_{i,j,s}=0$ in this case. Thus in the rest of the proof, we assume WLOG that $\frac{1}{2}(\boldsymbol{\beta}_i+\boldsymbol{\beta}_j)\notin\mathbb{B}_s$ and $\mathbb{P}_s(\Delta^{\boldsymbol{v}}_{i\to j,s}(t))>0$. In this case we have $\rho_s(\partial_{i,j})>0$ and $D_{i,j,s}>0$. We upper bound $W^{\boldsymbol{v}}_{i\to j,s}+W^{\boldsymbol{v}}_{j\to i,s}$ by $W^{\boldsymbol{v}}_{i\to j,s}$ (as $W^{\boldsymbol{v}}_{j\to i,s}\le 0$). It remains to find an upper bound of $W^{\boldsymbol{v}}_{i\to j,s}$ that satisfies the desired analytical properties.

Similarly to Section VI-D, a convenient coordinate system is used. In particular, we choose the origin to be $\frac{1}{2}(\boldsymbol{\beta}_i+\boldsymbol{\beta}_j)$, the principal axis to be the direction of $\boldsymbol{\beta}_i-\boldsymbol{\beta}_j$, and the secondary axis to be the direction that is orthogonal to $\boldsymbol{\beta}_i-\boldsymbol{\beta}_j$ and in the plane $\mathcal{L}_{i,j,s}$; see the left panel of Figure 5. Under this coordinate system, we have the representation

$$\begin{aligned}&\boldsymbol{\beta}_i=(d_{i,j},0,\dots,0),\quad \boldsymbol{\beta}_j=(-d_{i,j},0,\dots,0),\\ &\boldsymbol{\beta}^*_s=(b^*_1,b^*_2,0,\dots,0)\end{aligned}\tag{24}$$

for some $b^*_1,b^*_2\in\mathbb{R}$. The orientation of the secondary axis can be chosen to satisfy $b^*_2<0$. We note that the boundary $\partial_{i,j}(\boldsymbol{\beta})$ rotates around $\frac{1}{2}(\boldsymbol{\beta}_i+\boldsymbol{\beta}_j)$ by an angle $\psi(t)$ that satisfies

$$\tan(\psi(t))=\frac{t\sin(\theta)}{d_{i,j}-t\cos(\theta)},\tag{25}$$

where $\theta\equiv\theta^{\boldsymbol{v}}_{i,j}\in[0,\pi/2]$ is the (unsigned) angle between $\boldsymbol{v}_j$ and $\boldsymbol{\beta}_i-\boldsymbol{\beta}_j$. Moreover, since we assume $\mathbb{P}_s(\Delta^{\boldsymbol{v}}_{i\to j}(t))>0$, the directions $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ have the following coordinate representation:

$$\boldsymbol{v}_j=\big(\cos(\theta),-\sin(\theta),0,\dots,0\big)=-\boldsymbol{v}_i.$$

We now proceed to upper bound the function $W^{\boldsymbol{v}}_{i\to j,s}$. Define the polyhedron set

$$\widetilde{\Delta}^{\boldsymbol{v}}_{i\to j}(t):=\big\{\boldsymbol{x}\in\Delta^{\boldsymbol{v}}_{i\to j}(t):x_1\le D_{i,j,s}\tan(\psi(t))\big\}.\tag{26}$$

The set $\widetilde{\Delta}^{\boldsymbol{v}}_{i\to j}(t)$ is sandwiched between the two hyperplanes $x_1=0$ and $x_1=D_{i,j,s}\tan(\psi(t))$; see the right panel of Figure 5 for an illustration. With the above notations, we can upper bound $W^{\boldsymbol{v}}_{i\to j,s}$ as follows:

$$\begin{aligned}W^{\boldsymbol{v}}_{i\to j,s}&\overset{\text{(i)}}{\le}\int_{\widetilde{\Delta}^{\boldsymbol{v}}_{i\to j}(t)}\Big[2\langle\boldsymbol{x},\boldsymbol{\beta}_i+t\boldsymbol{v}_i-\boldsymbol{\beta}_j-t\boldsymbol{v}_j\rangle\\ &\quad+\big(\|\boldsymbol{\beta}_j+t\boldsymbol{v}_j\|^2-\|\boldsymbol{\beta}_i+t\boldsymbol{v}_i\|^2\big)\Big]f_s(\boldsymbol{x})d\boldsymbol{x}\\ &\overset{\text{(ii)}}{=}\int_{\widetilde{\Delta}^{\boldsymbol{v}}_{i\to j}(t)}2\langle\boldsymbol{x},\boldsymbol{\beta}_i+t\boldsymbol{v}_i-\boldsymbol{\beta}_j-t\boldsymbol{v}_j\rangle f_s(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &=\int_{\widetilde{\Delta}^{\boldsymbol{v}}_{i\to j}(t)}\Big[2\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\|x_1-4t\cos(\theta)x_1\\ &\quad+4t\sin(\theta)x_2\Big]f_s(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\end{aligned}\tag{27}$$

where step (i) holds because the integrand in the definition of $W^{\boldsymbol{v}}_{i\to j,s}$ is non-positive and thus integrating over a smaller set $\widetilde{\Delta}^{\boldsymbol{v}}_{i\to j}(t)\subseteq\Delta^{\boldsymbol{v}}_{i\to j}(t)$ does not decrease the value of the integral, and step (ii) holds since under the current coordinate system, $\boldsymbol{\beta}_i=-\boldsymbol{\beta}_j$ and $\boldsymbol{v}_i=-\boldsymbol{v}_j$.

To proceed, we let

$$D(t):=\max\big\{x_2:\boldsymbol{x}\in\widetilde{\Delta}^{\boldsymbol{v}}_{i\to j}(t)\cap\mathbb{B}_s\big\}.\tag{28}$$

denote maximum of the second coordinate of the set $\widetilde{\Delta}^{\boldsymbol{v}}_{i\to j}(t)\cap\mathbb{B}_s$ under the current coordinate system. The following lemma, proved at the end of this section, characterizes the limit property of $D(t)$.

Lemma 7 (Negative second coordinate at the boundary)**.** *Suppose that $\boldsymbol{v}$ satisfies $\|\boldsymbol{v}_i\|=\|\boldsymbol{v}_j\|=1$ and $\boldsymbol{v}_i=$*

$-\boldsymbol{v}_j \in \mathcal{L}_{i,j,s}$, $\frac{\boldsymbol{\beta}_i+\boldsymbol{\beta}_j}{2} \notin \mathbb{B}_s$ *and* $\Delta^{\boldsymbol{v}}_{i\to j}(t) \cap \mathbb{B}_s \neq \emptyset$. *We have* $\lim_{t\to 0} D(t) = -D_{i,j,s}$.

Lemma 7 ensures that $\lim_{t\to 0} D(t) = -D_{i,j,s} < 0$. Consequently, when t is sufficiently small, we have $D(t) < 0$ by the continuity.

Continuing from the last display equation (27), we obtain our final upper bound $\widehat{W}^{\boldsymbol{v}}_{i,j,s}(t)$:

$$W^{\boldsymbol{v}}_{i\to j,s} \le \int_{\tilde{\Delta}^{\boldsymbol{v}}_{i\to j}(t)} \Big[2\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\| x_1 - 4t\cos(\theta)x_1 + 4t\sin(\theta)D(t)\Big] f_s(\boldsymbol{x})\mathrm{d}\boldsymbol{x} =: \widehat{W}^{\boldsymbol{v}}_{i,j,s}(t).$$

To establish the analytical properties of $\widehat{W}^{\boldsymbol{v}}_{i,j,s}(t)$, we follow a similar argument as in Section VI-D. Define the integral

$$\rho^{\boldsymbol{v}}_{i\to j,s}(z,t) := \int_{\boldsymbol{x}\in\tilde{\Delta}^{\boldsymbol{v}}_{i\to j}(t), x_1 = z} f_s(z, x_2, \ldots, x_d)\mathrm{d}x_2 \ldots \mathrm{d}x_d,$$

and rewrite $\widehat{W}^{\boldsymbol{v}}_{i,j,s}(t)$ compactly as follows:

$$\widehat{W}^{\boldsymbol{v}}_{i,j,s}(t) = \int_0^{D_{i,j,s}\tan(\psi(t))} \Big[2\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\| x_1 - 4t\cos(\theta)x_1 + 4t\sin(\theta)D(t)\Big] \rho^{\boldsymbol{v}}_{i\to j,s}(x_1,t)\mathrm{d}x_1.$$

We have the sandwich bound $m(t) \le \rho^{\boldsymbol{v}}_{i\to j,s}(x_1,t) \le M(t)$, valid for $x_1 \in \big[0, D_{i,j,s}\tan(\psi(t))\big]$, where $m(t) := \min_{x_1\in\left[0, D_{i,j,s}\tanh(\psi(t))\right]} \rho^{\boldsymbol{v}}_{i\to j,s}(x_1,t)$ and $M(t) := \max_{x_1\in\left[0, D_{i,j,s}\tan(\psi(t))\right]} \rho^{\boldsymbol{v}}_{i\to j,s}(x_1,t)$.

Moreover, the functions $m(\cdot)$ and $M(\cdot)$ satisfy

$$\lim_{t\to 0} m(t) = \lim_{t\to 0} M(t) = \frac{\mathrm{ReVol}(\partial_{i,j})}{\mathrm{Vol}(\mathbb{B}_s)} = \rho_s(\partial_{i,j}).$$

With some algebra as well as the non-positivity of $D(t)$, we obtain the following sandwich bound for $\widehat{W}^{\boldsymbol{v}}_{i,j,s}$ over a small neighborhood of 0:

$$\begin{aligned}\widehat{W}^{\boldsymbol{v}}_{i,j,s}(t) \ge\ & 2m(t)d_{i,j}D^2_{i,j,s}\tan^2(\psi(t)) \\ & + 4tD_{i,j,s}\sin(\theta)D(t)M(t)\tan(\psi(t)) \\ & - 2tD^2_{i,j,s}\cos(\theta)M(t)\tan^2(\psi(t));\end{aligned}$$

and

$$\begin{aligned}\widehat{W}^{\boldsymbol{v}}_{i,j,s}(t) \le\ & 2M(t)d_{i,j}D^2_{i,j,s}\tan^2(\psi(t)) \\ & + 4tD_{i,j,s}\sin(\theta)D(t)m(t)\tan(\psi(t)) \\ & - 2tD^2_{i,j,s}\cos(\theta)m(t)\tan^2(\psi(t)).\end{aligned}$$

Proposition 4 follows immediately from the limit properties of $m(t)$, $M(t)$, $D(t)$ and $\tanh(\psi(t))$.

Proof of Lemma 7. We use the same notations and coordinate system as before. Observe that in equation (28), the maximum that defines $D(t)$ must be attained at a point in $\widetilde{\Delta}^{\boldsymbol{v}}_{i\to j}(t) \cap \mathbb{S}_s$, where $\mathbb{S}_s$ is the hypersphere of the ball $\mathbb{B}_s$; see the right panel of Figure 5. We claim that the maximum must also be attained by a point in the hyperplane $\mathcal{L}_{i,j,s}$. Indeed, recalling the representation in equation (24), we see that each point $\boldsymbol{x} \in \widetilde{\Delta}^{\boldsymbol{v}}_{i\to j}(t) \cap \mathbb{S}_s$ must satisfy

$$(x_1 - b^*_1)^2 + (x_2 - b^*_2)^2 + \sum_{j\ge 3} x_j^2 = r^2 \quad \text{and}$$

$$x_1 \in \big[0, D_{i,j,s}\tan(\psi(t))\big]$$

From the above equation it is clear that for each fixed $z \in [0, D_{i,j,s}\tan(\psi(t))]$, over the set $\mathbb{S}_s \cap \widetilde{\Delta}^{\boldsymbol{v}}_{i\to j}(t) \cap \{\boldsymbol{x} : x_1 = z\}$, the maximum of x_2 is attained exactly when $x_j = 0$ for all $j \ge 3$, that is, $\boldsymbol{x} \in \mathcal{L}_{i,j,s}$. Combining these observations, we conclude that

$$D(t) = \max\big\{x_2 : \boldsymbol{x} \in \widetilde{\Delta}^{\boldsymbol{v}}_{i\to j}(t) \cap \mathbb{S}_s \cap \mathcal{L}_{i,j,s}\big\}.$$

Note that the set $\widetilde{\Delta}^{\boldsymbol{v}}_{i\to j}(t) \cap \mathbb{S}_s \cap \mathcal{L}_{i,j,s}$ is compact, which is represented by the solid blue segment in $\mathbb{B}_s$ in the right panel of Figure 5; as $t \to 0$, this set shrinks continuously to a single point $(0, -D_{i,j,s}, \ldots, 0)$. It thus follows that $\lim_{t\to 0} D(t) = -D_{i,j,s}$. This completes the proof of Lemma 7. □

F. Proofs of Lemmas 5 and 6

For convenience we restate the bounds in equations (20) and (21) as follows:

$$\begin{cases} m_{i,s} \le k\lambda r, & \text{if } s \in B_i, \\ m_{i,s} \ge 1 - k^2\lambda r, & \text{if } s \in A_i, \end{cases} \tag{29}$$

where we recall that $m_{i,s} = \mathbb{P}(\mathcal{V}_i)$ is the probability mass of the set $\mathcal{V}_i$ with respect to the uniform distribution on the ball $\mathbb{B}_s \equiv \mathbb{B}_{\boldsymbol{\beta}^*_s}(r)$, and $\boldsymbol{c}_s$ is the corresponding center of mass. Applying the simple geometric result in Lemma 10, we further obtain the bound

$$\|\boldsymbol{\beta}^*_s - \boldsymbol{c}_{i,s}\| \le \frac{r\cdot(1-m_{i,s})}{m_{i,s}} \le \begin{cases} \frac{r}{m_{i,s}}, & \text{if } s \in B_i, \\ \frac{k^2\lambda r^2}{m_{i,s}}, & \text{if } s \in A_i, \end{cases} \tag{30}$$

where the last step follows from the fact that $m_{i,s} \le 1$ and equation (29). We are ready to prove the two lemmas.

Proof of Lemma 5. We have the following decomposition of the vector $\boldsymbol{\mu}$:

$$\begin{aligned}\boldsymbol{\mu} =& \frac{\sum_{s\in A_i} m_{i,s}\boldsymbol{c}_{i,s} + \sum_{s\in B_i} m_{i,s}\boldsymbol{c}_{i,s}}{\sum_{s\in A_i} m_{i,s} + \sum_{s\in B_i} m_{i,s}} - \frac{\sum_{s\in A_i} m_{i,s}\boldsymbol{c}_{i,s}}{\sum_{s\in A_i} m_{i,s}} \\ =& \frac{\sum_{s\in B_i} m_{i,s}\boldsymbol{c}_{i,s}}{\sum_{s\in A_i} m_{i,s} + \sum_{s\in B_i} m_{i,s}} \\ & - \frac{\sum_{s\in B_i} m_{i,s}\sum_{s\in A_i} m_{i,s}\boldsymbol{c}_{i,s}}{(\sum_{s\in A_i} m_{i,s} + \sum_{s\in B_i} m_{i,s})\sum_{s\in A_i} m_{i,s}} \\ =& \frac{\sum_{s\in B_i} m_{i,s}(\boldsymbol{c}_{i,s} - \boldsymbol{\beta}^*_s)}{\sum_{s\in A_i} m_{i,s} + \sum_{s\in B_i} m_{i,s}} \\ & - \frac{\sum_{s\in B_i} m_{i,s}\sum_{s\in A_i} m_{i,s}(\boldsymbol{c}_{i,s} - \boldsymbol{\beta}^*_s)}{(\sum_{s\in A_i} m_{i,s} + \sum_{s\in B_i} m_{i,s})\sum_{s\in A_i} m_{i,s}} \\ & + \frac{\sum_{s\in B_i} m_{i,s}\boldsymbol{\beta}^*_s}{\sum_{s\in A_i} m_{i,s} + \sum_{s\in B_i} m_{i,s}} \\ & - \frac{\sum_{s\in B_i} m_{i,s}\sum_{s\in A_i} m_{i,s}\boldsymbol{\beta}^*_s}{(\sum_{s\in A_i} m_{i,s} + \sum_{s\in B_i} m_{i,s})\sum_{s\in A_i} m_{i,s}}.\end{aligned}$$

It follows that

$$\begin{aligned}\|\boldsymbol{\mu}\| \le& \frac{\sum_{s\in B_i} m_{i,s}\|\boldsymbol{c}_{i,s} - \boldsymbol{\beta}^*_s\|}{\sum_{s\in A_i} m_{i,s} + \sum_{s\in B_i} m_{i,s}} \\ & + \frac{\sum_{s\in B_i} m_{i,s}\sum_{s\in A_i} m_{i,s}\|\boldsymbol{c}_{i,s} - \boldsymbol{\beta}^*_s\|}{(\sum_{s\in A_i} m_{i,s} + \sum_{s\in B_i} m_{i,s})\sum_{s\in A_i} m_{i,s}}\end{aligned}$$

$$+\frac{2\sum_{s\in B_i} m_{i,s}}{\sum_{s\in A_i} m_{i,s}+\sum_{s\in B_i} m_{i,s}}\Delta_{\max}.$$
$$\overset{\text{(i)}}{\leq}\frac{k^* r}{1-k^2\lambda r}+\frac{k^*\cdot(k\lambda r)\cdot(k^2\lambda r^2)}{(1-k^2\lambda r)^2}+\frac{2k^* k\lambda r}{1-k^2\lambda r}\Delta_{\max},$$

where the last step holds due to the inequalities (29) and (30) as well as the fact that $|A_i|, |B_i| \in [1, k^*]$. This completes the proof of Lemma 5. □

Proof of Lemma 6. We have the following decomposition of the vector $\boldsymbol{\nu}$:

$$\begin{aligned}\boldsymbol{\nu} =& \frac{\sum_{s\in A_i} m_{i,s}\boldsymbol{c}_{i,s}}{\sum_{s\in A_i} m_{i,s}} - \frac{1}{|A_i|}\sum_{s\in A_i}\boldsymbol{\beta}_s^* \\ =& \frac{\sum_{s\in A_i} m_{i,s}(\boldsymbol{c}_{i,s}-\boldsymbol{\beta}_s^*)}{\sum_{s\in A_i} m_{i,s}} \\ &+ \frac{\sum_{s\in A_i}\left(m_{i,s}-\frac{1}{|A_i|}\sum_{s'\in A_i} m_{i,s'}\right)\boldsymbol{\beta}_s^*}{\sum_{s\in A_i} m_{i,s}}.\end{aligned}$$

Using the inequalities (29) and (30) as well as the triangle inequality, we obtain that

$$\|\boldsymbol{\nu}\| \leq \frac{k^2\lambda r^2}{1-k^2\lambda r}+\frac{k^2\lambda r}{1-k^2\lambda r}\Delta_{\max},$$

thereby proving Lemma 6. □

VII. Conclusion

In this paper, we characterize the structures of all local minima in the k-means problem for general values of k. We show that under an appropriate separation condition of the ground truth clusters, the local minima are always composed of one-fit-many, many-fit-one or almost-empty type associations between the fitted and ground truth centers.

Several future directions are of interests for both theory and applications. An immediate direction is to generalize our results from the population case to the finite sample case, and from balanced spherical GMMs to more general mixture models with imbalanced clusters, general covariance matrices and heavy-tailed distributions. To transfer the population case to the finite sample case, we might be able to adapt a set of general techniques based on localization and uniform concentration arguments; see [44, 54]. Also, while we have focused on the k-means formulation, we expect that similar structural results hold for a much broader class of clustering formulations, including the maximum likelihood formulation of mixture problems. On the computational side, we have discussed the implications of our results for improving clustering algorithms. Rigorously justifying these algorithms (which are largely heuristic so far) in a broad range of models would be interesting.

Finally, it would be of great interest to establish similar structural results for other non-convex optimization problems that arise in machine learning and statistics applications [55, 46, 47, 45, 44].

Appendix

Equivalence to the partition-based formulation

A common way of formulating the k-means clustering problem is as follows: given a set of observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, we find a partition $\boldsymbol{S} = \{S_1, \ldots, S_k\}$ of these observations such that the within-cluster sum of squared distances is minimized:

$$\min_{\boldsymbol{S}} \sum_{j=1}^{k}\sum_{\boldsymbol{x}\in S_j}\|\boldsymbol{x}-\boldsymbol{\mu}_j\|^2, \tag{31}$$

where $\boldsymbol{\mu}_j = \frac{1}{|S_j|}\sum_{\boldsymbol{x}\in S_j}\boldsymbol{x}$ is the mean of points in cluster i. Meanwhile, the formulation (1) used in this paper, restated below,

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}\min_{j\in[k]}\|\boldsymbol{x}_i-\boldsymbol{\beta}_j\|^2,$$

is based on optimizing over the centers $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$. Note that each solution $\boldsymbol{\beta}$ induces a partition $\{S_j(\boldsymbol{\beta})\}$ of the observations via the Voronoi diagram. Also note that the sum of squared distances to a set of points is minimized by the mean of these points. Combining this observations, we obtain that for any $\boldsymbol{\beta}$:

$$\begin{aligned}\min_{\boldsymbol{S}} \sum_{j=1}^{k}\sum_{\boldsymbol{x}\in S_j}\|\boldsymbol{x}-\boldsymbol{\mu}_j\|^2 \leq& \sum_{j=1}^{k}\sum_{\boldsymbol{x}\in S_j(\boldsymbol{\beta})}\|\boldsymbol{x}-\boldsymbol{\beta}_j\|^2 \\ =& \sum_{i=1}^{n}\min_{j\in[k]}\|\boldsymbol{x}_i-\boldsymbol{\beta}_j\|^2.\end{aligned} \tag{32}$$

On the other hand, for any partition $\boldsymbol{S} = \{S_1, \ldots, S_k\}$ of the data points and its corresponding means $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)$, we have

$$\begin{aligned}\sum_{j=1}^{k}\sum_{\boldsymbol{x}\in S_j}\|\boldsymbol{x}-\boldsymbol{\mu}_j\|^2 \geq& \sum_{i=1}^{n}\min_{j\in[k]}\|\boldsymbol{x}_i-\boldsymbol{\mu}_j\|^2 \\ \geq& \min_{\boldsymbol{\beta}}\sum_{i=1}^{n}\min_{j\in[k]}\|\boldsymbol{x}_i-\boldsymbol{\beta}_j\|^2.\end{aligned} \tag{33}$$

Taking the minimum over $\boldsymbol{\beta}$ of both sides of equation (32), and the minimum over $\boldsymbol{S}$ for equation (33), we conclude that the two formulations (31) and (1) have the same optimal values. Moreover, an optimal solution for one formulation induces an optimal solution for the other. Hence these two formulations are equivalent.

Proof of Proposition 1

In this section we prove Proposition 1, which states that under the Stochastic Ball Model with $k^* = k$, the ground truth centers $\boldsymbol{\beta}^*$ is the global minimum of the k-means objective function G.

Proof. We begin by upper bounding the objective value of the ground truth:

$$\begin{aligned}G(\boldsymbol{\beta}^*) =& \frac{1}{k}\sum_{s\in[k]}\int\min_{i\in[k]}\|\boldsymbol{x}-\boldsymbol{\beta}_i^*\|^2 f_s(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ \leq& \frac{1}{k}\sum_{s\in[k]}\int\|\boldsymbol{x}-\boldsymbol{\beta}_s^*\|^2 f_s(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ \leq& r^2,\end{aligned}$$

where the second inequality follows from the fact that each true cluster $\mathbb{B}_s$ has radius r.

Now let $\boldsymbol{\beta}$ be a global minimum of G. By optimality of $\boldsymbol{\beta}$, we have for each $s \in [k]$:

$$\begin{aligned} r^2 &\geq G(\boldsymbol{\beta}) \\ &\overset{\text{(i)}}{\geq} \frac{1}{k} \int \min_{i\in[k]} \|\boldsymbol{x} - \boldsymbol{\beta}_i\|^2 f_s(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &\overset{\text{(ii)}}{\geq} \frac{1}{k} \int \min_{i\in[k]} \left(\frac{1}{2}\|\boldsymbol{\beta}_s^* - \boldsymbol{\beta}_i\|^2 - \|\boldsymbol{x} - \boldsymbol{\beta}_s^*\|^2 \right) f_s(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &\overset{\text{(iii)}}{\geq} \frac{1}{2k} \min_{i\in[k]} \|\boldsymbol{\beta}_s^* - \boldsymbol{\beta}_i\|^2 - \frac{r^2}{k}. \end{aligned}$$

where step (i) holds by ignoring $k-1$ clusters, step (ii) holds by the inequality $(a-b)^2 \geq \frac{1}{2}a^2 - b^2$, and step (iii) holds because f_s is a probability density. From the above equation we obtain that

$$\min_{i\in[k]} \|\boldsymbol{\beta}_s^* - \boldsymbol{\beta}_i\| < 2\sqrt{k}r, \qquad \forall s \in [k];$$

that is, each true center $\boldsymbol{\beta}_s^*$ is $2\sqrt{k}r$-close to at least one $\boldsymbol{\beta}_i$. We further observe that each $\boldsymbol{\beta}_i$ is $2\sqrt{k}r$-close to at most one true center $\boldsymbol{\beta}_s^*$; otherwise, by the triangle inequality we woud have $\Delta_{\min} \leq \|\boldsymbol{\beta}_s^* - \boldsymbol{\beta}_{s'}^*\| \leq \|\boldsymbol{\beta}_s^* - \boldsymbol{\beta}_i\| + \|\boldsymbol{\beta}_{s'}^* - \boldsymbol{\beta}_i\| < 4\sqrt{k}r$, contradicting the SNR assumption $\eta_{\min} := \frac{\Delta_{\min}}{r} \geq 6\sqrt{k}$. Since the number of $\boldsymbol{\beta}_i$'s is equal to that of $\boldsymbol{\beta}_s^*$'s, we deduce that each $\boldsymbol{\beta}_i$ is $2\sqrt{k}r$-close to *exactly one* $\boldsymbol{\beta}_s^*$. Without loss of generality, we may assume that

$$\|\boldsymbol{\beta}_s^* - \boldsymbol{\beta}_s\| < 2\sqrt{k}r, \qquad \forall s \in [k].$$

When the above inequality and the SNR assumption $\eta_{\min} := \frac{\Delta_{\min}}{r} \geq 6\sqrt{k}$ hold, we have for each pairs $(s, s') \in [k] \times [k]$ with $s \neq s'$ and each $\boldsymbol{x} \in \mathbb{B}_s$:

$$\begin{aligned} \|\boldsymbol{x} - \boldsymbol{\beta}_s\| &\leq \|\boldsymbol{x} - \boldsymbol{\beta}_s^*\| + \|\boldsymbol{\beta}_s^* - \boldsymbol{\beta}_s\| \\ &< r + 2\sqrt{k}r \\ &\leq 6\sqrt{k}r - r - 2\sqrt{k}r \\ &< \|\boldsymbol{\beta}_s^* - \boldsymbol{\beta}_{s'}^*\| - \|\boldsymbol{x} - \boldsymbol{\beta}_s^*\| - \|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^*\| \\ &\leq \|\boldsymbol{x} - \boldsymbol{\beta}_{s'}\|, \end{aligned}$$

which implies that $\mathbb{B}_s \subseteq \mathcal{V}_s(\boldsymbol{\beta})$. Applying Lemma 2 to the global minimum $\boldsymbol{\beta}$, we obtain that

$$\boldsymbol{\beta}_s = \frac{\int_{\mathcal{V}_s(\boldsymbol{\beta})} \boldsymbol{x} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}}{\int_{\mathcal{V}_s(\boldsymbol{\beta})} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}} = \frac{\int_{\mathbb{B}_s} \boldsymbol{x} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}}{\int_{\mathbb{B}_s} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}} = \boldsymbol{\beta}_s^*, \qquad \forall s \in [k].$$

thereby proving that $\boldsymbol{\beta}^*$ is the only global minimum. □

Proof of Proposition 2

In this section we prove Proposition 2, which states that under the one-dimensional Stochastic Ball Model in Figure 2 with $r < 0.4$, the solution $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3) = (-2 - \frac{r}{2}, -2 + \frac{r}{2}, 1)$ is a local minimum of the k-means objective function G.

Proof. Observe that $\mathcal{V}_1(\boldsymbol{\beta}) = (-\infty, -2]$, $\mathcal{V}_2(\boldsymbol{\beta}) = [-2, \frac{-1+r/2}{2}]$, $\mathcal{V}_3(\boldsymbol{\beta}) = [\frac{-1+r/2}{2}, \infty]$, $\partial_{1,2}(\boldsymbol{\beta}) = -2$ and $\partial_{2,3}(\boldsymbol{\beta}) = \frac{-1+r/2}{2}$. When $r < 0.4$, it is easy to see that for any $\boldsymbol{b} = (b_1, b_2, b_3) \in \mathbb{R}^3$ in a small neighborhood of $\boldsymbol{\beta}$, the Voronoi boundary $\partial_{2,3}(\boldsymbol{b})$ remains strictly between $-2+r$ and $-r$, and $\partial_{1,2}(\boldsymbol{b})$ remains strictly between $-2-r$ and $-2+r$. Therefore, for any such $\boldsymbol{b}$ we can explicitly write down its objective value:

$$\begin{aligned} G(\boldsymbol{b}) = &\int_{-2-r}^{\frac{b_1+b_2}{2}} (x-b_1)^2 \mathrm{d}x + \int_{\frac{b_1+b_2}{2}}^{-2+r} (x-b_2)^2 \mathrm{d}x \\ &+ \int_{-r}^{r} (x-b_3)^2 \mathrm{d}x + \int_{2-r}^{2+r} (x-b_3)^2 \mathrm{d}x. \end{aligned}$$

We compute the derivative and Hessian for G at $\boldsymbol{b}$:

$$\nabla_{\boldsymbol{b}} G = \begin{bmatrix} -2\int_{-2-r}^{\frac{b_1+b_2}{2}} (x-b_1)\mathrm{d}x \\ -2\int_{\frac{b_1+b_2}{2}}^{-2+r} (x-b_2)\mathrm{d}x \\ -2\int_{-r}^{r}(x-b_3)\mathrm{d}x - 2\int_{2-r}^{2+r}(x-b_3)\mathrm{d}x \end{bmatrix}$$

and

$$\nabla_{\boldsymbol{b}}^2 G = \begin{bmatrix} \frac{b_1-b_2}{2} + 2(\frac{b_1+b_2+2+r}{2}) & \frac{b_1-b_2}{2} & 0 \\ \frac{b_1-b_2}{2} & \frac{b_1-b_2}{2} + 2(-2+r-\frac{b_1+b_2}{2}) & 0 \\ 0 & 0 & 8r \end{bmatrix}.$$

Evaluating the above expressions at $b_1 = \beta_1 = 2 = \frac{r}{2}$, $b_2 = \beta_2 = -2 + \frac{r}{2}$ and $b_3 = \beta_3 = 1$, we find that the derivative vanishes and the Hessian is positive definite:

$$\begin{aligned} &\nabla_{\boldsymbol{b}} G\big|_{\boldsymbol{b}=\boldsymbol{\beta}} = 0, \\ &\nabla_{\boldsymbol{b}}^2 G\big|_{\boldsymbol{b}=\boldsymbol{\beta}} = \begin{bmatrix} 1.5r & -0.5r & 0 \\ -0.5r & 1.5r & 0 \\ 0 & 0 & 8r \end{bmatrix} \succ 0. \end{aligned}$$

Therefore, $\boldsymbol{\beta}$ is indeed a local minimum of G. □

Proofs for Section V

In this section, we prove the technical lemmas stated in Section V.

Proof of Lemma 1

Proof. Our goal is to derive the existence and expression of the derivative of the function

$$H^{\boldsymbol{v}}(t) := G(\boldsymbol{\beta} + t\boldsymbol{v}) = \int_{\boldsymbol{x}} \min_{i\in[k]} \|\boldsymbol{x} - \boldsymbol{\beta}_i - t\boldsymbol{v}_i\|^2 f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

at $t = 0$. We make use of the following measure-theoretic version of the Leibniz integral rule.

Proposition 5 (Leibniz's integral rule)**.** *Let T be an open subset of $\mathbb{R}$, and X be a measure space. Suppose $g : T \times X \to \mathbb{R}$ satisfies the following conditions: (i) $g(t, \boldsymbol{x})$ is a Lebesgue-integrable function of $\boldsymbol{x}$ for each $t \in T$; (ii) for almost all $\boldsymbol{x} \in X$, the partial derivative $\frac{\partial}{\partial t} g(t, \boldsymbol{x})$ exists for all $t \in T$; (iii) There is an integrable function $\theta : X \to \mathbb{R}$ such that $|\frac{\partial}{\partial t} g(t, \boldsymbol{x})| \leq \theta(\boldsymbol{x})$ for all $t \in T$ and almost every $\boldsymbol{x} \in X$, then we have*

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_X g(t, \boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_X \frac{\partial}{\partial t} g(t, \boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

We verify the above three conditions in the proposition for $H^{\boldsymbol{v}}$. Without loss of generality, assume that $\|\boldsymbol{v}_i\| \leq 1, \forall i \in [k]$. Let $\Delta := \min_{i\neq j} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|$, which satisfies $\Delta > 0$ by the assumption that $\{\boldsymbol{\beta}_j\}_{j=1}^k$ are pairwise distinct. For condition (i), we see that the function $g(t, \boldsymbol{x}) := \min_{i\in[k]} \|\boldsymbol{x} - \boldsymbol{\beta}_i - t\boldsymbol{v}_i\|^2 f(\boldsymbol{x})$ is integrable in $\boldsymbol{x}$ for each bounded t, since the

density f has bounded second moment. For condition (ii), note that when $t \in T := [-\frac{\Delta}{4}, \frac{\Delta}{4}]$, the perturbed solution $\boldsymbol{\beta} + t\boldsymbol{v}$ remains pairwise disjoint, hence the Voronoi boundary $\partial(\boldsymbol{\beta}+t\boldsymbol{v})$ has measure 0. For all $t \in T$ and all $\boldsymbol{x} \notin \partial(\boldsymbol{\beta}+t\boldsymbol{v})$, the minimizer in the definition of $g(t+\epsilon, \boldsymbol{x})$ remains fixed when $|\epsilon|$ is sufficiently small, hence the partial derivative $\frac{\partial}{\partial t} g(t, \boldsymbol{x})$ exists at all $t \in T$ and satisfies

$$\begin{aligned} &\boldsymbol{x} \in \mathcal{V}_i(\boldsymbol{\beta} + t\boldsymbol{v}) \\ \implies &\frac{\partial}{\partial t} g(t, \boldsymbol{x}) = -2\langle \boldsymbol{v}_i, \boldsymbol{x} - \boldsymbol{\beta}_i - t\boldsymbol{v}_i \rangle f(\boldsymbol{x}). \end{aligned} \tag{34}$$

Finally for condition (iii), for each $\boldsymbol{x} \in \text{support}(f) = \cup_{s\in[k]} \mathbb{B}_s(r)$, we have the bound $|\langle \boldsymbol{v}_i, \boldsymbol{x} - \boldsymbol{\beta}_i - t\boldsymbol{v}_i\rangle| \le \max_{s\in[k]} \|\boldsymbol{\beta}_s^*\| + r + \|\boldsymbol{\beta}_i\| + \frac{\Delta}{4}$ when $t \in T$, hence $|\frac{\partial}{\partial t} g(t, \boldsymbol{x})|$ is bounded by an integrable function. Applying the Leibniz's integral rule and equation (34), we obtain that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} H^{\boldsymbol{v}}(0) &= \int_{\boldsymbol{x}} \frac{\partial}{\partial t} g(0, \boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= -\sum_{i=1}^{k} \int_{\mathcal{V}_i(\boldsymbol{\beta})} 2\langle \boldsymbol{v}_i, \boldsymbol{x} - \boldsymbol{\beta}_i \rangle f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \end{aligned}$$

as claimed. □

Proof of Lemma 2

Proof. In view of the decomposition (10) and Remark 1, we have the following upper bound for $H^{\boldsymbol{v}}$:

$$H^{\boldsymbol{v}}(t) \le h^{\boldsymbol{v}}(t) := \sum_{i=1}^{k} \int_{\mathcal{V}_i(\boldsymbol{\beta})} \|\boldsymbol{x} - \boldsymbol{\beta}_i - t\boldsymbol{v}_i\|^2 f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{35}$$

which satisfies $h^{\boldsymbol{v}}(0) = H^{\boldsymbol{v}}(0)$. Since $\boldsymbol{\beta}$ is a local minimum of G, we know that $t = 0$ is local minimum of $H^{\boldsymbol{v}}$ for all $\boldsymbol{v}$, hence Lemma 3 ensures that $t = 0$ is also a local minimum of $h^{\boldsymbol{v}}$.

Suppose that we have $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ and $\mathcal{V}_1(\boldsymbol{\beta}) = \mathcal{V}_2(\boldsymbol{\beta})$ has a positive measure with respect to f, and that all other $\boldsymbol{\beta}_j, j \ge 3$ are pairwise distinct and different from $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. We may partition $\mathcal{V}_1(\boldsymbol{\beta}) = \mathcal{V}_2(\boldsymbol{\beta})$ into two disjoint sets S_1 and S_2, each with positive measure. For $i \in \{1, 2\}$ denote by $\boldsymbol{s}_i := \frac{\int_{S_i} \boldsymbol{x} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}}{\int_{S_i} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}}$ the center of mass of S_i with respect to f. We can choose the partition in such a way that $\boldsymbol{s}_1 \neq \boldsymbol{\beta}_1$ and $\boldsymbol{s}_2 \neq \boldsymbol{\beta}_2$. Fix a direction $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{0}, \ldots, \boldsymbol{0})$ with $\boldsymbol{v}_1 = \boldsymbol{s}_1 - \boldsymbol{\beta}_1$ and $\boldsymbol{v}_2 = \boldsymbol{s}_2 - \boldsymbol{\beta}_2$. In this case the upper bound $h^{\boldsymbol{v}}$ can be written as

$$\begin{aligned} h^{\boldsymbol{v}}(t) &= \int_{S_1} \|\boldsymbol{x} - \boldsymbol{\beta}_1 - t\boldsymbol{v}_1\|^2 f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &\quad + \int_{S_2} \|\boldsymbol{x} - \boldsymbol{\beta}_2 - t\boldsymbol{v}_2\|^2 f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \text{constant} \\ &= \int_{S_1} \|\boldsymbol{x} - \boldsymbol{s}_1\|^2 f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_{S_1} \|\boldsymbol{s}_1 - \boldsymbol{\beta}_1 - t\boldsymbol{v}_1\|^2 f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &\quad + \int_{S_2} \|\boldsymbol{x} - \boldsymbol{s}_2\|^2 f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_{S_2} \|\boldsymbol{s}_2 - \boldsymbol{\beta}_2 - t\boldsymbol{v}_2\|^2 f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &\quad + \text{constant}. \end{aligned}$$

(In the calculation above we have avoided double counting the contribution from $\mathcal{V}_1(\boldsymbol{\beta}) = \mathcal{V}_2(\boldsymbol{\beta})$, and the constant part corresponds to the contribution from $\mathcal{V}_i(\boldsymbol{\beta})$ for $i \ge 3$) With the above choices of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, we see that $h^{\boldsymbol{v}}(0) > h^{\boldsymbol{v}}(t)$ for all $t \in (0, 1)$ and hence $t = 0$ is not a local minimum of $h^{\boldsymbol{v}}$, which is a contradiction. Therefore, we must have $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$ whenever $\mathcal{V}_i(\boldsymbol{\beta}) \cup \mathcal{V}_j(\boldsymbol{\beta})$ has a positive measure. The more general statement in Lemma 2 can be established in a similar manner.

Now suppose that $\boldsymbol{\beta}_i$ has a Voronoi set $\mathcal{V}_i(\boldsymbol{\beta})$ with a positive measure. In this case the center of mass $\boldsymbol{c}_i := \frac{\int_{\mathcal{V}_i(\boldsymbol{\beta})} \boldsymbol{x} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}}{\int_{\mathcal{V}_i(\boldsymbol{\beta})} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}}$ is well-defined. Choose the direction $\boldsymbol{v} = (\boldsymbol{0}, \ldots, \boldsymbol{0}, \boldsymbol{c}_i - \boldsymbol{\beta}_i, \boldsymbol{0}, \ldots, \boldsymbol{0})$. Since $t = 0$ is a local minimum of $H^{\boldsymbol{v}}$, its derivative must vanish at $t = 0$. Using the derivative expression from Lemma 1,[4] we obtain that

$$\begin{aligned} 0 = \frac{\mathrm{d}}{\mathrm{d}t} H^{\boldsymbol{v}}(0) &= -\int_{\mathcal{V}_i(\boldsymbol{\beta})} 2\langle \boldsymbol{v}_i, \boldsymbol{x} - \boldsymbol{\beta}_i \rangle f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= -2\langle \boldsymbol{v}_i, \boldsymbol{c}_i - \boldsymbol{\beta}_i \rangle \int_{\mathcal{V}_i(\boldsymbol{\beta})} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= -2\|\boldsymbol{c}_i - \boldsymbol{\beta}_i\|^2 \int_{\mathcal{V}_i(\boldsymbol{\beta})} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \end{aligned}$$

where the last step follows from our choice of $\boldsymbol{v}$. Since $\int_{\mathcal{V}_i(\boldsymbol{\beta})} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ is the measure of $\mathcal{V}_i(\boldsymbol{\beta})$ and positive, we must have $\boldsymbol{\beta}_i = \boldsymbol{c}_i$ as claimed. □

We state and prove several technical lemmas that are used in Section VI.

Proof of Lemma 4

Recall that $m_{i,s}$ and $\boldsymbol{c}_{i,s}$ denote the mass and the center of mass of the set $\mathcal{V}_i$ with respect to the density f_s. We similarly define

$$\begin{aligned} \widetilde{m_{i,s}} &= \sum_{s' \in T_i : s' \neq s} m_{i,s'} \qquad \text{and} \\ \widetilde{\boldsymbol{c}_{i,s}} &= \frac{\sum_{s' \in T_i : s' \neq s} m_{i,s'} \boldsymbol{c}_{i,s'}}{\sum_{s' \in T_i : s' \neq s} m_{i,s'}}, \end{aligned}$$

which are the mass and the center of mass of the set $\mathcal{V}_i$ with respect to the density $\sum_{s' \neq s} f_{s'}$. With this notation, the local minimum $\boldsymbol{\beta}$ must satisfy the necessary condition

$$\boldsymbol{\beta}_i = \frac{m_{i,s} \boldsymbol{c}_{i,s} + \widetilde{m_{i,s}} \widetilde{\boldsymbol{c}_{i,s}}}{m_{i,s} + \widetilde{m_{i,s}}}, \tag{36}$$

which follows from Lemma 2 and the text thereafter. Rearranging the expression (36) gives

$$\widetilde{\boldsymbol{c}_{i,s}} = \frac{(\widetilde{m_{i,s}} + m_{i,s})\boldsymbol{\beta}_i - m_{i,s} \boldsymbol{c}_{i,s}}{\widetilde{m_{i,s}}}.$$

It then follows from the triangle inequality that

$$\begin{aligned} \|\widetilde{\boldsymbol{c}_{i,s}} - \boldsymbol{\beta}_i\| &= \frac{m_{i,s}}{\widetilde{m_{i,s}}} \|\boldsymbol{c}_{i,s} - \boldsymbol{\beta}_i\| \\ &\le \frac{m_{i,s}}{\widetilde{m_{i,s}}} (\|\boldsymbol{c}_{i,s} - \boldsymbol{\beta}_s^*\| + \|\boldsymbol{\beta}_s^* - \boldsymbol{\beta}_i\|). \end{aligned} \tag{37}$$

[4]Lemma 1 is applicable for the following reason: we can ignore those $\mathcal{V}_i(\boldsymbol{\beta})$'s with zero measure in the integrals defining G and $H^{\boldsymbol{v}}$, in which case we have just established that $\boldsymbol{\beta}$ must have pairwise distinct components and thus satisfy the premise of Lemma 1.

We are now ready to prove Lemma 4, whose assumption states that $\rho_s(\partial_{j,\ell}) > \lambda = \frac{c}{\sqrt{r\Delta_{\max}}}$ for some $(s,j,\ell) \in T_i \times [k] \times [k]$. Observation 1 ensures that such an s is unique, hence for all other $s' \in T_i \setminus \{s\}$, we must have $\rho_{s'}(\partial_{j,\ell}) \le \lambda, \forall (j,\ell)$. Observation 2 ensures that for all these s', if $\boldsymbol{\beta}^*_{s'} \in \mathcal{V}_i$ then $s' \in A_i$. In view of these properties and equation (36), we can see that $\widetilde{\boldsymbol{c}_{i,s}}$ is similar to $\boldsymbol{\beta}_i$ except that the density of the s-th true cluster is ignored. Therefore, we can follow the same arguments for proving Part 2(b) of Theorem 4 as well as the simplification step in (12) to obtain that

$$\|\widetilde{\boldsymbol{c}_{i,s}} - \boldsymbol{b}_i^-\| \le \Delta_{\max} \frac{(4k^* + 2ck^2)(1 + \frac{2k^*}{k})}{\sqrt{\eta_{\max}}} \tag{38}$$

On the other hand, under the assumption of the lemma, Part 1 of Theorem 4 ensures that

$$\|\boldsymbol{\beta}_i - \boldsymbol{\beta}^*_s\| \le \frac{k}{\lambda} + 3r \le \Delta_{\max} \frac{4k^*}{c\sqrt{\eta_{\max}}}. \tag{39}$$

We proceed by bounding $\|\boldsymbol{b}_i^- - \boldsymbol{\beta}^*_s\|$ as follows:

$$\begin{aligned}
\|\boldsymbol{b}_i^- - \boldsymbol{\beta}^*_s\| \le& \|\boldsymbol{b}_i^- - \widetilde{\boldsymbol{c}_{i,s}}\| + \|\widetilde{\boldsymbol{c}_{i,s}} - \boldsymbol{\beta}_i\| + \|\boldsymbol{\beta}_i - \boldsymbol{\beta}^*_s\| \\
\le& \|\boldsymbol{b}_i^- - \widetilde{\boldsymbol{c}_{i,s}}\| + \frac{m_{i,s}}{\widetilde{m_{i,s}}}(\|\boldsymbol{c}_{i,s} - \boldsymbol{\beta}^*_s\| + \|\boldsymbol{\beta}^*_s - \boldsymbol{\beta}_i\|) \\
&+ \|\boldsymbol{\beta}_i - \boldsymbol{\beta}^*_s\| \qquad (40) \\
\le& \Delta_{\max} \frac{(4k^* + 2ck^2)(1 + \frac{2k^*}{k})}{\sqrt{\eta_{\max}}} \\
&+ \frac{m_{i,s}}{\widetilde{m_{i,s}}} \cdot \frac{\Delta_{\max}}{\eta_{\max}} + \left(\frac{m_{i,s}}{\widetilde{m_{i,s}}} + 1\right) \Delta_{\max} \frac{4k^*}{c\sqrt{\eta_{\max}}}, \qquad (41)
\end{aligned}$$

where step (40) follows from the bound (37), and step (41) follows from the bounds (38) and (39) as well as the fact that $\boldsymbol{c}_{i,s} \in \mathbb{B}_s$ so $\|\boldsymbol{c}_{i,s} - \boldsymbol{\beta}^*_s\| \le r = \frac{\Delta_{\max}}{\eta_{\max}}$. Now, note that since $|A_i \setminus \{s\}| \ge 2$ by assumption, there exists some $s' \in A_i \subseteq T_i$ such that $s' \neq s$. We have established above that this s' must satisfy $\rho_{s'}(\partial_{j,\ell}) \le \lambda, \forall (j,\ell)$, hence applying Part 2 of Theorem 4 we obtain that $m_{i,s'} = \mathbb{P}_{s'}(\mathcal{V}_i) \ge 0.5$, which further implies $\frac{m_{i,s}}{\widetilde{m_{i,s}}} \le 2$. Continuing from the above display equation, we obtain

$$\begin{aligned}
\|\boldsymbol{b}_i^- - \boldsymbol{\beta}^*_s\| \le& \Delta_{\max} \frac{(4k^* + 2ck^2)(1 + \frac{2k^*}{k})}{\sqrt{\eta_{\max}}} \\
&+ 3\Delta_{\max} \frac{4k^*}{c\sqrt{\eta_{\max}}} + \Delta_{\max} \frac{2}{\eta_{\max}} \\
\le& \Delta_{\max} \frac{2ck^2(1 + \frac{2k^*}{k}) + 4k^*(2.5 + \frac{2k^*}{k})}{\sqrt{\eta_{\max}}}, \qquad (42)
\end{aligned}$$

where the last step follows from the assumption that $c > 3$ and $\eta_{\max} \ge 4ck^2$. Combining the inequalities (39) and (42), we obtain

$$\begin{aligned}
\|\boldsymbol{b}_i^- - \boldsymbol{\beta}_i\| \le& \|\boldsymbol{b}_i^- - \boldsymbol{\beta}^*_s\| + \|\boldsymbol{\beta}^*_s - \boldsymbol{\beta}_i\| \\
\le& \Delta_{\max} \frac{2ck^2(1 + \frac{2k^*}{k}) + 4k^*(3 + \frac{2k^*}{k})}{\sqrt{\eta_{\max}}}, \qquad (43)
\end{aligned}$$

thereby proving the first bound in Lemma 4.

To prove the second bound in Lemma 4, we observe that by definition of $\boldsymbol{b}_i^+$ and $\boldsymbol{b}_i^-$, there holds

$$\begin{aligned}
\boldsymbol{b}_i^+ :=& \frac{1}{1 + |A_i \setminus \{s\}|} \left(\sum_{s' \in A_i \setminus \{s\}} \boldsymbol{\beta}^*_{s'} + \boldsymbol{\beta}^*_s \right) \\
=& \frac{|A_i \setminus \{s\}|}{1 + |A_i \setminus \{s\}|} \boldsymbol{b}_i^- + \frac{1}{1 + |A_i \setminus \{s\}|} \boldsymbol{\beta}^*_s.
\end{aligned}$$

whence $\|\boldsymbol{b}_i^+ - \boldsymbol{\beta}^*_s\| = \frac{|A_i \setminus \{s\}|}{1 + |A_i \setminus \{s\}|} \|\boldsymbol{b}_i^- - \boldsymbol{\beta}^*_s\|$. It follows that

$$\begin{aligned}
\|\boldsymbol{b}_i^+ - \boldsymbol{\beta}_i\| \le& \|\boldsymbol{b}_i^+ - \boldsymbol{\beta}^*_s\| + \|\boldsymbol{\beta}^*_s - \boldsymbol{\beta}_i\| \\
\le& \|\boldsymbol{b}_i^- - \boldsymbol{\beta}^*_s\| + \|\boldsymbol{\beta}^*_s - \boldsymbol{\beta}_i\| \\
\le& \Delta_{\max} \frac{2ck^2(1 + \frac{2k^*}{k}) + 4k^*(3 + \frac{2k^*}{k})}{\sqrt{\eta_{\max}}},
\end{aligned}$$

where the last step follows from equation (43).

It remains to show that $|A_i \setminus \{s\}| \ge 2$. Note that $A_i \neq \emptyset$ under the assumption $|W_i \setminus \{s\}| \ge 1$ of the lemma. For the sake of deriving a contradiction, assume that $W_i \setminus \{s\} = \{s'\}$, in which case $\boldsymbol{b}_i^- = \boldsymbol{\beta}^*_{s'}$. It then follows from inequality (42) that $\|\boldsymbol{\beta}^*_{s'} - \boldsymbol{\beta}^*_s\| \le \Delta_{\max} \frac{2ck^2(1 + \frac{2k^*}{k}) + 4k^*(3 + \frac{2k^*}{k})}{\sqrt{\eta_{\max}}}$, contradicting the separation assumption on $\eta_{\min}$ in Theorem 3. This completes the proof of Lemma 4.

Controlling the Volume

In this section, we show that the intersection of a Voronoi set and a ground truth cluster must be small if (i) the true center is not in the Voronoi set and (ii) the intersection of the true cluster and the boundary of the Voronoi set is small. This is formalized in the following lemma.

Lemma 8 (Controlling the volume of intersection)**.** *Let μ be the uniform distribution on $\mathbb{B}_{\mathbf{0}}(r)$. Let P be a closed polyhedron with at most k facets satisfying $\mathbf{0} \notin \mathrm{int}(P)$. If each facet F of P satisfies $\frac{1}{r^d V_d} \mathrm{ReVol}(F \cap \mathbb{B}_{\mathbf{0}}(r)) \le \lambda$, then we have $\mu(P) \le k\lambda r$.*

Proof. Introduce the shorthand $\mathbb{B} := \mathbb{B}_{\mathbf{0}}(r)$. We may assume that $P \cap \mathbb{B} \neq \emptyset$, because otherwise the lemma is trivially true. We claim that one may shift the polyhedron P by a distance r so that its intersection with the ball $\mathbb{B}$ has zero measure. That is, there exists a unit vector $\boldsymbol{v}$ such that $\big(P + (r+\epsilon)\boldsymbol{v}\big) \cap \mathbb{B} = \emptyset$ for all $\epsilon > 0$. We further claim that $\boldsymbol{v}$ can be chosen in such a way that the intersection $P \cap \mathbb{B}$ is enclosed by the original boundary and the shifted boundary; that is, $P \cap \mathbb{B} \subseteq (\partial P \cap \mathbb{B}) + L_r$, where we $L_r := \{t\boldsymbol{v} : t \in [0,r]\}$ is a line segment. Figure 6 provides an illustration of these two claims, whose proof is deferred to the end of this section. Therefore, we have the bound

$$\begin{aligned}
\mu(P) =& \frac{\mathrm{Vol}(P \cap \mathbb{B})}{\mathrm{Vol}(\mathbb{B})} \le \frac{\mathrm{Vol}\big((\partial P \cap \mathbb{B}) + L_r\big)}{\mathrm{Vol}(\mathbb{B})} \\
\le& \frac{r \sum_{F \in \mathcal{F}} \mathrm{ReVol}(F \cap \mathbb{B})}{r^d V_d} \le rk\lambda,
\end{aligned}$$

where $\mathcal{F} := \{F : F \text{ is a facet of } P\}$ satisfies $|\mathcal{F}| \le k$ by assumption. This completes the proof of the lemma.

Let us prove the two claims above. Because P is convex and $\mathbf{0} \notin \mathrm{int}(P)$, the separating hyperplane theorem ensures that

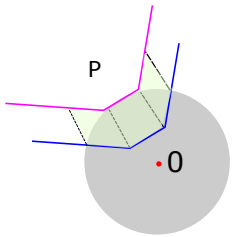

Figure 6. Shifting the boundary of the polyhedron P to bound the volume of $P \cap \mathbb{B}_{\mathbf{0}}(r)$.

there exists some unit vector $\boldsymbol{v}$ such that $\langle \boldsymbol{x}, \boldsymbol{v} \rangle \geq 0, \forall \boldsymbol{x} \in P$. Therefore, for all $\boldsymbol{x} \in P$ and $\epsilon > 0$, we have

$$\begin{aligned}\|\boldsymbol{x} + (r+\epsilon)\boldsymbol{v}\|^2 &= \|\boldsymbol{x}\|^2 + 2(r+\epsilon)\langle \boldsymbol{x}, \boldsymbol{v}\rangle + (r+\epsilon)^2\|\boldsymbol{v}\|^2 \\ &\geq 0 + 0 + (r+\epsilon)^2 > r^2,\end{aligned}$$

whence $\boldsymbol{x} + (r+\epsilon)\boldsymbol{v} \notin \mathbb{B}$, proving the first claim. To prove the second claim, fix an arbitrary $\boldsymbol{x} \in P \cap \mathbb{B}$ and consider the half line $\ell := \{\boldsymbol{x} - t\boldsymbol{v} : t \geq 0\}$. Note that ℓ must intersect the boundary ∂P; otherwise we would have $\ell \subseteq P$ and hence the separating hyperplane property implies that $\langle \boldsymbol{x} - t\boldsymbol{v}, \boldsymbol{v}\rangle \geq 0$ for all $t \geq 0$, which cannot hold as $\boldsymbol{v}$ has unit norm. Since P is convex, ℓ intersects ∂P at a unique point, say $\boldsymbol{x}_0 = \boldsymbol{x} - t_0\boldsymbol{v}$. We must have $t_0 \leq r$; otherwise we would have $\boldsymbol{x} = \boldsymbol{x}_0 + t_0\boldsymbol{v} \in P + (r+\epsilon)\boldsymbol{v}$ for some $\epsilon > 0$ and hence $\boldsymbol{x} \notin \mathbb{B}$ by the first claim, which is a contradiction. Using the separating hyperplane property $0 \leq \langle \boldsymbol{x} - t_0\boldsymbol{v}, \boldsymbol{v}\rangle \leq \langle \boldsymbol{x}, \boldsymbol{v}\rangle$ again, we have

$$\begin{aligned}\|\boldsymbol{x}_0\|^2 =& \|\boldsymbol{x} - t_0\boldsymbol{v}\|^2 \\ =& \|\boldsymbol{x}\|^2 + \langle -t_0\boldsymbol{v}, \boldsymbol{x}\rangle + \langle \boldsymbol{x} - t_0\boldsymbol{v}, -t_0\boldsymbol{v}\rangle \leq r^2 + 0 + 0\end{aligned}$$

and thus $\boldsymbol{x}_0 \in \mathbb{B} \cap \partial P$. Combining pieces, we conclude that $\boldsymbol{x} = \boldsymbol{x}_0 + t_0\boldsymbol{v} \in (\partial P \cap \mathbb{B}) + L_r$. As $\boldsymbol{x} \in P \cap \mathbb{B}$ is arbitrary, we have $P \cap \mathbb{B} \subseteq (\partial P \cap \mathbb{B}) + L_r$ as claimed. □

Below we state a analogous version of Lemma 8 under the Gaussian density. This result is used in the proof of our main Theorem 5 for the Gaussian Mixture model.

Lemma 9 (Controlling the volume of intersection, Gaussian)**.** *Let μ be the probability measure with respect to a spherical Gaussian distribution in $\mathbb{R}^d$ with mean $\mathbf{0}$ and variance σ^2. Let P be a closed polyhedron with at most k facets satisfying $\mathbf{0} \notin \mathrm{int}(P)$. If each facet F of P satisfies $\rho_1(F) \leq \lambda$, where $\rho_1(F)$ is defined as in equation (48) with $\boldsymbol{\beta}_1^* = \mathbf{0}$, then we have $\mu(P \cap \mathbb{B}_{\mathbf{0}}(r)) \leq k\lambda r$ for any $r > 0$.*

Proof. We fix r and introduce the shorthand $\mathbb{B} \equiv \mathbb{B}_{\mathbf{0}}(r)$. From the proof of Lemma 8, there exists a unit vector $\boldsymbol{v} = (v_1, \ldots, v_d)$ satisfying (i) $\langle \boldsymbol{v}, \boldsymbol{x}\rangle \geq 0$ for all $\boldsymbol{x} \in P$; (ii) $P \cap \mathbb{B} \subseteq \partial P \cap \mathbb{B} + L_r$, where $L_r := \{t\boldsymbol{v} : t \in [0, r]\}$. It follows that

$$\mu(P \cap \mathbb{B}) \leq \mu(\partial P \cap \mathbb{B} + L_r) \leq \sum_{F \in \mathcal{F}} \mu(F \cap \mathbb{B} + L_r),$$

where $\mathcal{F} := \{F : F \text{ is a facet of } P\}$. Since by assumption $|\mathcal{F}| \leq k$ and each facet $F \in \mathcal{F}$ satisfies $\rho_1(F) \leq \lambda$, the lemma follows if we can show that $\mu(F \cap \mathbb{B} + L_r) \leq \rho_1(F) \cdot r$.

Without loss of generality, we may use an orthonormal basis $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d\}$ such that $\boldsymbol{w}_1$ is the normal vector of F and $\langle \boldsymbol{v}, \boldsymbol{w}_1\rangle \geq 0$. Under this basis, we have $x_1 = \langle \boldsymbol{x}, \boldsymbol{w}_1\rangle = c$ for all $\boldsymbol{x} \in F$, where c is a fixed number. We define a vertical slice at $x_1 = c$ as follows:

$$S_F := \{(x_2, \ldots, x_d) : (c, x_2, \ldots, x_d) \in F \cap \mathbb{B}\}.$$

With this notation, the definition of $\rho_1(F)$ is equivalent to

$$\begin{aligned}&\rho_1(F) \\ =& \int_{(x_2,\ldots,x_d)\in S_F} \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^d} \exp\left(-\frac{c^2 + \sum_{j=2}^d x_j^2}{2\sigma^2}\right) \mathrm{d}x_2 \ldots \mathrm{d}x_d,\end{aligned}$$

and the set $F \cap \mathbb{B} + L_r$ admits the explicit expression

$$\begin{aligned}F \cap \mathbb{B} + L_r = \big\{\boldsymbol{z} : \boldsymbol{z} = (c, x_2, \ldots, x_d) + t\boldsymbol{v}, \\ (x_2, \ldots, x_d) \in S_F, t \leq r\big\}.\end{aligned} \tag{44}$$

We are now ready to bound the quantity of interest

$$\begin{aligned}&\mu(F \cap \mathbb{B} + L_r) \\ =& \int_{\boldsymbol{z}\in F\cap\mathbb{B}+L_r} \frac{1}{(\sqrt{2\pi\sigma^2})^d} \exp\left(-\frac{\sum_{j=1}^d z_j^2}{2\sigma^2}\right) \mathrm{d}z_1 \ldots \mathrm{d}z_d.\end{aligned}$$

Using the expression (44), we may rewrite the above integral by a change of variable from $(z_1, z_2, \ldots, z_d)$ to $(t, x_2, \ldots, x_d)$. To this end, note that the corresponding Jacobian matrix is

$$J = \begin{bmatrix} v_1 & 0 & 0 & \ldots & 0 \\ v_2 & 1 & 0 & \ldots & 0 \\ \vdots & & & & \\ v_d & 0 & 0 & \ldots & 1 \end{bmatrix},$$

which is a lower triangular matrix satisfying $|\det(J)| = |v_1|$. Consequently, we have

$$\begin{aligned}&\mu(F \cap \mathbb{B} + L_r) \\ =& \int_0^r \Bigg[\int_{(x_2,\ldots,x_d)\in S_F} \frac{1}{(\sqrt{2\pi\sigma^2})^d} \cdot \\ &\exp\left(-\frac{(c+tv_1)^2 + \sum_{j=2}^d (x_j + tv_j)^2}{2\sigma^2}\right) \mathrm{d}x_2 \ldots \mathrm{d}x_d\Bigg] |v_1| \mathrm{d}t.\end{aligned}$$

To upper bound the Gaussian density above, we note that for any $t \geq 0$ and $(c, x_2, \ldots, x_d) \in F \subseteq P$, the property of the separating hyperplane ensures that $t(v_1 c + \sum_{j=2}^d v_j x_j) \geq 0$, whence

$$\begin{aligned}&\exp\left(-\frac{(c+tv_1)^2 + \sum_{j=2}^d (x_j + tv_j)^2}{2\sigma^2}\right) \\ =& \exp\left(-\frac{c^2 + \sum_{j=2}^d (x_j)^2}{2\sigma^2}\right) \exp\left(-\frac{t(v_1 c + \sum_{j=2}^d v_j x_j)}{\sigma^2}\right) \cdot \\ &\exp\left(-\frac{t^2}{2\sigma^2}\right) \\ \leq& \exp\left(-\frac{c^2 + \sum_{j=2}^d x_j^2}{2\sigma^2}\right).\end{aligned}$$

It follows that

$$\begin{aligned}&\mu(F \cap \mathbb{B} + L_r) \\ \leq& \int_0^r \Bigg[\int_{(x_2,\ldots,x_d)\in S_F} \frac{1}{(\sqrt{2\pi\sigma^2})^d} \cdot\end{aligned}$$

$$\exp\left(-\frac{c^2+\sum_{j=2}^d (x_j)^2}{2\sigma^2}\right)\mathrm{d}x_2\ldots\mathrm{d}x_d\Big]|v_1|\mathrm{d}t$$
$$=\int_0^r \rho_1(F)\cdot|v_1|\mathrm{d}t \le \rho_1(F)\cdot r,$$

where the last step holds since $\boldsymbol{v}$ is a unit vector with $|v_1|\le 1$. This completes the proof of Lemma 9. □

Controlling the distance to the center

In this section, we prove the following result:

Lemma 10 (Bound on the center of mass, ball)**.** *Let μ be the uniform distribution over the ball $\mathbb{B}_{\mathbf{0}}(r)\subset\mathbb{R}^d$. Suppose that a subset $S\subset\mathbb{R}^d$ has probability measure $\mu(S)>0$. Let $\boldsymbol{c}_S$ be the center of mass of the set S with respect to μ. We have the bound*

$$\|\boldsymbol{c}_S\| \le \frac{r\cdot\mu(\mathbb{R}^d\setminus S)}{\mu(S)}.$$

Proof. Recall the expression of the center of mass $\boldsymbol{c}_S = \frac{\int \boldsymbol{x}\mathbb{1}_S(\boldsymbol{x})\mathrm{d}\mu}{\mu(S)}$. We have

$$\begin{aligned}\mu(S)\cdot\|\boldsymbol{c}_S\| &= \left\|\int \boldsymbol{x}\mathbb{1}_S(\boldsymbol{x})\mathrm{d}\mu\right\| \\ &\overset{\text{(i)}}{=} \left\|-\int \boldsymbol{x}\mathbb{1}_{\mathbb{R}^d\setminus S}(\boldsymbol{x})\mathrm{d}\mu\right\| \\ &\overset{\text{(ii)}}{\le} \int\|\boldsymbol{x}\|\mathbb{1}_{\mathbb{R}^d\setminus S}(\boldsymbol{x})\mathrm{d}\mu \\ &\overset{\text{(iii)}}{\le} r\mu(\mathbb{R}^d\setminus S),\end{aligned}$$

where step (i) holds since μ has mean $\mathbf{0}$, and step (ii) holds by the Jensen's inequality, and step (iii) holds because $\|\boldsymbol{x}\|\le r$ for all $\boldsymbol{x}\in\text{support}(\mu)=\mathbb{B}_{\mathbf{0}}(r)$. Rearranging the inequality proves the desired bound. □

Lemma 11 (Bound on the center of mass, Gaussian)**.** *Let μ be the Gaussian distribution $\mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I}_d)$. Suppose that a subset $S\subset\mathbb{R}^d$ has probability measure $\mu(S)>0$. Let $\boldsymbol{c}_S$ be the center of mass of the set S with respect to μ. We have the bound*

$$\|\boldsymbol{c}_S\| \le \frac{2\sigma\cdot\sqrt{\mu(\mathbb{R}^d\setminus S)}}{\mu(S)}.$$

Proof. Recall the variational characterization of the center of mass $\boldsymbol{c}_S$:

Since $\mathbf{0}$ is the mean of μ, we have

$$\begin{aligned}0\le&\|\boldsymbol{c}_S\|^2 \\ =&\int\|\boldsymbol{x}-\boldsymbol{c}_S\|^2\mathrm{d}\mu-\int\|\boldsymbol{x}\|^2\mathrm{d}\mu \\ =&\int\|\boldsymbol{x}-\boldsymbol{c}_S\|^2\mathbb{1}_S(\boldsymbol{x})\mathrm{d}\mu-\int\|\boldsymbol{x}\|^2\mathbb{1}_S(\boldsymbol{x})\mathrm{d}\mu \\ &+\int\|\boldsymbol{x}-\boldsymbol{c}_S\|^2\mathbb{1}_{\mathbb{R}^d\setminus S}(\boldsymbol{x})\mathrm{d}\mu-\int\|\boldsymbol{x}\|^2\mathbb{1}_{\mathbb{R}^d\setminus S}(\boldsymbol{x})\mathrm{d}\mu \\ \overset{\text{(i)}}{\le}&\int\|\boldsymbol{x}-\boldsymbol{c}_S\|^2\mathbb{1}_{\mathbb{R}^d\setminus S}(\boldsymbol{x})\mathrm{d}\mu-\int\|\boldsymbol{x}\|^2\mathbb{1}_{\mathbb{R}^d\setminus S}(\boldsymbol{x})\mathrm{d}\mu \\ =&\mu(\mathbb{R}^d\setminus S)\|\boldsymbol{c}_S\|^2-2\int\langle\boldsymbol{x},\boldsymbol{c}_S\rangle\mathbb{1}_{\mathbb{R}^d\setminus S}(\boldsymbol{x})\mathrm{d}\mu. \qquad (45)\end{aligned}$$

where step (i) follows from the variational characterization of the center of mass

$$\boldsymbol{c}_S=\text{argmin}_{\boldsymbol{z}\in\mathbb{R}^d}\int\|\boldsymbol{x}-\boldsymbol{z}\|^2\mathbb{1}_S(\boldsymbol{x})\mathrm{d}\mu.$$

Rearranging equation (45) gives

$$\begin{aligned}\mu(S)\|\boldsymbol{c}_S\|^2 &\le -2\int\langle\boldsymbol{x},\boldsymbol{c}_S\rangle\mathbb{1}_{\mathbb{R}^d\setminus S}(\boldsymbol{x})\mathrm{d}\mu \\ &\overset{\text{(ii)}}{\le} 2\|\boldsymbol{c}_S\|\sqrt{\int\left\langle\boldsymbol{x},\frac{\boldsymbol{c}_S}{\|\boldsymbol{c}_S\|}\right\rangle^2\mathrm{d}\mu}\cdot\sqrt{\int\mathbb{1}_{\mathbb{R}^d\setminus S}(\boldsymbol{x})\mathrm{d}\mu} \\ &\overset{\text{(iii)}}{=} 2\|\boldsymbol{c}_S\|\sigma\cdot\sqrt{\mu(\mathbb{R}^d\setminus S)},\end{aligned}$$

where step (ii) follows from Cauchy-Schwarz and step (iii) follows from the fact that any one-dimensional margin of $\mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I}_d)$ is the univariate Gaussian distribution $\mathcal{N}(0,\sigma^2)$. Rearranging the above equation proves the desired bound. □

Proof of Theorem 2

We prove a more general and quantitative version Theorem 2 that holds for any k^* and k.

Theorem 5 (General version of Theorem 2)**.** *Let t be any number satisfying $\varphi(t):=2\exp(-t^2\min(d,k+k^*)/8)<\frac{1}{4}$. Under the Gaussian Mixture Model, assume that $\eta_{\max}\ge 16tc^2k^4$ and $\eta_{\min}\ge\left[\frac{k^*}{k\sqrt{t}}\left(2+\frac{c\sqrt{t}}{\sqrt{\eta_{\max}}}\right)+3ck(k^*+k)\sqrt{t}\right]\sqrt{\eta_{\max}}+2(2k^*+1)\varphi(t)\eta_{\max}$ for some constant $c\ge 3$. If $\boldsymbol{\beta}=(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k)\in\mathbb{R}^{d\times k}$ is a local minimum of G, then the ground truth centers and fitted centers can be partitioned as $[k]=\bigcup_{a=1}^m S_a^*$ and $[k]=\bigcup_{a=0}^m S_a$, respectively, such that for each $a\in[m]$, exactly one of the following holds:*

- ***(many/one-fit-one association)*** *$|S_a|\ge 1$ and $S_a^*=\{s\}$ for some $s\in[k]$; moreover,*

$$\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_s^*\|\le\Delta_{\max}\left\{\frac{\frac{2k^*}{ck\sqrt{t}}+2ck(k^*+k)\sqrt{t}}{\sqrt{\eta_{\max}}}+7k^*\varphi(t)\right\}$$
$$\forall i\in S_a. \qquad (46)$$

- ***(one-fit-many association)*** *$S_a=\{i\}$ for some $i\in[k]$ and $|S_a^*|\ge 2$; moreover,*

$$\begin{aligned}&\left\|\boldsymbol{\beta}_i-\frac{1}{|S_a^*|}\sum_{s\in S_a^*}\boldsymbol{\beta}_s^*\right\| \\ \le&\Delta_{\max}\left\{\frac{\frac{2k^*}{ck\sqrt{t}}+3ck(k^*+k)\sqrt{t}}{\sqrt{\eta_{\max}}}+7k^*\varphi(t)\right\} \qquad (47)\end{aligned}$$

In addition, for each $i\in S_0$, we have $\mathbb{P}\big(\mathcal{V}_i(\boldsymbol{\beta})\big)\le\frac{ck\sqrt{t}}{\sqrt{\eta_{\max}}}+\varphi(t)$ ***(almost-empty association)****.*

Given Theorem 5, Theorem 2 follows immediately. The rest of this section is devoted to proving Theorem 5.

As before, we use $\mathbb{P}$ to denote the probability measure with respect to f and $\mathbb{P}_s$ to denote the probability measure with respect to f_s, where f is the density of the Gaussian mixture and f_s is the density of the s-th Gaussian component. Recall the population k-mean objective function:

$$G(\boldsymbol{\beta})=\int_{\boldsymbol{x}}\min_{j\in[k]}\|\boldsymbol{x}-\boldsymbol{\beta}_j\|^2 f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$= \frac{1}{k^*} \sum_{s=1}^{k^*} \int_{\boldsymbol{x}} \min_{j\in[k]} \|\boldsymbol{x} - \boldsymbol{\beta}_j\|^2 f_s(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Reduction to lower dimensions

We first argue that it suffices to prove the theorem in dimension $d' \le k + k^*$. Once this is established, then the theorem for $d > k^* + k$ dimensions can be deduced as follows. Suppose that $\boldsymbol{\beta}^* \in \mathbb{R}^{d\times k^*}$ is the ground truth solution and $\boldsymbol{\beta} \in \mathbb{R}^{d\times k}$ is a candidate solution. We may choose a coordinate system such that the first $d' = k^* + k$ dimensions correspond to $\mathrm{span}\{\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_{k^*}^*, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k\}$. In this case, for each $s \in [k^*]$ we have $\boldsymbol{\beta}_s^* = (\boldsymbol{\beta}_s^{*\prime}, \boldsymbol{0})$ for some $\boldsymbol{\beta}_s^{*\prime} \in \mathbb{R}^{d'}$ and for each $i \in [k]$, we have and $\boldsymbol{\beta}_i = (\boldsymbol{\beta}_i', \boldsymbol{0})$ for some $\boldsymbol{\beta}_i' \in \mathbb{R}^{d'}$. Moreover, thanks to Gaussian's rotational invariance, the d-dimensional Gaussian mixture is a product distribution with respect to the first d' dimensions and the last $d - d'$ dimensions, where the first d'-dimensional margin is itself a Gaussian mixture. Indeed, for any $\boldsymbol{x} = (\boldsymbol{x}', \boldsymbol{z}) \in \mathbb{R}^d$ with $\boldsymbol{x}' \in \mathbb{R}^{d'}$ and $\boldsymbol{z} \in \mathbb{R}^{d-d'}$, the density of the Gaussian mixture factorizes:

$$\begin{aligned} f(\boldsymbol{x}) &\propto \frac{1}{k^*} \sum_{s=1}^{k^*} \exp\left(\frac{\|\boldsymbol{x} - \boldsymbol{\beta}_s^*\|^2}{2\sigma^2}\right) \\ &= \frac{1}{k^*} \sum_{s=1}^{k^*} \exp\left(\frac{\|(\boldsymbol{x}', \boldsymbol{z}) - (\boldsymbol{\beta}_s^{*\prime}, \boldsymbol{0})\|^2}{2\sigma^2}\right) \\ &= \left[\frac{1}{k^*} \sum_{s=1}^{k^*} \exp\left(\frac{\|\boldsymbol{x}' - \boldsymbol{\beta}_s^{*\prime}\|^2}{2\sigma^2}\right)\right] \cdot \exp\left(\frac{\|\boldsymbol{z}\|^2}{2\sigma^2}\right). \end{aligned}$$

Now, if $\boldsymbol{\beta}$ is a local minimum of G, then $\boldsymbol{\beta}'$ is also a local minimum of G restricted to the first $d' = k^* + k$ dimensions. Applying the theorem with dimension d', we obtain bounds on the quantities $\|\boldsymbol{\beta}_i' - \boldsymbol{\beta}_s^{*\prime}\|$, $\|\boldsymbol{\beta}_i' - \sum_{s\in S} \boldsymbol{\beta}_s^{*\prime}\|$ and $\mathbb{P}(\mathcal{V}_i(\boldsymbol{\beta}'))$. We claim that these three quantities are equal to $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_s^*\|$, $\|\boldsymbol{\beta}_i - \sum_{s\in S} \boldsymbol{\beta}_s^*\|$ and $\mathbb{P}(\mathcal{V}_i(\boldsymbol{\beta}))$, respectively. Indeed, the first two equalities are immediate under our coordinate system; the last equality holds because the Gaussian mixture factorzes (shown above) and so do the Voronoi sets: $\mathcal{V}_i(\boldsymbol{\beta}) = \mathcal{V}_i(\boldsymbol{\beta}') \times \mathbb{R}^{d-d'}$. We conclude that the same collection of bounds hold in dimension d as well. In the rest of the proof, we can safely assume that $d \le k^* + k$, in which case $\min\{k^* + k, d\} = d$.

As in the proof for the Stochastic Ball Model, we first establish a general result, an analogue of Theorem 4, that provides a family of bounds parametrized by two positive numbers λ and t. To state this result, we introduce some additional notation. Set $r = t\sigma\sqrt{d}$ and let

$$\mathbb{B}_s(r) := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{\beta}_s^*\| \le r\}.$$

denote the ball centered at $\boldsymbol{\beta}_s^*$ with radius r. Recall that for each $i, j \in [k]$ and $s \in [k^*]$, the Voronoi boundary $\partial_{i,j}$ lies in a $(d-1)$-dimensional affine subspace $\mathcal{L}$. Since the distribution f_s of the s-th component is rotationally invariant, we may assume WLOG that $\mathcal{L} = \{\boldsymbol{x} \in \mathbb{R}^d : x_1 = z\}$ for some number z. Accordingly, we define the quantity

$$\rho_s(\partial_{i,j}) := \int \mathbb{1}\{(z, x_2, \ldots, x_d) \in \partial_{i,j} \cap \mathbb{B}_s(r)\} \cdot f_s(z, x_2, \ldots, x_d)\, \mathrm{d}x_2 \ldots \mathrm{d}x_d, \tag{48}$$

which is a measure of the relative probability mass of the Voronoi boundary $\partial_{i,j}$ when restricted to the ball $\mathbb{B}_s(r)$.[5] Also recall the function $\varphi(\cdot)$ defined in the statement of Theorem 5, which satisfies $\varphi(t) = 2\exp(-t^2 d/8)$ when $d \le k^* + k$.

We now state the family of bound in the Gaussian case:

Theorem 6 (Family of bounds for Gaussian)**.** *Under the Gaussian mixture model, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ be a local minima for the k-means objective function G defined in (3). Let $\lambda > 0$ and $t > 0$ be two arbitrary fixed numbers and set $r := t\sigma\sqrt{d}$. For each $i \in [k]$, define the sets:*

$$\begin{aligned} T_i &:= \big\{s \in [k^*] : \mathcal{V}_i \cap \mathbb{B}_s(r) \ne \emptyset\big\} \quad \text{and} \\ A_i &:= \big\{s \in [k^*] : \boldsymbol{\beta}_s^* \in \mathrm{int}(\mathcal{V}_i)\big\} \subseteq T_i. \end{aligned}$$

Then the following is true for each $i \in [k]$:

1) *If $\rho_s(\partial_{j,\ell}) > \lambda$ for some pair (j, ℓ) and $s \in T_i$, then*

$$\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_s^*\| \le \frac{k^*}{\lambda} + 3r.$$

2) *For each $s \in T_i$, if $\rho_s(\partial_{j,\ell}) \le \lambda$ for all pairs (j, ℓ), then the following bounds hold*

$$\begin{aligned} \mathbb{P}_s(\mathcal{V}_i) &\ge 1 - k^2\lambda r - \varphi(t), \quad && \forall s \in A_i, \\ \mathbb{P}_s(\mathcal{V}_i) &\le k\lambda + \varphi(t), && \forall s \in T_i \setminus A_i. \end{aligned}$$

Furthermore, if $\rho_s(\partial_{j,\ell}) \le \lambda$ for all pair (j, ℓ) and and $s \in T_i$, then:

a) *When $|A_i| = 0$, we have*

$$\mathbb{P}(\mathcal{V}_i) \le k\lambda r + \varphi(t).$$

b) *When $|A_i| > 0$, we have*

$$\begin{aligned} &\|\boldsymbol{\beta}_i - \boldsymbol{b}_i\| \\ \le &\frac{2k^*\sigma}{1 - k^2\lambda r - \varphi(t)} \\ &+ \frac{2\sigma\sqrt{k^2\lambda r + \varphi(t)}(1 + (k^*k - k^2)\lambda r + (k^* - 1)\varphi(t))}{(1 - k^2\lambda r - \varphi(t))^2} \\ &+ \frac{(2k^* + k)k\lambda r + (2k^* + 1)\varphi(t)}{1 - k^2\lambda r - \varphi(t)} \Delta_{\max}, \end{aligned}$$

where $\boldsymbol{b}_i := \frac{1}{|A_i|}\sum_{s\in A_i} \boldsymbol{\beta}_s^*$.

We prove Theorem 6 in Section 0c to follow. Note that Theorem 6 is similar to Theorem 4 except that the error bounds here have an additional error term $\varphi(t)$.

The procedure to derive the main Theorem 5 from Theorem 6 is exactly same as that for the Stochastic Ball Model. In particular, with r fixed to be $t\sigma\sqrt{d}$, we set $\lambda = \frac{c}{\sqrt{r\Delta_{\max}}}$. The assumptions on t and $\eta_{\max}$ ensure that $\lambda k^2 r < \frac{1}{4}$ and $\varphi(t) = 2\exp(-t^2 d/8) < \frac{1}{4}$. Moreover, Observations 1 and 2 still hold in the current setting. We then construct, in three steps, a partition of the fitted centers $\{\boldsymbol{\beta}_i\}_{i\in[k]}$ as well as the true centers $\{\boldsymbol{\beta}_s^*\}_{s\in[k^*]}$. We provide a brief description here and reuse all the notation from the proof of Theorem 3.

[5] Note that when f_s is the uniform distribution over $\mathbb{B}_s(r)$, the definition here reduces to $\rho_s(\partial_{i,j}) = \frac{1}{V_d r^d}\mathrm{ReVol}(\partial_{i,j} \cap \mathbb{B}_s(r))$ and hence is consistent with our previous definition in the ball model.

a) Step 1 (almost empty association): We consider the set

$$S_0 := \big\{ i \in [k] : \rho_s(\partial_{j,\ell}) \le \lambda, \quad \forall (s,j,\ell) \in T_i \times [k] \times [k]; |A_i| = 0 \big\}.$$

Part 2(a) of Theorem 6 ensures that for all $i \in S_0$, we have $\mathbb{P}(\mathcal{V}_i) \le k\lambda r + \varphi(t) = \frac{ck\sqrt{t}}{\sqrt{\eta_{\max}}} + \varphi(t)$.

b) Step 2 (many/one fit one association):: We then consider the fitted centers indexed by the set

$$\mathcal{J} := \{ i \in [k] : |A_i| \le 1 \} \setminus S_0.$$

For each $i \in \mathcal{J}$, there are two complementary cases:

- $\rho_s(\partial_{j,k}) \le \lambda$ for all $(s,j,\ell) \in T_i \times [k] \times [k]$. We must have $|A_i| = 1$; say $A_i = \{s\}$. Applying Part 2(b) of Theorem 6, we have

$$\begin{aligned}
&\|\boldsymbol{\beta}_i - \boldsymbol{\beta}^*_s\| \\
=&\|\boldsymbol{\beta}_i - \boldsymbol{b}_i\| \\
\le& \frac{2k^*\sigma}{1 - k^2\lambda r - \varphi(t)} \\
&+ \frac{2\sigma\sqrt{k^2\lambda r + \varphi(t)}(1 + (k^*k - k^2)\lambda r + (k^* - 1)\varphi(t))}{(1 - k^2\lambda r - \varphi(t))^2} \\
&+ \frac{(2k^* + k)k\lambda r + (2k^* + 1)\varphi(t)}{1 - k^2\lambda r - \varphi(t)} \Delta_{\max} \\
\overset{\text{(i)}}{\le}& 4k^*\sigma + 4\sqrt{2}\sigma(1 + c(k^*k - k^2)/\sqrt{\eta_{\max}/t} + (k^* - 1)\varphi(t)) \\
&+ 2(c(2k^* + k)k/\sqrt{\eta_{\max}/t} + (2k^* + 1)\varphi(t))\Delta_{\max} \\
\overset{\text{(ii)}}{\le}& \Delta_{\max} \left\{ \frac{\frac{2k^*}{ck\sqrt{t}} + 2ck(k^* + k)\sqrt{t}}{\sqrt{\eta_{\max}}} + 7k^*\varphi(t) \right\} \qquad (49)
\end{aligned}$$

  where in step (i) we plug in $\lambda = \frac{c}{\sqrt{\Delta_{\max} r}}$ with $r = t\sigma\sqrt{d}$ and use the assumption that $k^2\lambda r + \varphi(t) < \frac{1}{2}$; in step (ii), we use the assumption that $\eta_{\max} \ge 16tc^2k^4$ to further simplify the bound.
- $\rho_s(\partial_{j,\ell}) > \lambda$ for some $(s,j,\ell) \in T_i \times [k] \times [k]$; that is, there exists some ground truth cluster $\mathbb{B}_s$ that encloses a Voronoi boundary with a large relative volume. Applying Part 1 of Theorem 4 and plugging the value of λ, we obtain that $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}^*_s\| \le \frac{k^*}{\lambda} + 3r \le \Delta_{\max} \frac{4k^*\sqrt{t}}{c\sqrt{\eta_{\max}}}$.

For each distinct $s \in [k^*]$ that appears in the above arguments, let $S^*_a = \{s\}$ and let the corresponding S_a index those $\boldsymbol{\beta}_i$'s for which either of the two cases holds.

c) Step 3 (one-fit-many association):: We are left with the fitted centers indexed by the set

$$\mathcal{K} := \{ i \in [k] : |A_i| \ge 2 \} = [k] \setminus \big(S_0 \cup \mathcal{J} \big).$$

Similarly to before, for each $i \in \mathcal{K}$, there are two complementary cases:

- $\rho_s(\partial_{j,\ell}) \le \lambda$ for all $(s,j,\ell) \in T_i \times [k] \times [k]$. Applying Part 2(b) of Theorem 4 and following the same steps as in equation (49), we obtain that

$$\|\boldsymbol{\beta}_i - \boldsymbol{b}_i\| \le \Delta_{\max} \left\{ \frac{\frac{2k^*}{ck\sqrt{t}} + 2ck(k^* + k)\sqrt{t}}{\sqrt{\eta_{\max}}} + 7k^*\varphi(t) \right\}.$$

  In this case, we let $S_a = \{i\}$ and $S^*_a = A_i$.
- $\rho_s(\partial_{j,\ell}) > \lambda$ for some $(s,j,\ell) \in T_i \times [k] \times [k]$. We then have an analogous lemma as Lemma 4 to show that $\boldsymbol{\beta}_i$ is close to the mean of all the true centers contained in its Voronoi set, regardless of whether we include or exclude $\boldsymbol{\beta}^*_s$. The proof is exactly the same as in Section A; we omit its proof.

  Lemma 12 (Proximity to mean of true centers, Gaussian case)**.** *Under the assumption of Theorem 5, let $\boldsymbol{\beta}$ be a local minimum of G. The following is true for each $i \in [k]$. If $\rho_s(\partial_{j,\ell}) > \lambda = \frac{c}{\sqrt{r\Delta_{\max}}}$ for some $(s,j,\ell) \in T_i \times [k] \times [k]$ and $|A_i \setminus \{s\}| \ge 1$, then we have the bounds*

$$\|\boldsymbol{\beta}_i - \boldsymbol{b}^-_i\| \le \Delta_{\max} \left\{ \frac{\frac{2k^*}{ck\sqrt{t}} + 3ck(k^* + k)\sqrt{t}}{\sqrt{\eta_{\max}}} + 7k^*\varphi(t) \right\}$$

and

$$\|\boldsymbol{\beta}_i - \boldsymbol{b}^+_i\| \le \Delta_{\max} \left\{ \frac{\frac{2k^*}{ck\sqrt{t}} + 3ck(k^* + k)\sqrt{t}}{\sqrt{\eta_{\max}}} + 7k^*\varphi(t) \right\},$$

where $\boldsymbol{b}^-_i := \frac{1}{|A_i \setminus \{s\}|} \sum_{s' \in A_i \setminus \{s\}} \boldsymbol{\beta}^*_{s'}$ and $\boldsymbol{b}^+_i := \frac{1}{|A_i \cup \{s\}|} \sum_{s' \in A_i \cup \{s\}} \boldsymbol{\beta}^*_{s'}$; moreover, $|A_i \setminus \{s\}| \ge 2$.

  In this case, we let $S_a = \{i\}$; also let $S^*_a = A_i \setminus \{s\}$ if the index s has appeared in the sets $\{S^*_{a'}\}$ constructed previously in Step 2 or in this step, and set $S^*_a = A_i \cup \{s\}$.

The reason that the above construction corresponds to a partition for the fitted centers $\{\boldsymbol{\beta}_i\}_{i \in [k]}$ as well as a partition for the true centers $\{\boldsymbol{\beta}^*_s\}_{s \in [k^*]}$ is the same as the proof in Theorem 3.

Proof of Theorem 6

To establish Theorem 6 for the Gaussian model, we follow the same strategy as in the proof of Theorem 4 for the ball model. The only technical hurdle is that each Gaussian component distribution has unbounded support. Our main idea is to identify a bounded ball, namely $\mathbb{B}_s(r)$, that contains most of the probability mass of the s-th Gaussian component. Using a standard concentration inequality for χ^2 random variables (e.g., [56, Example 2.28]), we know that when $t > 2$, there holds the tail bound

$$\mathbb{P}_s\Big(\mathbb{B}_s(r)^{\complement} \Big) \le \varphi(t) = 2\exp(-t^2 d/8), \qquad (50)$$

where $S^{\complement}$ denotes the complement of a set $S \subseteq \mathbb{R}^d$. By restricting each s-th ground truth component to the ball $\mathbb{B}_s(r)$ and treating the tail mass in equation (50) as additional error terms, we can repeat most of the arguments used in the proof of the ball model. In what follows, we sketch the analysis and point out the minor modifications needed to adapt the proof of Theorem 4 to the Gaussian case.

The main step in the proof for the Ball model involves constructing smooth upper bounds for the function $W^{\boldsymbol{v}}_{i \to j,s} + W^{\boldsymbol{v}}_{j \to i,s}$, as done in Proposition 3 and Proposition 4. These two propositions still hold in the Gaussian case under the definition (48) of the "relative volume" $\rho_s(\partial_{i,j})$. In particular, the value of the integral defining $W^{\boldsymbol{v}}_{i \to j,s} + W^{\boldsymbol{v}}_{j \to i,s}$ does not increase if we restrict integration to the small set subset $(\mathbb{B}_s(r))$,

as the integrand is non-positive. Consequently, we can establish the two key inequalities (18) and (19), restated below:

$$d_{i,j} \cdot \rho_s(\partial_{i,j}) \le \frac{k^*}{2} \qquad \text{and} \qquad \frac{D^2_{i,j,s}}{d_{i,j}} \cdot \rho_s(\partial_{i,j}) \le \frac{k^*}{2}. \tag{51}$$

We can then derive the structural properties of a local minimum $\boldsymbol{\beta}$ from the inequalities (51). As in the proof of Theorem 4, for each $i \in [k]$ indexing the fitted center $\boldsymbol{\beta}_i$ and its Voronoi set $\mathcal{V}_i$, we consider two complementary cases.

Case 1: there exist some $(s, j, \ell) \in T_i \times [k] \times [k]$ such that $\rho_s(\partial_{j,\ell}) > \lambda$.: In this case, following exactly the same argument as in the Ball model proof, we can derive from the inequalities (51) that $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}^*_s\| \le \frac{k^*}{\lambda} + 3r$. This proves Part 1 of Theorem 6.

Case 2: for all $(s, j, \ell) \in T_i \times [k] \times [k]$ there holds $\rho_s(\partial_{j,\ell}) \le \lambda$.: Recall that $m_{i,s}$ is the probability mass of $\mathcal{V}_i$ with respect to the Gaussian density f_s and $\boldsymbol{c}_{i,s}$ is the corresponding center of mass of $\mathcal{V}_i$. If we restrict the density f_s onto the ball $\mathbb{B}_s(r)$, the values of $m_{i,s}$ and $\boldsymbol{c}_{i,s}$ do not change much; in particular, we can control the amount of change using the tail bound (50). With this in mind, we proceed by considering two sub cases.

- Case 2(a): $A_i = \emptyset$, in which case $T_i = B_i$. Following the same argument for deriving equation (20) and using the Gaussian Lemma 9 in place of Lemma 8, we obtain that

$$m_{i,s} \le k\lambda r + \mathbb{P}_s\big(\mathbb{B}_s(r)^{\complement}\big) \le k\lambda r + \varphi(t), \qquad \forall s \in B_i, \tag{52}$$

  where the second RHS term accounts for the tail probability on $\mathbb{B}_s(r)^{\complement}$. It follows that

$$\mathbb{P}(\mathcal{V}_i) = \frac{1}{k}\sum_{s \in T_i} m_{i,s} \le k\lambda r + \varphi(t).$$

  This proves Part 2(a) of Theorem 6.

- Case 2(b): $A_i \neq \emptyset$. By Lemma 2, $\boldsymbol{\beta}$ must satisfy the expression

$$\boldsymbol{\beta}_i = \frac{\sum_{s \in [k^*]} \boldsymbol{c}_{i,s} m_{i,s}}{\sum_{s \in [k^*]} m_{i,s}}.$$

  Using this expression, we have the decomposition $\boldsymbol{\beta}_i - \boldsymbol{b}_i = \boldsymbol{\mu} + \boldsymbol{\nu}$ for some vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ as in equation (22). To bound $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, we follow our general strategy to decompose the Gaussian density f_s into two parts, one supported on the ball $\mathbb{B}_s(r)$ and the other the tail, where the tail probability is bounded by $\varphi(t)$ as in equation (50). By doing so, we can establish analogous versions of the bounds (29) and (30) as given below:

$$\begin{cases} m_{i,s} \le k\lambda r + \varphi(t), & \text{if } s \in B_i, \\ m_{i,s} \ge 1 - k^2\lambda r - \varphi(t), & \text{if } s \in A_i, \end{cases}$$

  and

$$\|\boldsymbol{\beta}^*_s - \boldsymbol{c}_{i,s}\| \le \begin{cases} \frac{2\sigma}{m_{i,s}}, & \text{if } s \in B_i, \\ \frac{2\sigma\sqrt{k^2\lambda r + \varphi(t)}}{m_{i,s}}, & \text{if } s \in A_i, \end{cases}$$

  where the bound on $\|\boldsymbol{\beta}^*_s - \boldsymbol{c}_{i,s}\|$ follows from Lemma 11. Using the above two bounds, we can further establish analogous versions of Lemmas 5 and 6 as given below:

$$\begin{aligned} \|\boldsymbol{\mu}\| &\le \frac{2k^*\sigma}{1 - k^2\lambda r - \varphi(t)} \\ &\quad + \frac{2k^*(k\lambda r + \varphi(t))\sigma\sqrt{k^2\lambda r + \varphi(t)}}{(1 - k^2\lambda r - \varphi(t))^2} \\ &\quad + \frac{2k^*(k\lambda r + \varphi(t))}{1 - k^2\lambda r - \varphi(t)}\Delta_{\max}, \\ \|\boldsymbol{\nu}\| &\le \frac{2\sigma\sqrt{k^2\lambda r + \varphi(t)}}{1 - k^2\lambda r - \varphi(t)} \\ &\quad + \frac{k^2\lambda r + \varphi(t)}{1 - k^2\lambda r - \varphi(t)}\Delta_{\max}. \end{aligned}$$

  It follows that an analogue of inequality (**??**) holds:

$$\begin{aligned} &\|\boldsymbol{\beta}_i - \boldsymbol{b}_i\| \\ \le &\|\boldsymbol{\mu}\| + \|\boldsymbol{\nu}\| \\ \le &\frac{2k^*\sigma}{1 - k^2\lambda r - \varphi(t)} \\ &+ \frac{2\sigma\sqrt{k^2\lambda r + \varphi(t)}(1 + (k^*k - k^2)\lambda r + (k^* - 1)\varphi(t))}{(1 - k^2\lambda r - \varphi(t))^2} \\ &+ \frac{(2k^* + k)k\lambda r + (2k^* + 1)\varphi(t)}{1 - k^2\lambda r - \varphi(t)}\Delta_{\max}. \end{aligned}$$

  This proves Part 2(b) of Theorem 6.

Additional Examples

In this section, we provide two concrete examples that give additional insights on the behaviors of the local minima of the k-means objective and corroborate the results in our main theorems.

Our main theorem assumes certain separation conditions in terms of the SNRs $\eta_{\min}$ and $\eta_{\max}$. The first example shows that if the SNR is too small, then a local minimum may fail to have the structure described in Theorem 3. Therefore, a separation condition on the true clusters is in general necessary.

Example 1 (Small Separation)**.** *Consider the Stochastic Ball Model with $k = 3$ in dimension $d = 1$, where the ground truth cluster centers are $\beta^*_1 = -1$, $\beta^*_2 = 0$ and $\beta^*_3 = 1$, and the radius r of the balls satisfies $(\frac{9\sqrt{2}}{2} - \frac{1}{4})r > 1$. Let $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ be a candidate solution with $\beta_1 = -\frac{2}{3} - \frac{1}{6}r$, $\beta_2 = \frac{2}{3} + \frac{1}{6}r$ and $\beta_3 > 0$ sufficiently large.*

When β_3 is sufficiently large, the minimization $\min_{i \in [k]} \|x - \beta_i\|^2$ in the objective G is never attained by $i = 3$. In this case, the only effective variables for G are the first two centers β_1 and β_2. The Voronoi boundary $\partial_{1,2}(\boldsymbol{\beta})$ (which is 0) intersects the second ground truth cluster. Note that these properties continue to hold under small perturbation of $\boldsymbol{\beta}$. Consequently, for any solution $\boldsymbol{b}$ in a small neighborhood of $\boldsymbol{\beta}$, its objective value has the following expression:

$$\begin{aligned} G(\boldsymbol{b}) = &\frac{1}{6r}\bigg[\int_{-1-r}^{-1+r}(x - b_1)^2\mathrm{d}x + \int_{-r}^{\frac{b_1+b_2}{2}}(x - b_1)^2\mathrm{d}x \\ &+ \int_{\frac{b_1+b_2}{2}}^{r}(x - b_2)^2\mathrm{d}x + \int_{1-r}^{1+r}(x - b_2)^2\mathrm{d}x\bigg]. \end{aligned}$$

We compute the gradient and Hessian of G at $\boldsymbol{b}$ (recall that only the first two coordinate of $\boldsymbol{b}$ are effective):

$$\nabla_{\boldsymbol{b}} G = \frac{1}{6r}\begin{bmatrix} -2\int_{-1-r}^{-1+r}(x - b_1)\mathrm{d}x - 2\int_{-r}^{\frac{b_1+b_2}{2}}(x - b_1)\mathrm{d}x \\ -2\int_{\frac{b_1+b_2}{2}}^{r}(x - b_2)\mathrm{d}x - 2\int_{1-r}^{1+r}(x - b_2)\mathrm{d}x \end{bmatrix},$$

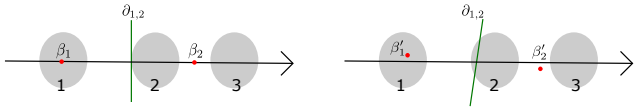

Figure 7. Two-dimensional Stochastic Ball Model, where the true centers are $\boldsymbol{\beta}_1^* = (-1, 0)$, $\boldsymbol{\beta}_2^* = (0, 0)$ and $\boldsymbol{\beta}_3^* = (1, 0)$. The third fitted center $\boldsymbol{\beta}_3$ is far away from the origin, so the only effective variables are $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Left panel: $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2 = \frac{1}{2}(\boldsymbol{\beta}_2^* + \boldsymbol{\beta}_3^*)$. The green line is the Voronoi boundary $\partial_{1,2}(\boldsymbol{\beta})$. Right panel: A perturbed solution $(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ and the corresponding Voronoi boundary.

$$\nabla^2_{\boldsymbol{b}} G = \frac{1}{6r}\begin{bmatrix} 6r + (b_1 + b_2) - \frac{b_2 - b_1}{2} & -\frac{b_2 - b_1}{2} \\ -\frac{b_2 - b_1}{2} & 6r - (b_1 + b_2) + \frac{b_1 - b_2}{2} \end{bmatrix}.$$

Evaluating these expressions at $\beta_1 = -\frac{2}{3} - \frac{1}{6}r$ and $\beta_2 = \frac{2}{3} + \frac{1}{6}r$, we find that the gradient vanishes $\nabla_{\boldsymbol{b}} G \mid_{\boldsymbol{b}=\boldsymbol{\beta}} = 0$ and the Hessian is

$$\nabla^2_{\boldsymbol{b}} G \mid_{\boldsymbol{b}=\boldsymbol{\beta}} = \frac{1}{6r}\begin{bmatrix} \frac{35}{6}r - \frac{2}{3} & -\frac{2}{3} - \frac{1}{6}r \\ -\frac{2}{3} - \frac{1}{6}r & \frac{37}{6}r + \frac{2}{3} \end{bmatrix}.$$

When $(\frac{9\sqrt{2}}{2} - \frac{1}{4})r > 1$ or equivalently $\eta_{\min} < \frac{9\sqrt{2}}{2} - \frac{1}{4}$, the Hessian is positive definite, so $\boldsymbol{\beta}$ is a local minimum of G. Moreover, one can verify that $G(\boldsymbol{\beta}) < G(\boldsymbol{\beta}^*)$, so $\boldsymbol{\beta}$ is not a global minimum. We see that the spurious local minimum $\boldsymbol{\beta}$ does not have the structure described in Theorem 3, as $\boldsymbol{\beta}$ involves a 2-fit-3 association.

The second example shows that in higher dimensions, there exists a local minimum $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ such that $\boldsymbol{\beta}_1$ *approximately* equals $\boldsymbol{\beta}_1^*$, and $\boldsymbol{\beta}_2$ *approximately* equals $(\boldsymbol{\beta}_2^* + \boldsymbol{\beta}_3^*)/2$ — a structure guaranteed by Theorem 3 — but neither approximation is exact. Therefore, the non-zero approximation errors that appear in Theorem 3, is necessary in general.

Example 2 (Approximation Errors)**.** *Consider the Stochastic Ball model with $k = 3$ in dimension $d = 2$, where the true cluster centers are $\boldsymbol{\beta}_1^* = (-1, 0)$, $\beta_2^* = (0, 0)$ and $\beta_3^* = (1, 0)$, where the radius r of the balls satisfies $r \geq \frac{1}{4}$. Let $\boldsymbol{\beta}$ be a candidate solution with $\boldsymbol{\beta}_1 = (-1, 0)$, $\boldsymbol{\beta}_2 = (\frac{1}{2}, 0)$ and $\boldsymbol{\beta}_3$ sufficiently far away from the origin. See the left panel of Figure 7 for an illustration.*

As in Example 1, here the only effective variables are $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Assume first that $r = \frac{1}{4}$. In this case, the Voronoi boundary $\partial_{1,2}$ is at $x_1 = -\frac{1}{4} = \boldsymbol{\beta}_2^* - r$, the left boundary of the second true cluster. We claim that $\boldsymbol{\beta}$ is a local minimum of G; the proof is deferred to the end of this section. Now, let us increase the radius r by a sufficient small amount, in which case the objective function becomes $\widetilde{G}$. By the continuity, there exists a local minimum $\widetilde{\boldsymbol{\beta}}$ of $\widetilde{G}$ near the original local minima $\boldsymbol{\beta}$. Recall that by Lemma 2, $\widetilde{\boldsymbol{\beta}}_1$ and $\widetilde{\boldsymbol{\beta}}_2$ must lie at the center of mass of their Voronoi sets $\mathcal{V}_1(\widetilde{\boldsymbol{\beta}})$ and $\mathcal{V}_2(\widetilde{\boldsymbol{\beta}})$, respectively. It is then not hard to see that the new Voronoi boundary $\widetilde{\partial}_{1,2}$ corresponding to $\widetilde{\boldsymbol{\beta}}$ necessarily intersects the interior of the second true cluster $\mathbb{B}_2$. It follows that $\mathcal{V}_1(\widetilde{\boldsymbol{\beta}}) = \mathbb{B}_1 \cup D$ and $\mathcal{V}_2(\widetilde{\boldsymbol{\beta}}) = (\mathbb{B}_2 \cup \mathbb{B}_3) \setminus D$ for some subset $D \subset \mathbb{B}_2$ with a positive measure. Applying Lemma 2 again, we conclude that $\widetilde{\boldsymbol{\beta}}_1$ is close but not equal to $\boldsymbol{\beta}_1^*$, and that $\widetilde{\boldsymbol{\beta}}_2$ is close but not equal to $(\boldsymbol{\beta}_1^* + \boldsymbol{\beta}_2^*)/2$.

Proof of the claim. Let $t \in (0, 1/8)$ be a sufficiently small number, and $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^2$ be two arbitrary vectors satisfying $\|\boldsymbol{v}_1\|^2 + \|\boldsymbol{v}_2\|^2 = 1$. Consider perturbing $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ to $\boldsymbol{\beta}_1' = \boldsymbol{\beta}_1 + t\boldsymbol{v}_1$ and $\boldsymbol{\beta}_2' = \boldsymbol{\beta}_2 + t\boldsymbol{v}_2$, respectively. Since Voronoi sets only change by a small amount when the perturbation t is small, we find that $\Delta^{\boldsymbol{v}}_{2\to 1}(t) \subseteq \mathbb{B}_2$ is the only set of points that change their association from one Voronoi set to another; see the right panel of Figure 7. Using the expression (10) for the directional k-means objective, we can write $G(\boldsymbol{\beta}')$ as

$$\begin{aligned} G(\boldsymbol{\beta}') =& H^{\boldsymbol{v}}(t) \\ =& \frac{1}{3}\Big[\int_{\mathcal{V}_1(\boldsymbol{\beta})\cap\mathbb{B}_1} \|\boldsymbol{x} - \boldsymbol{\beta}_1 - t\boldsymbol{v}_1\|^2 \mathrm{d}\boldsymbol{x} \\ &+ \int_{\mathcal{V}_2(\boldsymbol{\beta})\cap(\mathbb{B}_2\cup\mathbb{B}_3)} \|\boldsymbol{x} - \boldsymbol{\beta}_2 - t\boldsymbol{v}_2\|^2 \mathrm{d}\boldsymbol{x}\Big] \\ &+ \frac{1}{3}\int_{\Delta^{\boldsymbol{v}}_{2\to 1}(t)} \big(\|\boldsymbol{x} - \boldsymbol{\beta}_1 - t\boldsymbol{v}_1\|^2 - \|\boldsymbol{x} - \boldsymbol{\beta}_2 - t\boldsymbol{v}_2\|^2\big)\mathrm{d}\boldsymbol{x}. \end{aligned}$$

A quick calculation shows that

$$\begin{aligned} & G(\boldsymbol{\beta}') - G(\boldsymbol{\beta}) \\ =& H^{\boldsymbol{v}}(t) - H^{\boldsymbol{v}}(0) \\ =& \frac{1}{3}\Big[t^2 - \underbrace{\int_{\Delta^{\boldsymbol{v}}_{2\to 1}(t)} \Big(\|\boldsymbol{x} - \boldsymbol{\beta}_2 - t\boldsymbol{v}_1\|^2 - \|\boldsymbol{x} - \boldsymbol{\beta}_1 - t\boldsymbol{v}_2\|^2\Big)\mathrm{d}\boldsymbol{x}}_{K(t)}\Big]. \end{aligned}$$

We decompose the term $K(t)$ as follows:

$$\begin{aligned} K(t) =& \int_{\Delta^{\boldsymbol{v}}_{2\to 1}(t)} \underbrace{\Big(\|\boldsymbol{x} - \boldsymbol{\beta}_2\|^2 - \|\boldsymbol{x} - \boldsymbol{\beta}_1\|^2\Big)}_{\kappa_1}\mathrm{d}\boldsymbol{x} \\ &+ \int_{\Delta^{\boldsymbol{v}}_{2\to 1}(t)} \underbrace{t^2\Big(\|\boldsymbol{v}_2\|^2 - \|\boldsymbol{v}_1\|^2\Big)}_{\kappa_2}\mathrm{d}\boldsymbol{x} \\ &+ \int_{\Delta^{\boldsymbol{v}}_{2\to 1}(t)} \underbrace{2t\Big(\langle \boldsymbol{v}_1 - \boldsymbol{v}_2, \boldsymbol{x}\rangle + \langle \boldsymbol{v}_2, \boldsymbol{\beta}_2\rangle - \langle \boldsymbol{v}_1, \boldsymbol{\beta}_1\rangle\Big)}_{\kappa_3}\mathrm{d}\boldsymbol{x} \end{aligned}$$

For all $\boldsymbol{x} \in \Delta^{\boldsymbol{v}}_{2\to 1}(t) \subseteq \mathcal{V}_2(\boldsymbol{\beta})$, we have $\|\boldsymbol{x} - \boldsymbol{\beta}_2\| \leq \|\boldsymbol{x} - \boldsymbol{\beta}_1\|$, hence $\kappa_1 \leq 0$. We also have $\kappa_2 \leq t^2$ since $\|\boldsymbol{v}_2\|^2 - \|\boldsymbol{v}_1\|^2 \leq \|\boldsymbol{v}_1\|^2 \leq 1$. To bound κ_3 we observe that $\langle \boldsymbol{v}_1 - \boldsymbol{v}_2, \boldsymbol{x}\rangle \leq \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|\|\boldsymbol{x}\| \leq 4r = 1$ for all $\boldsymbol{x} \in \Delta^{\boldsymbol{v}}_{2\to 1}(t) \subseteq \mathbb{B}_2$, $\langle \boldsymbol{\beta}_2, \boldsymbol{v}_2\rangle \leq \|\boldsymbol{\beta}_2\|\|\boldsymbol{v}_2\| \leq \frac{1}{2}$, and $\langle \boldsymbol{\beta}_1, \boldsymbol{v}_1\rangle \leq \|\boldsymbol{\beta}_1\|\|\boldsymbol{v}_1\| \leq 1$; it follows that $\kappa_3 \leq 5t$. Combining pieces, we obtain that

$$G(\boldsymbol{\beta}') - G(\boldsymbol{\beta}) \geq \big[t^2 - 6t \cdot \mathrm{Vol}(\Delta^{\boldsymbol{v}}_{2\to 1}(t))\big]/3. \tag{53}$$

It remains to control the volume of the set $\Delta^{\boldsymbol{v}}_{2\to 1}(t)$, which is illustrated in Figure 8. Observe that the distance between the old mid point $(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2)/2$ and the new Voronoi boundary $\partial_{1,2}(\boldsymbol{\beta}')$ can be bounded as

$$\begin{aligned} d_1 :=& \operatorname{dist}\left(\frac{\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2}{2}, \partial_{1,2}(\boldsymbol{\beta}')\right) \\ \leq& \left\|\frac{\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2}{2} - \frac{\boldsymbol{\beta}_1' + \boldsymbol{\beta}_2'}{2}\right\| \\ \leq& t. \end{aligned}$$

Moreover, the (unsigned) angle ψ between the old and new Voronoi boundaries $\partial_{1,2}(\boldsymbol{\beta})$ and $\partial_{1,2}(\boldsymbol{\beta}')$ satisfies

$$\tan\psi = \left|\frac{t(v_{2,2} - v_{1,2})}{2 + t(v_{2,1} - v_{1,1})}\right| \leq \frac{t}{1 - t} \leq 2t.$$

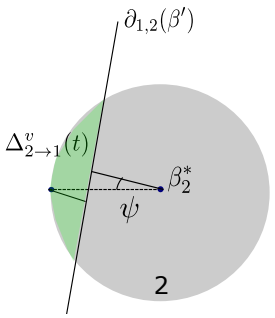

Figure 8. Illustration of $\Delta^{\boldsymbol{v}}_{2\to1}(t)$. The new boundary $\partial_{1,2}(\boldsymbol{\beta}')$ has an angle ψ with the original Voronoi boundary $\partial_{1,2}(\boldsymbol{\beta})$.

From these two observations and the fact that $r = 1/4$, elementary geometry shows that the distance d_2 between $\boldsymbol{\beta}^*_2$ and $\partial_{1,2}(\boldsymbol{\beta}')$ satisfies

$$\begin{aligned} d_2 =& r\cos\psi - d_1 \\ =& \frac{r}{\sqrt{1+\tan^2\psi}} - d_1 \geq \frac{r}{1+2t} - 4r \geq r(1-6t), \end{aligned}$$

whence

$$\mathrm{Vol}(\Delta^{\boldsymbol{v}}_{2\to1}(t)) \leq 2\cdot\sqrt{r^2-d_2^2}\cdot(r-d_2) \leq 12t\sqrt{12t}.$$

Combining with equation (53) shows that $G(\boldsymbol{\beta}') > G(\boldsymbol{\beta})$ when t is sufficiently small. As this inequality holds for arbitrary perturbation direction $(\boldsymbol{v}_1, \boldsymbol{v}_2)$, we conclude that $\boldsymbol{\beta}$ is a local minimum of G. □

References

[1] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[2] S. Dasgupta, "The hardness of k-means clustering," Department of Computer Science and Engineering, University of California, San Diego, Tech. Rep., 2008.

[3] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The planar k-means problem is NP-hard," in *International Workshop on Algorithms and Computation*. Springer, 2009, pp. 274–285.

[4] P. Awasthi, M. Charikar, R. Krishnaswamy, and A. K. Sinop, "The hardness of approximation of Euclidean k-means," in *31st International Symposium on Computational Geometry. arXiv preprint arXiv:1502.03316*, 2015, pp. 754–767.

[5] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[6] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[7] D. Steinley, "Local optima in k-means clustering: what you don't know may hurt you." *Psychological Methods*, vol. 8, no. 3, p. 294, 2003.

[8] ——, "Profiling local optima in k-means clustering: Developing a diagnostic technique." *Psychological methods*, vol. 11, no. 2, p. 178, 2006.

[9] K. Chaudhuri, S. Dasgupta, and A. Vattani, "Learning mixtures of Gaussians using the k-means algorithm," *arXiv preprint arXiv:0912.0086*, 2009.

[10] J. Xu, D. J. Hsu, and A. Maleki, "Global analysis of expectation maximization for mixtures of two Gaussians," in *Advances in Neural Information Processing Systems*, 2016, pp. 2676–2684.

[11] C. Daskalakis, C. Tzamos, and M. Zampetakis, "Ten steps of em suffice for mixtures of two gaussians," in *Conference on Learning Theory*. PMLR, 2017, pp. 704–710.

[12] C. Jin, Y. Zhang, S. Balakrishnan, M. J. Wainwright, and M. I. Jordan, "Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences," in *Advances in Neural Information Processing Systems*, 2016, pp. 4116–4124.

[13] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[14] L. Bottou and Y. Bengio, "Convergence properties of the k-means algorithms," in *Advances in Neural Information Processing Systems*, 1995, pp. 585–592.

[15] N. Srebro, "Are there local maxima in the infinite-sample likelihood of Gaussian mixture estimation?" in *International Conference on Computational Learning Theory*, 2007, pp. 628–629.

[16] D. Steinley, "K-means clustering: a half-century synthesis," *British Journal of Mathematical and Statistical Psychology*, vol. 59, no. 1, pp. 1–34, 2006.

[17] A. Kumar, Y. Sabharwal, and S. Sen, "A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions," in *Annual Symposium on Foundations of Computer Science*, vol. 45. IEEE Computer Society Press, 2004, pp. 454–462.

[18] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "A local search approximation algorithm for k-means clustering," *Computational Geometry*, vol. 28, no. 2-3, pp. 89–112, 2004.

[19] G. W. Milligan, "An examination of the effect of six types of error perturbation on fifteen clustering algorithms," *Psychometrika*, vol. 45, no. 3, pp. 325–342, 1980.

[20] S. Har-Peled and B. Sadri, "How fast is the k-means method?" *Algorithmica*, vol. 41, no. 3, pp. 185–202, 2005.

[21] D. Arthur and S. Vassilvitskii, "How slow is the k-means method?" in *Symposium on Computational Geometry*, vol. 6, no. 32, 2006, pp. 1–10.

[22] A. Kumar and R. Kannan, "Clustering with spectral norm and the k-means algorithm," in *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE, 2010, pp. 299–308.

[23] Y. Lu and H. H. Zhou, "Statistical and computational guarantees of Lloyd's algorithm and its variants," *arXiv preprint arXiv:1612.02099*, 2016.

[24] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the Eighteenth*

Annual ACM-SIAM Symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[25] R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy, "The effectiveness of Lloyd-type methods for the k-means problem," *Journal of the ACM (JACM)*, vol. 59, no. 6, p. 28, 2012.

[26] S. Dasgupta and L. Schulman, "A probabilistic analysis of EM for mixtures of separated, spherical Gaussians," *Journal of Machine Learning Research*, vol. 8, no. Feb, pp. 203–226, 2007.

[27] J. Peng and Y. Xia, "A new theoretical framework for k-means-type clustering," in *Foundations and Advances in Data Mining*. Springer, 2005, pp. 79–96.

[28] J. Peng and Y. Wei, "Approximating k-means-type clustering via semidefinite programming," *SIAM Journal on Optimization*, vol. 18, no. 1, pp. 186–205, 2007.

[29] E. Elhamifar, G. Sapiro, and R. Vidal, "Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery," in *Advances in Neural Information Processing Systems*, 2012, pp. 19–27.

[30] A. Nellore and R. Ward, "Recovery guarantees for exemplar-based clustering," *Information and Computation*, vol. 245, pp. 165–180, 2015.

[31] P. Awasthi, A. S. Bandeira, M. Charikar, R. Krishnaswamy, S. Villar, and R. Ward, "Relax, no need to round: Integrality of clustering formulations," in *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*. ACM, 2015, pp. 191–200.

[32] X. Li, Y. Li, S. Ling, T. Strohmer, and K. Wei, "When do birds of a feather flock together? k-means, proximity, and conic programming," *Mathematical Programming*, vol. 179, no. 1-2, pp. 295–341, 2020.

[33] Y. Fei and Y. Chen, "Hidden integrality of SDP relaxation for sub-Gaussian mixture models," in *Conference on Learning Theory (COLT)*, 2018, arXiv preprint arXiv:1803.06510.

[34] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[35] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the EM algorithm: From population to sample-based analysis," *The Annals of Statistics*, vol. 45, no. 1, pp. 77–120, 2017.

[36] W. Qian, Y. Zhang, and Y. Chen, "Global convergence of least squares EM for demixing two log-concave densities," in *Advances in Neural Information Processing Systems*, 2019, pp. 4795–4803.

[37] J. Kwon, W. Qian, C. Caramanis, Y. Chen, and D. Davis, "Global convergence of the EM algorithm for mixtures of two component linear regression," in *Conference on Learning Theory*, 2019, pp. 2055–2110.

[38] J. Xu, D. J. Hsu, and A. Maleki, "Benefits of overparameterization with EM," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 662–10 672.

[39] X. Yi and C. Caramanis, "Regularized EM algorithms: A unified framework and statistical guarantees," in *Advances in Neural Information Processing Systems*, 2015, pp. 1567–1575.

[40] J. M. Klusowski, D. Yang, and W. D. Brinda, "Estimating the coefficients of a mixture of two linear regressions by expectation maximization," *IEEE Transactions on Information Theory*, 2019.

[41] B. Yan, M. Yin, and P. Sarkar, "Convergence of gradient EM on multi-component mixture of Gaussians," in *Advances in Neural Information Processing Systems*, 2017, pp. 6956–6966.

[42] J. Kwon and C. Caramanis, "EM converges for a mixture of many linear regressions," *arXiv preprint arXiv:1905.12106*, 2019.

[43] R. Kannan, H. Salmasian, and S. Vempala, "The spectral method for general mixture models," in *International Conference on Computational Learning Theory*. Springer, 2005, pp. 444–457.

[44] S. Mei, Y. Bai, and A. Montanari, "The landscape of empirical risk for non-convex losses," *The Annals of Statistics*, vol. 46, no. 6A, pp. 2747–2774, 2018.

[45] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Advances in Neural Information Processing Systems*, 2016, pp. 3873–3881.

[46] I. Safran and O. Shamir, "Spurious local minima are common in two-layer ReLU neural networks," in *International Conference on Machine Learning*, 2018, pp. 4430–4438.

[47] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Advances in Neural Information Processing Systems 29*, 2016.

[48] G. H. Ball and D. J. Hall, "PROMENADE – an on-line pattern recognition system," Stanford Research Institution, Menlo Park, CA, Tech. Rep., 1967.

[49] M. M. Astrahan, "Speech analysis by clustering, or the hyperphoneme method," Stanford University, Department of Computer Science, Tech. Rep., 1970.

[50] A. R. Barakbah and A. Helen, "Optimized k-means: an algorithm of initial centroids optimization for k-means," in *Proc. Seminar on Soft Computing, Intelligent System, and Information Technology (SIIT), Surabaya*, 2005.

[51] A. R. Barakbah and Y. Kiyoki, "A pillar algorithm for k-means optimization by distance maximization for initial centroid designation," in *2009 IEEE Symposium on Computational Intelligence and Data Mining*. IEEE, 2009, pp. 61–68.

[52] P. Fränti and S. Sieranoja, "K-means properties on six clustering benchmark datasets," *Applied Intelligence*, vol. 48, no. 12, pp. 4743–4759, 2018.

[53] R.-D. Buhai, A. Risteski, Y. Halpern, and D. Sontag, "Benefits of overparameterization in single-layer latent variable generative models," *arXiv preprint arXiv:1907.00030*, 2019.

[54] R. Dwivedi, N. Ho, K. Khamaru, M. I. Jordan, M. J. Wainwright, and B. Yu, "Singularity, misspecification, and the convergence rate of EM," *arXiv preprint arXiv:1810.00828*, 2018.

[55] Y. Zhang, H.-W. Kuo, and J. Wright, "Structured local minima in sparse blind deconvolution," in *Advances in*

Neural Information Processing Systems 31, 2018, pp. 2328–2337.

[56] M. J. Wainwright, *High-Dimensional Statistics: A Non-asymptotic Viewpoint*. Cambridge University Press, 2019, vol. 48.

Wei Qian received her Ph.D. and M.S. degrees in Operations Research from Cornell University in 2020, and B.S. degree in mathematics from The University of Michigan, Ann Arbor in 2014. Her research work lie in machine learning, reinforcement learning and optimization, with applications in transportation systems.

Yuqian Zhang is an Assistant Professor with the Department of Electrical and Computer Engineering at Rutgers University. She was a postdoctoral scholar with the Tripods Center for Data Science at Cornell University. She obtained her Ph.D. and M.S. in Electrical Engineering from Columbia University, and B.S. in Information Engineering from Xi'an Jiaotong University. Her research leverages physical models in data driven computations, convex and nonconvex optimization, solving problems in machine learning, computer vision, signal processing.

Yudong Chen is an Associate Professor with the School of Operations Research and Information Engineering at Cornell University. He obtained his Ph.D. degree in Electrical and Computer Engineering in 2013 from The University of Texas at Austin, and M.S. and B.S. degrees in Control Science and Engineering from Tsinghua University. He was a postdoctoral scholar in the Electrical Engineering and Computer Sciences department at the University of California, Berkeley from 2013 to 2015. He has served as area chairs for AAAI, AISTATS and NeurIPS, and received a National Science Foundation CAREER award. His research work lies in machine learning, reinforcement learning, high-dimensional statistics, and optimization, with applications in network scheduling, wireless communication, and financial systems.