Local Minima Structures in Gaussian Mixture Models

Yudong Chen[1], Dogyoon Song[2], Xumei Xi[3], and Yuqian Zhang[4]

[1]Department of Computer Sciences, University of Wisconsin-Madison
[2]Department of Electrical Engineering and Computer Science, University of Michigan
[3]School of Operations Research and Information Engineering, Cornell University
[4]Department of Electrical and Computer Engineering, Rutgers University

March 12, 2024

Abstract

We investigate the landscape of the negative log-likelihood function of Gaussian Mixture Models (GMMs) with a general number of components in the population limit. As the objective function is non-convex, there can exist multiple spurious local minima that are not globally optimal, even for well-separated mixture models. Our study reveals that all local minima share a common structure that partially identifies the cluster centers (i.e., means of the Gaussian components) of the true location mixture. Specifically, each local minimum can be represented as a non-overlapping combination of two types of sub-configurations: (1) fitting a single mean estimate to multiple Gaussian components or (2) fitting multiple estimates to a single true component. These results apply to settings where the true mixture components satisfy a certain separation condition, and are valid even when the number of components is over- or under-specified. We also present a more fine-grained analysis for the setting of one-dimensional GMMs with three components, which provide sharper approximation error bounds with improved dependence on the separation parameter.

Contents

1   Introduction

Mixture models, such as the Gaussian mixture model (GMM), are a class of latent variable models that offer a flexible approach to approximate complex multi-modal distributions. These models have been widely used in statistical inference with heterogeneous data. A standard method for estimating the parameters of GMM is maximum likelihood estimation, which seeks the global minimum of the negative log-likelihood function of the model. The statistical properties of the maximum likelihood estimators, including its asymptotic consistency [38] and finite-sample error rates [10, 29, 21], have been well studied.

However, estimating GMMs poses significant computational challenges, and the extent of these challenges is not fully understood. The negative log-likelihood function of GMM is non-convex and generally has multiple local minima. Consequently, standard iterative algorithms, such as the Expectation-Maximization (EM) [15], are only guaranteed to converge to a local minimum [41, 23]. A recent study [22] shows that for GMMs with three or more well-separated components, there provably exist spurious local minima that may be arbitrarily far from the global minimum in both Euclidean distance and in likelihood values; moreover, randomly initialized EM algorithms converge to a spurious local minimum with high probability. For certain special cases of GMMs, such as those with two equally weighted components, it has been shown that the negative log-likelihood function in fact has no spurious local minimum, and the EM converges to the global minimum from arbitrary initialization [14, 43]. These global convergence results, however, are the exceptions rather than the norm as demonstrated in [22]. Despite the negative theoretical results, iterative methods such as the EM and its variants are routinely applied to GMMs in practice.

Motivated by the above challenges and ubiquity of iterative methods, in this paper we perform a more fine-grained investigation of the likelihood landscape and local minima structures of GMMs. Specifically, we seek to answer the following question:

> *"Do spurious local minima possess informative structures related to the global minimum?"*

1.1   Our Contributions

We consider the problem of estimating the component means of a GMM with a general number of equally weighted components. Suppose that one fits a mixture of k Gaussians with means $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ to data generated by a true mixture of k_* Gaussians with true means $\boldsymbol{\Theta}^* = (\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{k_*}^*)$. We investigate the structures of the local minima of the negative log-likelihood function in the population limit (i.e., the sample size $n \to \infty$).

Our main contribution is a proof that *all* local minima $\boldsymbol{B}$ of the negative likelihood share a similar structure that partially identifies the means $\boldsymbol{\Theta}^*$ of the true mixture model. Specifically, each local minimum only involves two types of sub-configurations: either a single estimated center $\boldsymbol{\beta}_i$ is close to the average of several true component means $\{\boldsymbol{\theta}_s^*\}$, or several estimated centers $\{\boldsymbol{\beta}_i\}$ are close to a single true center $\boldsymbol{\theta}_s^*$. Moreover, these sub-configurations involve disjoint sets of estimates and component means. Notably, this result holds even when the number of components in the fitting model, k, is different from the number of components in the true model, k_*.

To illustrate the above structure, let us consider an example scenario where $k = 5$ and $k_* = 4$. A local minimum $\boldsymbol{B}$ of the negative log-likelihood has the form

$$\boldsymbol{\beta}_1 \approx \frac{1}{2}\left(\boldsymbol{\theta}_1^* + \boldsymbol{\theta}_2^*\right), \boldsymbol{\beta}_2 \approx \boldsymbol{\beta}_3 \approx \boldsymbol{\beta}_4 \approx \boldsymbol{\theta}_3^* \text{ and } \boldsymbol{\beta}_5 \approx \boldsymbol{\theta}_4^*. \tag{1}$$

The relationship between $\boldsymbol{B}$ and $\boldsymbol{\Theta}^*$ can be summarized by the association graph given in Figure 1. This association graph can be decomposed into three disjoint subgraphs, each of which is a star graph containing only one mean estimate $\boldsymbol{\beta}_i$, or only one true component mean $\boldsymbol{\theta}_s^*$. We refer to the first type of subgraphs

3

as "one-fits-many" (one $\boldsymbol{\beta}_i$ fits many $\boldsymbol{\theta}_s^*$'s), and the second type "many-fit-one". The main theorem in this paper, Theorem 2, states that *all* local minima of the population negative log-likelihood exhibit a similar combinatorial structure (potentially plus some unassociated estimates). In particular, each local minimum $\boldsymbol{B}$ corresponds to a *disjoint* union of complete bipartite graphs $(\mathcal{T}_a, \mathcal{S}_a)_{a=1,2,\dots}$ (potentially plus a degenerate pair $(\mathcal{T}_0, \mathcal{S}_0)$ with $\mathcal{S}_0 = \emptyset$); each bipartite graph is between a subset of estimated centers $\mathcal{T}_a \subseteq [k]$ and a subset of true component means $\mathcal{S}_a \subseteq [k_*]$, such that at least one of $\mathcal{T}_a$ and $\mathcal{S}_a$ is a singleton, as shown in Figure 1. Moreover, Theorem 2 provides upper bounds on the approximation errors in equation (1), characterizing how the errors depend on the separation between the true mixture components.
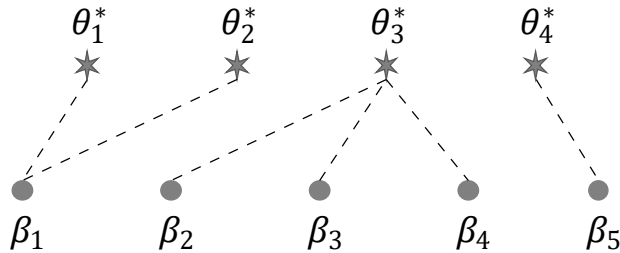
Figure 1: Association between mean estimates $\{\boldsymbol{\beta}_i\}$ and true component means $\{\boldsymbol{\theta}_s^*\}$ in a local minimum.

The above results apply to GMMs with any values of k and k_*. Specializing to the exemplar setting of one-dimensional GMM with $k_* = 3$ components, we show that sharper results can be obtained via a more fine-grained analysis. In particular, we show that the approximation errors diminish at a quadratic exponential rate with respect to the separation between true component means (Theorem 3).

The above findings complement the negative results in [22] on the existence of spurious local minima arbitrarily far from the global optimum $\boldsymbol{\Theta}^*$ *in Euclidean distance*. Our results show that all local minima are close to $\boldsymbol{\Theta}^*$ in a different sense, as they partially recover the structure of $\boldsymbol{\Theta}^*$. Consequently, standard iterative algorithms such as the EM algorithm and gradient methods, starting from an arbitrary initial solution (except from a set of measure zero [41, 23, 26]), converge to a solution that is informative of $\boldsymbol{\Theta}^*$.

While there has been a long line of research on algorithms and theoretical guarantees for learning GMMs, we emphasize that our focus in this paper is not on obtaining quantitative improvements on these results. Rather, we aim to understand the structural properties of the non-convex likelihood function and iterative methods for optimizing the likelihood. Nevertheless, we expect that our structural results will have important implications for developing better algorithms, especially in the overparameterized regime where $k \gg k_*$; we elaborate on this point in Section 5.

1.2 Related Work

In 2006, Srebro posed the question of whether the population negative log-likelihood function of GMMs has spurious local minima [34]. In 2016, Jin *et al.* answered this question in the negative for the general case with $k_* \geq 3$ [22]. Motivated by the computational considerations in estimating GMMs, recent works seek to understand the finer properties of the likelihood function as well as those of the EM algorithm—arguably the most popular algorithm for GMMs.

One line of work investigates the *local* behaviors of the likelihood in a neighborhood around the global optimum, which is relevant to the EM starting from a sufficiently good initial solution. The work in [3] proposes a general framework for establishing the local geometric convergence of the EM; implicit in their results is that the negative log-likelihood function of a two-component GMM has no other local minima near the global minimum. The extension to multiple-component settings is considered in the work [46]. Further work in this line studies GMMs with additional structures [48, 40, 20], the EM with unknown mixture weights and covariances [9], confidence intervals constructed using the EM [11], and the setting where the number of components is under-specified [19].

Another line of work studies *global* properties of GMMs and the EM in certain restricted settings, mostly that with $k_* = 2$ (typically equally weighted) components. In this setting, the work in [14, 43] proves that

the EM initialized at a random solution converges to the global minimum, showing that the negative log-likelihood function has no spurious local minimum that is not globally optimal in this case. This fact is further investigated in the work [28], which proposes a general framework for transferring the properties of the population likelihood function to its empirical counterpart. Extensions to mixtures of two log-concave distributions [32] or two linear regressions [25] have also been considered. A more recent set of papers study the delicate behaviors of the EM when the two components have small or no separation, or when the number of components is mis-specified [18, 17, 24, 42]; in these settings, the EM may exhibit a slower, non-parametric statistical error rate. The above global results for $k_* = 2$ usually do not generalize to $k_* \geq 3$, as spurious local minima provably exist in the latter [22]. Moreover, additional spurious local minima may arise when the mixture weights are not equal [44].

Our recent work [33] studies the related problem of optimizing the (non-smooth) k-means objective function, which can be viewed as a limit of the log-likelihood function of GMMs when the posited variance of Gaussian components goes to zero. The results in [33] and this paper are similar in spirit: spurious local minima provably exist but possess additional hidden structures. However, the theoretical results in this paper are substantially sharper and apply to the over/under-parameterized regime $k \neq k_*$. The proofs in this paper rely on quite different techniques and are also considerably simpler than [33] in several aspects, taking advantage of the rich structures in the smooth log-likelihood objective function for GMMs.

We remark that several recent works [27, 16] have developed new algorithms for k-means and GMMs that explicitly leverage the one-fits-many/many-fit-one structures studied in [33] and this paper.

1.3   Organization

This paper is organized as follows. In Section 2, we provide formal definitions and notation that are used throughout the paper. In Section 3, we present preliminary analytical results for the likelihood landscape of GMMs. Section 4 contains the main technical results of this work. Specifically, in Section 4.1, we state the structures of local minima, and in Section 4.2, we discuss their implications. Additionally, in Section 4.3, we explore how we can improve the approximation error bounds in the main theorem in the example setting of a one-dimensional GMM. Section 5 provides discussion on the our results and future directions.

The remaining sections are dedicated to technical details and proofs. In particular, Sections 6, 7 and 8 contain the proofs for the theoretical results in Sections 3, 4.1 and 4.3, respectively. We defer subsidiary items, such as proofs of technical lemmas, to the Appendix.

2   Problem setup

In this section, after introducing the basic notation, we describe the problem setup of Gaussian mixture models and the associated maximum likelihood estimation approach.

2.1   Notation

We let $\mathbb{N}_+$ denote the set of positive integers, and $\mathbb{N} := \mathbb{N}_+ \cup \{0\}$. For each $m \in \mathbb{N}_+$, let $[m] := \{1, 2, \ldots, m\}$ and $[m]_0 := [m] \cup \{0\}$. Given a set $\mathcal{I} \subseteq \mathbb{N}_+$, let $|\mathcal{I}|$ denote the cardinality of $\mathcal{I}$ and $\mathcal{I}(i)$ denote the i-th smallest integer in $\mathcal{I}$, so $\mathcal{I}(1) < \cdots < \mathcal{I}(|\mathcal{I}|)$. Let $\mathbb{R}$ denote the real numbers and $\mathbb{R}_+ := \{x \in \mathbb{R} : x > 0\}$. We generally use curly letters (e.g., $\mathcal{T}$) to denote sets. For a set $\mathcal{T} \subset \mathbb{R}^d$, let $\operatorname{int} \mathcal{T}$ denote its interior.

We use boldface lowercase letters, e.g., $\boldsymbol{x}$, to denote deterministic (column) vectors, of which x_i is the i-th element. In particular, $\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ is the i-th standard basis vector in $\mathbb{R}^d$, and $\mathbf{1}_n = (1, 1, \ldots, 1) \in \mathbb{R}^d$ is the vector of all ones. We let $\|\boldsymbol{x}\|$ denote the Euclidean ℓ_2 norm of $\boldsymbol{x}$.

We use boldface capital letters, e.g., $\boldsymbol{X}$, to denote matrices. In particular, $\boldsymbol{I}_d$ is the d-by-d identity matrix. We identify a matrix with an ordered set of column vectors, and let $\boldsymbol{X}_j \in \mathbb{R}^d$ denote the j-th column vector of $\boldsymbol{X}$. For any $\boldsymbol{X} \in \mathbb{R}^{d \times m}$ and any $\mathcal{I} \subseteq [m]$, we let $\boldsymbol{X}(\mathcal{I}) \in \mathbb{R}^{d \times |\mathcal{I}|}$ denote a matrix such that $\boldsymbol{X}(\mathcal{I})_i = \boldsymbol{X}_{\mathcal{I}(i)}$. We write $\boldsymbol{X} \succeq 0$ when $\boldsymbol{X}$ is symmetric positive semidefinite.

We use sans-serif letters, e.g., x and X, to denote random variables and vectors. For a probability measure ν, we write $\mathsf{x} \sim \nu$ to denote that x is distributed per ν. For any $\boldsymbol{u} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ such that $\boldsymbol{\Sigma} \succeq 0$, we denote by $\mathcal{N}(\boldsymbol{u}, \boldsymbol{\Sigma})$ the Gaussian distribution with mean $\boldsymbol{u}$ and covariance $\boldsymbol{\Sigma}$; moreover, we let $\phi(\cdot \mid \boldsymbol{u}, \boldsymbol{\Sigma})$ denote the probability density function of $\mathcal{N}(\boldsymbol{u}, \boldsymbol{\Sigma})$. With an underlying probability measure specified, we

use $\mathbb{E}$ and $\mathbb{P}$ to denote the expectation and probability. Given a sample space Ω and an event $A \subseteq \Omega$, the indicator function $\mathbb{1}_A : \Omega \to \{0, 1\}$ is defined such that $\mathbb{1}_A(x) = 1$ if and only if $x \in A$. We sometimes use $\mathbb{1}(x \in A)$ to denote $\mathbb{1}_A(x)$.

2.2 Gaussian Mixture Models

Consider a mixture of k_* equally weighted Gaussian distributions, denoted by ν^*, which has the density

$$f^*(\cdot) = \frac{1}{k_*} \sum_{s \in [k_*]} f_s^*(\cdot), \tag{2}$$

where $f_s^*(\cdot)$ denotes the density of the s-th mixture component, which is a Gaussian distribution. Specifically, we assume that $f_s^*(\cdot) := \phi(\cdot \mid \boldsymbol{\theta}_s^*, \sigma^2 \boldsymbol{I}_d)$ for some $\boldsymbol{\theta}_s^* \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}_+$. Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\} \subset \mathbb{R}^d$ be a collection of n i.i.d. samples generated from ν^*.

The goal of Gaussian mixture modeling is to fit a k-component mixture of Gaussians, denoted by ν, to the data $\boldsymbol{X}$. Assuming σ is known a priori, we may write the density of ν as

$$f(\cdot) = \frac{1}{k} \sum_{i \in [k]} f_i(\cdot), \tag{3}$$

where $f_i(\cdot) := \phi(\cdot \mid \boldsymbol{\beta}_i, \sigma^2 \boldsymbol{I}_d)$ for some $\boldsymbol{\beta}_i \in \mathbb{R}^d, i \in [k]$. We refer to $\{\boldsymbol{\beta}_i, \ i \in [k]\}$ as the *fitted centers*, or *mean estimates*. Note that the number of fitted components k may differ from the number of true components k_*. This covers the exact-parametrization ($k = k_*$), over-parametrization ($k > k_*$) and under-parametrization ($k < k_*$) settings.

To avoid cluttered notation, we have suppressed the dependence of f^* and $\{f_s^*\}$ on $\{\boldsymbol{\theta}_s^*\}$, and likewise for the dependence of f and $\{f_i\}$ on $\{\boldsymbol{\beta}_i\}$. We use $\mathbb{E}_*$ and $\mathbb{P}_*$ to denote the expectation and probability, respectively, under the true mixture model ν^*. Similarly, for each $s \in [k_*]$, we use $\mathbb{E}_s$ and $\mathbb{P}_s$ to denote the expectation and probability, respectively, under the s-th true Gaussian component with density f_s^*. Note that $\mathbb{E}_* = \frac{1}{k_*} \sum_{s \in [k_*]} \mathbb{E}_s$ and $\mathbb{P}_* = \frac{1}{k_*} \sum_{s \in [k_*]} \mathbb{P}_s$ by definition in (2). For clarity, we generally use s, s' to index true mixture components (e.g., $\boldsymbol{\theta}_s^*, f_s^*$), whereas i, j are used to index the components in the fitted model ν (e.g., $\boldsymbol{\beta}_i, f_i$).

For the true mixture model ν^*, we define the maximum and minimum component separations as

$$\begin{aligned} \Delta_{\max} &:= \max_{s, s' \in [k_*]} \left\| \boldsymbol{\theta}_s^* - \boldsymbol{\theta}_{s'}^* \right\|, \\ \Delta_{\min} &:= \min_{\substack{s, s' \in [k_*] \\ s \neq s}} \left\| \boldsymbol{\theta}_s^* - \boldsymbol{\theta}_{s'}^* \right\|. \end{aligned} \tag{4}$$

We refer to $\Delta_{\min}/\sigma$ as the *Signal-to-Noise Ratio* (SNR). Lastly, we define the condition number of ν^* as

$$\rho := \frac{\Delta_{\max}}{\Delta_{\min}}. \tag{5}$$

2.3 Maximum Likelihood Estimation

Given the data $\boldsymbol{X}$ from the true model ν^*, the maximum likelihood principle is a standard approach for fitting a model ν to $\boldsymbol{X}$. Let $\boldsymbol{\Theta}^* = (\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_{k_*}^*) \in \mathbb{R}^{d \times k_*}$ and $\boldsymbol{B} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_k) \in \mathbb{R}^{d \times k}$ denote the component mean parameters of the true and fitted models, respectively. The population negative log-likelihood function L—the infinite sample limit (i.e., $n \to \infty$) of the sample negative log-likelihood—is given by

$$\begin{aligned} L(\boldsymbol{B}) &\equiv L(\boldsymbol{B} \mid \boldsymbol{\Theta}^*) \\ &= -\mathbb{E}_* \left[\log f(\mathsf{x}) \right] \\ &= D_{\mathrm{KL}} \left(f^* \| f \right) - \mathbb{E}_* \left[\log f^*(\mathsf{x}) \right], \end{aligned} \tag{6}$$

6

where x is a random variable distributed per ν^*, $D_{\mathrm{KL}}\left(f^*\|f\right) := \mathbb{E}_*\left[\log\frac{f^*(\mathsf{X})}{f(\mathsf{X})}\right]$ is the Kullback-Leibler (KL) divergence between ν^* and ν, and the entropy term $-\mathbb{E}_*\left[\log f^*(\mathsf{x})\right]$ does not depend on $\boldsymbol{B}$. The maximum likelihood approach involves finding $\widehat{\boldsymbol{B}}$ that minimizes the negative log-likelihood, i.e.,

$$\widehat{\boldsymbol{B}} \in \arg\min_{\boldsymbol{B}\in\mathbb{R}^{d\times k}} L(\boldsymbol{B}). \tag{7}$$

When $k = k_*$, it is clear from (6) and the non-negativity of KL divergence that $\boldsymbol{\Theta}^*$ is a global minimizer of L. However, L is generally not convex and has local minima other than $\boldsymbol{\Theta}^*$ [22]. This fact poses a significant computational challenge for solving (7) because standard algorithms (such as the EM and gradient methods) are only guaranteed to find a local minimum, which can be substantially suboptimal.

Remark 1 (Euclidean invariance of L). It is easy to verify that the function L is invariant under a Euclidean transformation, i.e., rotations, translations, reflections, and a sequence thereof. That is, for any orthonormal matrix $\boldsymbol{U} \in \mathbb{R}^{d\times d}$ and vector $\boldsymbol{v} \in \mathbb{R}^d$, the following holds:

$$L\left(\boldsymbol{U}\boldsymbol{B} + \boldsymbol{v}\otimes\mathbf{1}_{k_*} \mid \boldsymbol{U}\boldsymbol{\Theta}^* + \boldsymbol{v}\otimes\mathbf{1}_{k_*}\right) = L(\boldsymbol{B}\mid\boldsymbol{\Theta}^*)$$

where $\boldsymbol{v}\otimes\mathbf{1}_{k_*} = \begin{bmatrix}\boldsymbol{v} & \cdots & \boldsymbol{v}\end{bmatrix} \in \mathbb{R}^{d\times k_*}$. In the analysis we frequently make use of this invariance property to choose a convenient coordinate system.

3 Preliminary Analysis

In this section, we present a preliminary analysis for the landscape of the likelihood function of Gaussian mixture models. In Section 3.1, we characterize the stationary points of the population negative log-likelihood function L. In Section 3.2, we derive an equivalent form of the stationary condition (Theorem 1), which immediately implies several useful properties of the stationary points of L. These results will be used in our analysis in subsequent sections.

3.1 Coefficients of Association and Optimality Conditions

Here we define the coefficients of association, which are then used to characterize the first-order and the second-order optimality conditions for L as well as the fixed points of the EM and gradient methods. Most of the materials in this subsection are not new, at least in the exact parameterization setting ($k = k_*$). Nevertheless, we collect these results here to set the stage for subsequent analysis.

3.1.1 Coefficients of Association

The coefficient of association, defined below, measures the relative strength of association between a data point $\boldsymbol{x}$ and each of the k estimated components $\{f_i : i \in [k]\}$ in the fitted mixture model.

Definition 1. *Let* $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{B} = (\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k) \in \mathbb{R}^{d\times k}$ *and* $\sigma \in \mathbb{R}_+$ *be given. For each* $i \in [k]$, *the* coefficient of association *between* $\boldsymbol{x}$ *and* $\boldsymbol{\beta}_i$ *at level* σ *is defined as*

$$\psi_i(\boldsymbol{x}) \equiv \psi_i(\boldsymbol{x};\boldsymbol{B},\sigma)$$

$$:= \frac{\frac{1}{k}f_i(\boldsymbol{x})}{f(\boldsymbol{x})} = \frac{\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{\beta}_i\|^2}{2\sigma^2}\right)}{\sum_{j\in[k]}\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{\beta}_j\|^2}{2\sigma^2}\right)}. \tag{8}$$

The association coefficient $\psi_i(\boldsymbol{x})$ takes the form of a soft (arg)min function and can be viewed as an approximation of the indicator function of $\boldsymbol{\beta}_i$ being the closest to $\boldsymbol{x}$ among all the fitted centers $\{\boldsymbol{\beta}_j : j \in [k]\}$; that is,

$$\psi_i(\boldsymbol{x}) \approx \mathbb{1}\left\{i = \arg\min_{j\in[k]}\|\boldsymbol{x}-\boldsymbol{\beta}_j\|^2\right\}.$$

The analysis in this paper makes use of this intuitive interpretation. From a Bayesian perspective, one may also interpret $\psi_i(\boldsymbol{x})$ as the posterior probability of a data point $\boldsymbol{x}$ belonging to the i-th fitted component,

given the current estimate $\boldsymbol{B}$ of the component means. As such, the quantity $\psi_i(\boldsymbol{x})$ appears in the E-step of the EM algorithm, as we shall see momentarily. With x denoting a random data point generated from the true distribution ν^*, the induced scalar random variables

$$\Psi_i := \psi_i(\mathsf{x}), \quad i \in [k], \tag{9}$$

correspond to the association coefficients between the random vector x and $\boldsymbol{B} = (\boldsymbol{\beta}_i : i \in [k])$.

3.1.2 First-order and Second-order Characterizations of Optimality for L

Here we discuss the optimality conditions that characterize the local minimizers of the population negative log-likelihood function L defined (6).

First-order necessary condition The gradient of L can be expressed in terms of the association coefficients, as stated in the following lemma.

Lemma 1. *Let* $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) \in \mathbb{R}^{d \times k}$ *and* $\sigma > 0$. *The partial derivatives of L at $\boldsymbol{B}$ admit the expression*

$$\frac{\partial}{\partial \boldsymbol{\beta}_i} L(\boldsymbol{B}) = \frac{1}{\sigma^2} \cdot \mathbb{E}_* \left[\Psi_i \cdot (\boldsymbol{\beta}_i - \mathsf{x}) \right], \quad i \in [k]. \tag{10}$$

It follows from Lemma 1 that $\boldsymbol{B}$ is a stationary point of L (i.e., $\nabla L(\boldsymbol{B}) = 0$) if and only if

$$\mathbb{E}_* \left[\Psi_i (\boldsymbol{\beta}_i - \mathsf{x}) \right] = 0, \quad \forall i \in [k], \tag{11}$$

which is equivalent to

$$\boldsymbol{\beta}_i = \frac{\mathbb{E}_* \left[\Psi_i \mathsf{x} \right]}{\mathbb{E}_* \left[\Psi_i \right]}, \qquad \forall i \in [k]. \tag{12}$$

Equations (11) and (12) are the first-order necessary condition for $\boldsymbol{B}$ being a local minimizer of for L.

Second-order necessary condition We can also write the Hessian of L using the coefficients of association, as done in the next lemma.

Lemma 2. *Let* $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) \in \mathbb{R}^{d \times k}$ *and* $\sigma > 0$. *The second-order partial derivatives of L at $\boldsymbol{B}$ admit the expression*

$$\begin{aligned} \frac{\partial^2}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_j} L(\boldsymbol{B}) &= \frac{1}{\sigma^4} \cdot \mathbb{E}_* \left[\Psi_i \Psi_j \cdot (\boldsymbol{\beta}_j - \mathsf{x})(\boldsymbol{\beta}_\mathsf{i} - \mathsf{x})^\top \right] \\ &\quad + \delta_{ij} \cdot \frac{1}{\sigma^4} \cdot \mathbb{E}_* \left[\Psi_i \cdot \left\{ \sigma^2 \cdot \boldsymbol{I}_d - (\boldsymbol{\beta}_i - \mathsf{x})(\boldsymbol{\beta}_\mathsf{i} - \mathsf{x})^\top \right\} \right] \end{aligned} \tag{13}$$

for all $i, j \in [k]$, *where δ_{ij} is the Kronecker delta (i.e., $\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ otherwise).*

The Hessian $\nabla^2 L(\boldsymbol{B})$ can be expressed as a $kd \times kd$ matrix whose (i,j)-th block (of size d-by-d) is $\frac{\partial^2}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_j} L(\boldsymbol{B})$ for $i, j \in [k]$. Any point $\boldsymbol{B}$ that satisfies

$$\nabla L(\boldsymbol{B}) = 0 \quad \text{and} \quad \nabla^2 L(\boldsymbol{B}) \succeq 0 \tag{14}$$

is called a second-order stationary point. Equation (14) is the second-order necessary condition for $\boldsymbol{B}$ being a local minimizer of L.

In summary, we have the following inclusion relation:

$$\{\boldsymbol{B} \in \mathbb{R}^{d \times k} : \boldsymbol{B} \text{ is a local minimizer of } L\} \subseteq \{\boldsymbol{B} \in \mathbb{R}^{d \times k} : \nabla L(\boldsymbol{B}) = 0 \text{ and } \nabla^2 L(\boldsymbol{B}) \succeq 0\}$$
$$\subseteq \{\boldsymbol{B} \in \mathbb{R}^{d \times k} : \nabla L(\boldsymbol{B}) = 0\}.$$

3.1.3 Connection between the Stationarity Condition and the EM Algorithm

The EM algorithm is a popular iterative method for optimizing the likelihood function. In the population setting, the EM update takes the form

$$\boldsymbol{\beta}_i \;\leftarrow\; \frac{\mathbb{E}_*\left[\Psi_i \mathsf{x}\right]}{\mathbb{E}_*\left[\Psi_i\right]} = \boldsymbol{\beta}_i - \frac{1}{\mathbb{E}_*\left[\Psi_i\right]} \cdot \frac{\partial}{\partial \boldsymbol{\beta}_i} L(\boldsymbol{B}), \quad \forall i \in [k], \tag{15}$$

where we have combined the E-step (computing Ψ_i) and the M-step (computing $\boldsymbol{\beta}_i$) into one update. The EM update (15) can be viewed as a fixed point iteration for solving the stationary condition (12), or as a gradient descent-like (or quasi-Newton) algorithm with a coordinate-dependent step size $1/\mathbb{E}_*\left[\Psi_i\right]$ [45]. Therefore, the fixed points of the EM algorithm correspond to the stationary points of L, and the *stable fixed points* of the EM correspond to the local minimizers.

Our results in Section 3.2 provide further characterizations of the stationary points and local minimizers of L. In light of the above discussion, these results immediately apply to the solution returned by the EM algorithm and other local algorithms including gradient descent and Newton methods.

3.2 Properties of the Stationary Points of L

Our first theorem provides a necessary and sufficient condition for the (first-order) stationary points of L. Recall that $\boldsymbol{\Theta}^* = \left(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots, \boldsymbol{\theta}_{k_*}^*\right)$ is the true parameters for the Gaussian mixture model whose density is given in (2).

Theorem 1 (Equivalent stationary condition)**.** *A point $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) \in \mathbb{R}^{d \times k}$ is a stationary point of L if and only if*

$$\sum_{j \in [k]} \boldsymbol{\beta}_j \sum_{s \in [k_*]} \mathbb{E}_s\left[\Psi_i \Psi_j\right] = \sum_{s \in [k_*]} \boldsymbol{\theta}_s^* \mathbb{E}_s\left[\Psi_i\right], \quad \forall i \in [k]. \tag{16}$$

We prove Theorem 1 in Section 6.2 using the Stein's identity.

Equation (16) is equivalent to the original stationary condition (12). However, the expression in (16) is often more useful as it exposes the relationship between the fitted centers $\{\boldsymbol{\beta}_j\}$ and the true centers $\{\boldsymbol{\theta}_s^*\}$. This result plays a key role in establishing the main results of this paper in Section 4.

Equation (16) links the estimates $\boldsymbol{B}$ to the true component centers $\boldsymbol{\Theta}^*$ by establishing a system of equations: for each $i \in [k]$, a convex combination of the estimates $\boldsymbol{\beta}_j$ weighted by $w_j \coloneqq \sum_{s \in [k_*]} \mathbb{E}_s\left[\Psi_i \Psi_j\right]$ is equated to a convex combination of true component centers $\boldsymbol{\theta}_s^*$ weighted by $w_s' \coloneqq \mathbb{E}_s\left[\Psi_i\right]$. This system of equations implicitly characterizes possible configurations of (first-order) stationary points. The four corollaries in Section 3.2.1 derive various useful properties of the stationary points and local minimizers of L from equation (16). Although these corollaries might seem intuitive, proving some of them could be challenging using other methods, particularly when $k \neq k_*$.

3.2.1 General Geometric Properties of Stationary Points

The result in the first corollary is probably well known. It states that for any stationary point of L, the weighted average of the fitted centers must equal the mean of the true centers.

Corollary 1 (Mean consistency)**.** *If $\boldsymbol{B} \in \mathbb{R}^{d \times k}$ is a stationary point of L, then*

$$\sum_{j \in [k]} \mathbb{E}_*\left[\Psi_j\right] \cdot \boldsymbol{\beta}_j = \frac{1}{k_*} \sum_{s \in [k_*]} \boldsymbol{\theta}_s^*.$$

Proof of Corollary 1. Adding up the equation (16) over $i \in [k]$ and using the fact that $\sum_{i \in [k]} \Psi_i = 1$ surely, we obtain that $\sum_{j \in [k]} \boldsymbol{\beta}_j \sum_{s \in [k_*]} \mathbb{E}_s\left[\Psi_j\right] = \sum_{s \in [k_*]} \boldsymbol{\theta}_s^*$. Since $\sum_{s \in [k_*]} \mathbb{E}_s\left[\Psi_j\right] = k_* \mathbb{E}_*\left[\Psi_j\right]$, the proof is complete. $\qquad\square$

The next corollary states that any stationary point of L must lie in the linear subspace spanned by the true component centers.

Corollary 2 (Linear span). *If $\boldsymbol{B} \in \mathbb{R}^{d \times k}$ is a stationary point of L, then we have*

$$\boldsymbol{\beta}_i \in \text{span}\left\{\boldsymbol{\theta}_s^*, s \in [k_*]\right\}, \quad i \in [k].$$

With this property, we can restrict ourselves in subsequent analysis to span $\{\boldsymbol{\theta}_s^*\}$, a k_*-dimensional subspace of $\mathbb{R}^d$. This property is particularly useful when k_* is much smaller than d. The proof of Corollary 2 is deferred to Section 6.3.

3.2.2 Two Extreme Cases of Fitting Gaussian Mixtures

We consider two extreme cases of the problem of fitting a mixture model of k Gaussians to data generated by a mixture of k_* Gaussians: the case with $k_* = 1$, and the case with $k = 1$. Theorem 1 readily identifies the stationary points in each of these cases. Specifically, the population log-likelihood L has a unique stationary point that is the global minimizer, when we fit multiple Gaussians to a single one (Corollary 3), or a single Gaussian to multiple ones (Corollary 4).

As we show shortly in Section 4, these two settings are the atomic cases of the general setting with arbitrary k and k_*. In particular, any local minimizer of L can be decomposed into a non-overlapping collection of the global minimizers of sub-problems in these two settings (plus a collection of vectors that are almost irrelevant, if any); see Theorem 2 and Section 4.2.

Case 1 ($k_* = 1$)   If the true model has only one component with mean $\boldsymbol{\theta}_1^*$, then L has a unique stationary point corresponding to the true center, regardless of the number k of fitted centers. There are no other local or global minimizers.

Corollary 3 (Fitting k Gaussians to one Gaussian). *If $k_* = 1$, then L has a unique stationary point $\boldsymbol{B} \in \mathbb{R}^{d \times k}$ such that*

$$\boldsymbol{\beta}_i = \boldsymbol{\theta}_1^*, \quad \forall i \in [k].$$

Proof of Corollary 3. Without loss of generality we may assume that $\boldsymbol{\theta}_1^* = \mathbf{0}$ (see Remark 1). If $\boldsymbol{B}$ is a stationary point of L, then Corollary 2 implies that $\boldsymbol{\beta}_i \in \text{span}\{\boldsymbol{\theta}_1^*\} = \{\mathbf{0}\}, \forall i \in [k]$. Conversely, if $\boldsymbol{\beta}_i = \mathbf{0}$ for all i, then $\psi_i(\boldsymbol{x}) = \frac{1}{k}$ for all $\boldsymbol{x} \in \mathbb{R}^d$ and the stationary condition (12) is satisfied by $\boldsymbol{B}$. $\square$

Corollary 3 is related to a recent line of work in [18, 17, 42] on the setting where the number of components in the mixture is over-specified (i.e., $k > k_*$). In the canonical over-specified setting where one fits a mixture of $k = 2$ Gaussians to data from a single Gaussian, the work above showed that the EM algorithm converges towards the true center from random initialization (albeit with a slower convergence rate and a larger statistical error $(d/n)^{1/4}$ than in the exact-specified setting). At the population level, Corollary 3 provides a more general result, applicable to any number $k \geq 1$ of specified components and any descent algorithms beyond the EM.

Case 2 ($k = 1$)   As a sanity check, we consider an under-specified setting with $k = 1$ and $k_* \geq 1$, that is, fitting a single Gaussian to a mixture of multiple Gaussians. In this case, we have $\psi_1(\boldsymbol{x}) = 1, \forall \boldsymbol{x} \in \mathbb{R}^d$, hence equation (16) immediately implies the following result (its proof is trivial):

Corollary 4 (Fitting One Gaussian to k_* Gaussians). *If $k = 1$, then L has a unique stationary point $\boldsymbol{B} = (\boldsymbol{\beta}_1) \in \mathbb{R}^{d \times 1}$ satisfying*

$$\boldsymbol{\beta}_1 = \frac{1}{k_*} \sum_{s \in [k_*]} \boldsymbol{\theta}_s^* = \mathbb{E}_*[\mathsf{x}].$$

We thus recover the elementary fact that the Maximum Likelihood Estimator of fitting a single Gaussian to a dataset is given by the sample mean.

4   Main Results: Combinatorial Structures of Local Minima

In this section, we present the main results of this work. Section 4.1 gives the main theorem (Theorem 2), which characterizes the common structures shared by all local minima of L with an arbitrary number of mixture components, $k_*, k \geq 1$. In Section 4.2, we discuss the implications of this result by connecting it to the decomposability of mixture learning problem instances. In Section 4.3, we show that the approximation error bounds in Theorem 2 can be exponentially improved in the example setting of a one-dimensional three-component mixture of Gaussian distributions.

4.1   Main Theorem Statement

To state the main theorem, we define the Voronoi cells induced by a solution $\boldsymbol{B}$ and the associated index function.

Definition 2 (Voronoi cells). *Let $\boldsymbol{B} = (\boldsymbol{\beta}_i)_{i=1}^{k} \in \mathbb{R}^{d \times k}$ and $i \in [k]$. The i-th Voronoi cell of $\boldsymbol{B}$ is defined as*

$$
\begin{aligned}
\mathcal{V}_i = \mathcal{V}_i(\boldsymbol{B}) &\coloneqq \left\{ \boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{\beta}_i\| \leq \|\boldsymbol{x} - \boldsymbol{\beta}_j\|, \forall j \in [k] \right\} \\
&= \left\{ \boldsymbol{x} \in \mathbb{R}^d : \psi_i(\boldsymbol{x}) \geq \psi_j(\boldsymbol{x}), \forall j \in [k] \right\}.
\end{aligned}
\tag{17}
$$

The index function induced by $\boldsymbol{B}$ is a function $\iota_{\boldsymbol{B}} : \mathbb{R}^d \to [k]$ such that

$$
\iota_{\boldsymbol{B}}(\boldsymbol{x}) \coloneqq \min \left\{ i \in [k] : \boldsymbol{x} \in \mathcal{V}_i(\boldsymbol{B}) \right\}.
$$

The maximum width of the Voronoi cells relative to $\boldsymbol{B}$ is defined as

$$
D_{\mathrm{cell}}(\boldsymbol{B}) \coloneqq \max_{j \in [k]} \max_{s \in [k_*] : \boldsymbol{\theta}_s^* \in \mathcal{V}_j} \|\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_j\|.
$$

Note that when $\{\boldsymbol{\beta}_i\}$ are distinct, the Voronoi cells partition the ambient subspace $\mathbb{R}^d$ (up to a set of measure zero). Also let $d_{\mathrm{eff}} \coloneqq \dim \operatorname{span} \{\boldsymbol{\theta}_s^* : s \in [k_*]\}$ denote the effective dimension of the true mixture ν^* (cf. Corollary 2); note that $d_{\mathrm{eff}} \leq \min\{d, k_*\}$. Finally, recall the signal-to-noise ratio $\Delta_{\min}/\sigma$ and the condition number $\rho = \Delta_{\max}/\Delta_{\min}$ defined in equations (4) and (5).

We are ready to present the main theorem of this paper.

Theorem 2 (Main theorem). *Let $\boldsymbol{B} \in \mathbb{R}^{d \times k}$ be a local minimizer of $L(\,\cdot\,|\,\boldsymbol{\Theta}^*)$. If*

$$
\frac{\Delta_{\min}}{\sigma} > 72(\sqrt{2\pi} + 1) \cdot k_* \cdot k^4,
\tag{18}
$$

and[1]

$$
\frac{\Delta_{\min}}{\sigma} > \sqrt{72(\sqrt{2\pi} + 1) \cdot k_*^2 \cdot k^3 \cdot (5k_* + 2k) \cdot \frac{D_{\mathrm{cell}}(\boldsymbol{B})}{\sigma}},
\tag{19}
$$

then there exist $q \in \mathbb{N}$, and two collections of sets $\mathbb{T} \coloneqq \{\mathcal{T}_a \subseteq [k] : a \in [q] \cup \{0\}\}$ and $\mathbb{S} \coloneqq \{\mathcal{S}_a \subseteq [k_*] : a \in [q]\}$, for which the following properties hold.

1. *(Simple partitions) There exist q_0 with $0 \leq q_0 \leq q$ such that the following properties hold:*

    (a) $\mathbb{T}$ *is a partition of* $[k]$.

    (b) $\mathbb{S}$ *is a partition of* $[k_*]$.

    (c) $|\mathcal{T}_a| = 1$ *for all* $a \in \{1, 2, \ldots, q_0\}$.

    (d) $|\mathcal{T}_a| \geq 2$ *and* $|\mathcal{S}_a| = 1$ *for all* $a \in \{q_0 + 1, \ldots, q\}$.

2. *(Mutual exclusiveness) Suppose that $i \in \mathcal{T}_a$ and $j \in \mathcal{T}_b$ for $a, b \in [q] \cup \{0\}$ with $a \neq b$. If $i \neq j$, then $\boldsymbol{\beta}_i \neq \boldsymbol{\beta}_j$.*

---

[1]Note that the SNR condition in (19) includes a solution-dependent quantity, $D_{\mathrm{cell}}(\boldsymbol{B})$, tied to the local minimizer $\boldsymbol{B}$. In 4.2.1, we discuss replacing this condition using a coarse universal upper bound, $D_{\mathrm{cell}}(\boldsymbol{B}) \leq \Delta_{\max} + \sigma k_* \sqrt{d_{\mathrm{eff}}}$, which eliminates the solution dependence but leads to a more restrictive SNR condition. We speculate a tighter universal upper bound for $D_{\mathrm{cell}}(\boldsymbol{B})$ is attainable and leave it as an open question.

3. *(Approximation error)*

   (a) *(one-fits-many) For each $a \in \{1, \cdots, q_0\}$,*

   $$\frac{1}{\sigma} \left\| \boldsymbol{\beta}_{i_a} - \frac{1}{|\mathcal{S}_a^\delta|} \sum_{s \in \mathcal{S}_a^\delta} \boldsymbol{\theta}_s^* \right\| \lesssim \left(\frac{D_{\text{cell}}(\boldsymbol{B})}{\sigma} \cdot k_*^2 \cdot k^3 \cdot (k_* + k) \right)^{1/2} + k_* \cdot d_{\text{eff}}^{1/2}, \qquad (20)$$

   where i_a denotes the unique element in $\mathcal{T}_a^\delta$.

   (b) *(many-fit-one) For each $a \in \{q_0 + 1, \cdots, q\}$,*

   $$\frac{1}{\sigma} \left\| \boldsymbol{\beta}_i - \boldsymbol{\theta}_{s_a}^* \right\| \lesssim \left(\frac{D_{\text{cell}}(\boldsymbol{B})}{\sigma} \cdot k_*^2 \cdot k^3 \cdot (k_* + k) \right)^{1/2}, \quad \forall i \in \mathcal{T}_a^\delta, \qquad (21)$$

   where s_a denotes the unique element in $\mathcal{S}_a^\delta$.

4. *(Non-association) Let $a \in \{1, \ldots, q\}$. If $s \in \mathcal{S}_a$ and $i \in \mathcal{T}_0$, then*

$$\mathbb{E}_s \left[\Psi_i \right] \lesssim \left(\frac{k^5 \cdot k_*}{k_* + k} \cdot \frac{\sigma}{\Delta_{\min}} \right)^{1/2}, \quad and$$

$$\mathbb{P}_s \left(\mathcal{V}_i \right) \lesssim \left(\frac{k^5 \cdot k_*}{k_* + k} \cdot \frac{\sigma}{\Delta_{\min}} \right)^{1/2}. \qquad (22)$$

Theorem 2 formalizes the structural results discussed in Section 1.1 and Figure 1: all local minimizers of L involves disjoint one-fit-many and many-fit-one associations between $\{\boldsymbol{\beta}_i\}$ and $\{\boldsymbol{\theta}_s^*\}$ (plus potentially some non-associations). In Section 4.2 below, we provide detailed discussions on the interpretation and implication of Theorem 2, as well the quantitative aspects and proof ideas of the theorem.

4.2 Implications and Discussions of Theorem 2

The main message of Theorem 2 is the following: as far as the local minimizers of the negative log-likelihood is concerned, fitting a GMM can be decomposed into multiple smaller subproblems that involve simple GMMs. To further explain this result, let us elaborate.

Mixture fitting problems   Let $k, k_* \in \mathbb{N}$ and $\boldsymbol{\Theta}^* = \left(\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{k_*}^* \right)$. We let $\texttt{GMM}(k, \boldsymbol{\Theta}^*)$ denote the problem instance of fitting k mean estimates to a true GMM with means $\boldsymbol{\Theta}^*$. We denote a *global* minimizer of $\texttt{GMM}(k, \boldsymbol{\Theta}^*)$ by

$$\widehat{\boldsymbol{B}}^{k, \boldsymbol{\Theta}^*} \coloneqq \arg \min_{\boldsymbol{B} \in \mathbb{R}^{d \times k}} L(\boldsymbol{B} \mid \boldsymbol{\Theta}^*) \qquad (23)$$

where the population negative log-likelihood L is defined in (6). In Section 3.2.2 we discussed two cases where the global minimizer is unique and simple:

- If $k_* = 1$, namely, $\boldsymbol{\Theta}^* = (\boldsymbol{\theta}_1^*)$, then $\widehat{\boldsymbol{B}}_i^{k, \boldsymbol{\Theta}^*} = \boldsymbol{\theta}_1^*$ for all $i \in [k]$.

- If $k = 1$, then $\widehat{\boldsymbol{B}}^{k, \boldsymbol{\Theta}^*} = \left(\hat{\boldsymbol{\beta}} \right)$ where $\hat{\boldsymbol{\beta}} = \frac{1}{k_*} \sum_{s \in [k_*]} \boldsymbol{\theta}_s^*$.

Implications of Theorem 2   For general values of k and k_*, Theorem 2 states that *all* local minimizers $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ of L possess a common combinatorial structure. Specifically, each of the fitted centers $\boldsymbol{\beta}_i$ ($i \in [k]$) must correspond to one of the three possibilities below, depending on which $\mathcal{T}_a$ the index i belongs to. Recall that for any $\boldsymbol{X} \in \mathbb{R}^{d \times m}$ and any $\mathcal{I} \subseteq [m]$, we let $\boldsymbol{X}(\mathcal{I}) \in \mathbb{R}^{d \times |\mathcal{I}|}$ denote a submatrix of X such that $\boldsymbol{X}(\mathcal{I})_i = \boldsymbol{X}_{\mathcal{I}(i)}$.

- *(One-fits-many)* First, suppose that $i \in \mathcal{T}_a$ for some $a \in [q_0]$, in which case $\boldsymbol{B}(\mathcal{T}_a) = \boldsymbol{\beta}_i$ is a singleton. With the notation defined in (23), the approximation bound (20) states that

$$\boldsymbol{B}(\mathcal{T}_a) = \boldsymbol{\beta}_i \approx \frac{1}{|\boldsymbol{\Theta}^*(\mathcal{S}_a)|} \sum_{s \in \mathcal{S}_a} \boldsymbol{\theta}_s^* = \widehat{\boldsymbol{B}}^{|\mathcal{T}_a|, \boldsymbol{\Theta}^*(\mathcal{S}_a)}, \tag{24}$$

where the last equality follows from Corollary 4. In other words, $\boldsymbol{\beta}_i$ is close to the mean of true centers in $\boldsymbol{\Theta}^*(\mathcal{S}_a)$, which in turn is the global minimizer of the one-fits-many subproblem $\mathtt{GMM}(1, \boldsymbol{\Theta}^*(\mathcal{S}_a))$.

- *(Many-fit-one)* Second, suppose that $i \in \mathcal{T}_a$ for some $a \in [q] \setminus [q_0]$, in which case $\boldsymbol{\Theta}^*(\mathcal{S}_a) = (\boldsymbol{\theta}_{s_a}^*)$ is a singleton. The approximation bound (21) states that

$$\boldsymbol{B}(\mathcal{T}_a) = (\boldsymbol{\beta}_i)_{i \in \mathcal{T}_a} \approx (\boldsymbol{\theta}_{s_a}^*, \ldots, \boldsymbol{\theta}_{s_a}^*) = \widehat{\boldsymbol{B}}_i^{|\mathcal{T}_a|, \boldsymbol{\Theta}^*(\mathcal{S}_a)}, \tag{25}$$

where the last equality follows from Corollary 3. In other words, the fitted centers $\{\boldsymbol{\beta}_i\}_{i \in \mathcal{T}_a}$ are all close to the true center $\boldsymbol{\theta}_{s_a}^*$, which in turn is the global minimizer of the many-fit-one subproblem $\mathtt{GMM}(k, \boldsymbol{\theta}_{s_a}^*)$. Here we remark "one-fits-one" is subsumed under "many-fit-one" configurations.

- *(Non-association)* Third, if $i \in \mathcal{T}_0$, then the inequalities in (22) state that $\boldsymbol{\beta}_i$ is not strongly associated with any of the true component means $\boldsymbol{\theta}_s^*$, $s \in [k_*]$. This implies that most of the data points generated from the true mixture model ν^* are far away from $\boldsymbol{\beta}_i$ in comparison with other fitted centers $\boldsymbol{\beta}_j$, $\forall j \in [k_*] \setminus \mathcal{T}_0$. In other words, $\boldsymbol{\beta}_i$ is virtually *not* used at all to fit any of the k_* components in the true mixture ν^*.

Moreover, the simple partition property (Theorem 2, Claim 1) ensures that every $i \in [k]$ belong to one and only one of $\mathcal{T}_a$, $a \in [q]_0$. As a result, the three configurations, namely, one-fits-many, many-fit-one, and non-associations, indeed exhaust all possibilities.

We highlight the symmetry between equations (24) and (25). Therefore, one can informally summarize the implications of Theorem 2 in terms of the decomposability of a given mixture problem instance. If $\boldsymbol{B}$ is any local minimizer of L, then there exist partitions $\mathbb{T} = \{\mathcal{T}_a \subseteq [k] : a \in [q]_0\}$ and $\mathbb{S} = \{\mathcal{S}_a \subseteq [k_*] : a \in [q]\}$ such that

$$\boldsymbol{B}(\mathcal{T}_a) \approx \widehat{\boldsymbol{B}}^{|\mathcal{T}_a|, \boldsymbol{\Theta}^*(\mathcal{S}_a)}, \qquad \forall a \in [q]$$

where for all $a \in [q]$, either $|\mathcal{T}_a| = 1$ or $|\mathcal{S}_a| = 1$ holds. That is, any local minimizer of L—identified as a set of vectors—can be decomposed into the global solutions of simple sub-problems in the 'one-fits-many' and 'many-fit-one' settings (up to some approximation error and plus a collection of non-associated estimates). While not mathematically precise, this decomposability property can be written schematically as

$$\mathtt{GMM}(k, \boldsymbol{\Theta}^*) \approx \bigoplus_{a=1}^{q} \mathtt{GMM}(|\mathcal{T}_a|, \boldsymbol{\Theta}^*(\mathcal{S}_a)),$$

with the understanding that the decomposition is with respect to a local minimizer of the negative log-likelihood function, and different minimizers lead to different decompositions.

We illustrate these results from Theorem 2 with an example. Suppose that we are fitting $k = 6$ estimated centers to data generated by a mixture of $k_* = 5$ Gaussians with true component means $\boldsymbol{\Theta}^* = (\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_5^*)$. A minimizer $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_6) \in \mathbb{R}^{d \times 6}$ of L correspond to $q_0 = 2$, $q = 3$ and the partitions $\mathbb{T} = \{\mathcal{T}_a : a \in [q]_0\}$ and $\mathbb{S} = \{\mathcal{S}_a : a \in [q]\}$ with

$$\mathcal{T}_0 = \{1, 2\}, \qquad \mathcal{T}_1 = \{3\}, \qquad \mathcal{T}_2 = \{4\}, \qquad \mathcal{T}_3 = \{5, 6\},$$
$$\mathcal{S}_1 = \{1, 2, 3\}, \qquad \mathcal{S}_2 = \{4\}, \qquad \mathcal{S}_3 = \{5\}.$$

The association between $\{\boldsymbol{\beta}_i\}$ and $\{\boldsymbol{\theta}_s^*\}$ is depicted by the bipartite graph in Figure 2. The graph can be partitioned into three types of star graphs indexed by: (i) $\mathcal{T}_0$ (non-associated centers); (ii) $(\mathcal{T}_1, \mathcal{S}_1)$ and $(\mathcal{T}_2, \mathcal{S}_2)$ (one-fits-many); (iii) $(\mathcal{T}_3, \mathcal{S}_3)$ (many-fit-one).

We highlight that our paper focuses on characterizing local minima structure, distinguishing sub-GMM components, and providing upper bounds on approximation errors and cross-association. We note that our results do not specify the quantity or sizes of these local minima or different GMM problems, i.e., values like $q_0, q, |\mathcal{T}_a|$, etc., are not addressed. Additionally, we conjecture that the non-association set could potentially be empty, but leave its proof as an open question for future exploration.
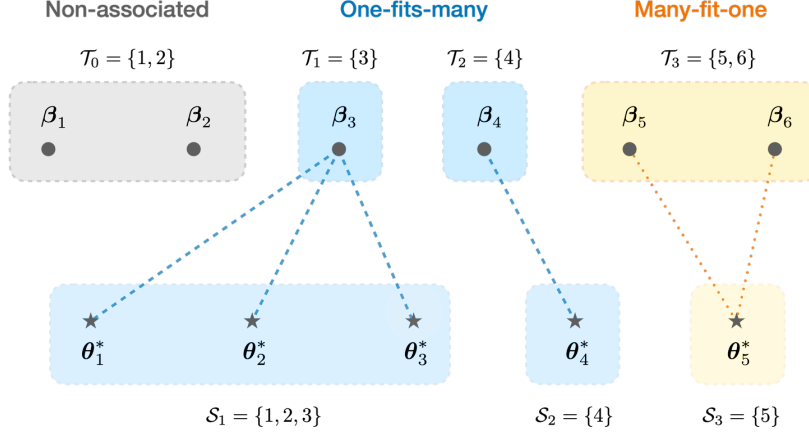
Figure 2: Illustration of the association between fitted centers $\{\boldsymbol{\beta}_i\}$ and true centers $\{\boldsymbol{\theta}_s^*\}$ in a local minimum.

4.2.1 Quantitative Remarks on Theorem 2

We discuss several quantitative aspects of Theorem 2, assuming $k = k_*$ for the convenience of discussion.

Universal upper bound on $D_{\text{cell}}(\boldsymbol{B})$ Theorem 2 and in particular the SNR requirement (19) involve a solution-specific quantity $D_{\text{cell}}(\boldsymbol{B})$ that depends on the local minimizer $\boldsymbol{B}$ under consideration. If one desires, the following universal upper bound for $D_{\text{cell}}(\boldsymbol{B})$ can be used instead. By utilizing the first-order optimality condition (12), we note that

$$
\boldsymbol{\beta}_i = \frac{\sum_{s=1}^{k_*} \mathbb{E}_s\left[\Psi_i \mathsf{x}\right]}{\sum_{s=1}^{k_*} \mathbb{E}_s\left[\Psi_i\right]}
$$
$$
= \frac{\sum_{s=1}^{k_*} \mathbb{E}_s\left[\Psi_i\right] \cdot \boldsymbol{\theta}_s^* + \mathbb{E}_s\left[\Psi_i \cdot (\mathsf{x} - \boldsymbol{\theta}_s^*)\right]}{\sum_{s=1}^{k_*} \mathbb{E}_s\left[\Psi_i\right]}.
$$

Observe that $\frac{\sum_{s=1}^{k_*} \mathbb{E}_s\left[\Psi_i\right] \cdot \boldsymbol{\theta}_s^*}{\sum_{s=1}^{k_*} \mathbb{E}_s\left[\Psi_i\right]}$ is in the convex hull of $\{\boldsymbol{\theta}_s^* : s \in [k_*]\}$. For each $s \in [k_*]$, we have

$$
\left\|\mathbb{E}_s\left[\Psi_i \cdot (\mathsf{x} - \boldsymbol{\theta}_s^*)\right]\right\| \leq \mathbb{E}_s\left[\left\|\Psi_i \cdot (\mathsf{x} - \boldsymbol{\theta}_s^*)\right\|\right]
$$
$$
\leq \mathbb{E}_s\left[\left\|\mathsf{x} - \boldsymbol{\theta}_s^*\right\|\right]
$$
$$
\leq \sigma \sqrt{d_{\text{eff}}}.
$$

Hence, we have $D_{\text{cell}}(\boldsymbol{B}) \leq \Delta_{\max} + \sigma k_* \sqrt{d_{\text{eff}}} \leq \Delta_{\max} + \sigma k_* \sqrt{k_*}$ for all local minimizers $\boldsymbol{B}$. Using this upper bound, the SNR requirement (19) can be simplied to $\Delta_{\min}/\sigma \geq C\rho \cdot k_*^6$ for some universal constant C. We note that this is only a crude upper bound, and we conjecture that a tighter universal upper bound for $D_{\text{cell}}(\boldsymbol{B})$ is attainable.

SNR requirements By the preceding remark, Theorem 2 requires the SNR to satisfy $\Delta_{\min}/\sigma \geq C\rho \cdot k_*^6$. The dependence on k_*^6 is likely an artifact of our analysis due to the use of the Voronoi-cell-based arguments and union bounds; see Propositions 1–3. While this paper does not focus on algorithmic guarantees, we remark that this SNR condition is sufficient for, e.g., spectral methods to learn the true model [39]. We conjecture the conclusions of Theorem 2 would hold as long as $\Delta_{\min}/\sigma$ is on the order of $k_*^{1/2}$ or above, which corresponds to the SNR range where clustering-based mixture-learning methods succeed [12, 39].

Approximation error bounds Observe that the approximation error bounds (20) and (21) are of order $\left(\frac{D_{\text{cell}}(\boldsymbol{B})}{\sigma}\right)^{1/2} \cdot k_*^3$. Here, we make three points interpreting this result.

(a) $\Delta_{\max}$ has little effects on the difficulty of mixture learning problem and our error bounds.

(b) Our analysis of the approximation error yields an upper bound that relies on the "cellular width" $D_{\mathrm{cell}}(\boldsymbol{B})$.
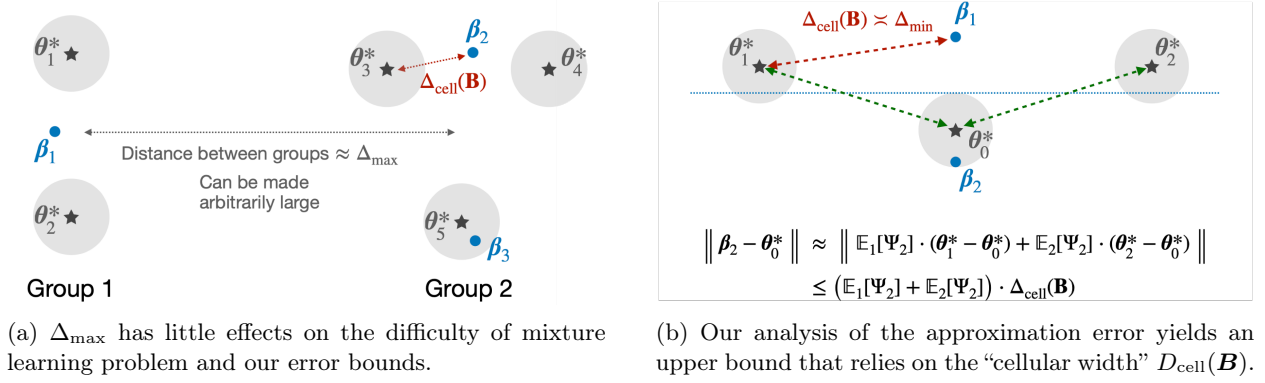
Figure 3: Illustrative examples that support our remarks on Theorem 2.

- Firstly, since $D_{\mathrm{cell}}(\boldsymbol{B}) \lesssim \Delta_{\max}$ (excluding the extra $\sigma k_* \sqrt{d_{\mathrm{eff}}}$ term), the approximation error is upper bounded by $\left(\frac{\Delta_{\max}}{\sigma}\right)^{1/2} \cdot k_*^3$.

- Secondly, this approximation error bound is no greater than $\frac{\Delta_{\min}}{\sigma}$ due to the condition (19). In particular, when $\Delta_{\min}$ is sufficiently large so that $D_{\mathrm{cell}}(\boldsymbol{B}) \cdot k_*^6 \ll \Delta_{\min}^2$, this becomes significantly smaller that $\Delta_{\min}$.

- Lastly, multiplying σ to both sides of (20) and (21) reveals that the approximation becomes exact in the noiseless limit $\sigma \to 0$.

Similarly to the SNR requirements, we believe these error bounds admit room for improvement in terms of their specific scaling with respect to the problem parameters, as showcased in Section 4.3 to follow.

Cross-association bounds We remark that the non-association property in Claim 4 of Theorem 2 actually holds beyond $i \in \mathcal{T}_0$. In Section 7, we present a more general Theorem 4, which establishes this property for $i \notin \mathcal{T}_0$. We did not include this result in the statement of Theorem 2 because Theorem 2 is derived from Theorem 4 by rearranging the (quasi-)partitions $(\mathbb{T}, \mathbb{S})$ via an iterative procedure described in Section 7.3.2 (Algorithm 2), and some of the (i, s) pairs could be affected by this procedure. In fact, the set $\mathcal{T}_0$ (corresponding to non-associated nodes in Figure 2) is likely an artifact of our analysis, and we conjecture that Theorem 2 in fact holds with $\mathcal{T}_0 = \emptyset$.

(In)dependence on problem parameters $\Delta_{\min}$ and $\Delta_{\max}$ Observe that the cross-association bounds (22) decreases as the SNR $\frac{\Delta_{\min}}{\sigma}$ increases. This is expected because the SNR determines the difficulty of separating one component in the mixture from others, and higher SNR makes separating components in the mixture easier. On the other hand, the maximum separation $\Delta_{\max}$ has minimal impact on the problem's difficulty and our error bounds. To illustrate, consider the scenario in Figure 3a with $k = 3$ and $k_* = 5$, where the mixture components $\{\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_5^*\}$ can be divided into two groups that are roughly $\Delta_{\max}$ away, and a local minimizer $\boldsymbol{B}$ satisfies $\boldsymbol{\beta}_1 \approx (\boldsymbol{\theta}_1^* + \boldsymbol{\theta}_2^*)/2$, $\boldsymbol{\beta}_2 \approx (\boldsymbol{\theta}_3^* + \boldsymbol{\theta}_4^*)/2$ and $\boldsymbol{\beta}_3 \approx \boldsymbol{\theta}_5^*$. Even if the distance $\approx \Delta_{\max}$ between the groups becomes significantly large, it barely affects the difficulty of estimating the component centers within each group. Specifically, the approximation errors $\|\boldsymbol{\beta}_2 - (\boldsymbol{\theta}_3^* + \boldsymbol{\theta}_4^*)/2\|$ and $\|\boldsymbol{\beta}_3 - \boldsymbol{\theta}_5^*\|$ are mostly determined by the configurations within group 2, and these errors remain almost constant when the inter-group distance $\Delta_{\max} \to \infty$.

Dependence on the auxiliary parameter $D_{\mathrm{cell}}(\boldsymbol{B})$ The approximation error bounds (20) and (21) depend on the "cellular width" parameter $D_{\mathrm{cell}}(\boldsymbol{B})$. This dependence stems from our proofs of Proposition 4 and Lemma 16. Using the "2-estimates-and-3-centers" scenario in Figure 3b, we outline below the key proof steps that yield the dependence on $D_{\mathrm{cell}}(\boldsymbol{B})$ and argue that this dependence might be unavoidable in general. Using the first-order optimality condition (11), we obtain $\sum_{s=0}^{2} \mathbb{E}_s [\Psi_i(\boldsymbol{\beta}_i - \mathsf{x})] = 0$ for all $i \in \{1, 2\}$. Therefore, the approximation error for $\boldsymbol{\beta}_2$ satisfies $\boldsymbol{\beta}_2 - \boldsymbol{\theta}_0^* \approx \mathbb{E}_1[\Psi_2] \cdot (\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*) + \mathbb{E}_2[\Psi_2] \cdot (\boldsymbol{\theta}_2^* - \boldsymbol{\theta}_0^*)$, because

15

$\mathbb{E}_0[\Psi_2] \approx 1$ and $\mathbb{E}_1[\Psi_2], \mathbb{E}_1[\Psi_2] \approx 0$. Suppose we increase $D_{\text{cell}}(\boldsymbol{B})$ by moving $\boldsymbol{\theta}_1^*$ and $\boldsymbol{\theta}_2^*$ away from each other along the horizontal direction. Note that the association coefficient $\mathbb{E}_1[\Psi_2]$, which is mostly determined by the distance difference $\|\boldsymbol{\beta}_2 - \boldsymbol{\theta}_1^*\|^2 - \|\boldsymbol{\beta}_1 - \boldsymbol{\theta}_1^*\|^2$, remain largely unchanged; similarly for $\mathbb{E}_2[\Psi_1]$. Consequently, the error $\boldsymbol{\beta}_2 - \boldsymbol{\theta}_0^*$ grows with $\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*$ and $\boldsymbol{\theta}_2^* - \boldsymbol{\theta}_0^*$, which in turn grow with $D_{\text{cell}}(\boldsymbol{B})$. While not a formal proof, the above observation suggests that some form of dependence on $D_{\text{cell}}(\boldsymbol{B})$ may be necessary in the presence of "one-fit-many".

4.2.2 Proof Ideas for Theorem 2

Here we outline the key ideas of the proof for Theorem 2, which will be detailed in Section 7. Our analysis centers on the expectation of the product of association coefficients, $\mathbb{E}_s[\Psi_i\Psi_j]$, for $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ with respect to the component measure f_s^*. This quantity reflects the strength of interaction between the estimates $\boldsymbol{\beta}_i$, $\boldsymbol{\beta}_j$ and the true cluster mean $\boldsymbol{\theta}_s^*$. Specifically, $\mathbb{E}_s[\Psi_i\Psi_j]$ is very close to 0 unless $\boldsymbol{\beta}_i, \boldsymbol{\beta}_j$ are both close to $\boldsymbol{\theta}_s^*$. Also recall that the Hessian $\nabla^2 L(\boldsymbol{B})$, whose expression (13) involves $\mathbb{E}_s[\Psi_i\Psi_j]$, must be positive semidefinite at a local minimizer $\boldsymbol{B}$. Combining these two facts allow us to extract structural information of $\boldsymbol{B}$.

We implement the above strategy by quantifying the connection between $\mathbb{E}_s[\Psi_i\Psi_j]$ and $\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\|$, as well as $\mathbb{E}_s[\Psi_i]$ and $\mathbb{P}_s(\mathcal{V}_i)$, through a set of auxiliary propositions (see Section 7.2.1) based on probabilistic and geometric arguments. Collecting these propositions, we argue that every *second-order stationary point* of L must possess a certain set of combinatorial and geometric properties, which depend on a thresholding parameter $\delta > 0$ that is used to distinguish between 'large' and 'small' values of $\mathbb{E}_s[\Psi_i\Psi_j]$ in our analysis. When the signal-to-noise ratio, $\Delta_{\min}/\sigma$, is sufficiently large, we choose a specific value of δ (see Section 7.2.2), and the combinatorial and geometric properties reduce to those presented in Theorem 2.

4.3 Improvability of the Approximation Error Bounds in Theorem 2

We believe it is possible to improve upon the approximation error bounds in Claim 3 of Theorem 2. To explore this possibility, we focus on the exemplary setting with $k = k_* = 3$ and $d = 1$, that is, ones fits three centers to a one-dimensional three-component GMM. By a refined analysis, we show that the coarse approximation error bounds, expressed in (20) and (21), can be sharpened exponentially to the form $e^{-\Omega(\Delta^2/\sigma^2)}$, where $\Delta = \Delta_{\min} = 1/2 \cdot \Delta_{\max}$.

Theorem 3 (Tight error bounds for 3-component GMM). *Let* $d = 1$, $k = k_* = 3$, $\sigma > 0$, *and* $\boldsymbol{\Theta}^* = (\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_3^*) = (-\Delta, 0, \Delta)$ *for some* $\Delta > 0$. *If* $\Delta/\sigma > 2^{16} \cdot 3^{10} \cdot (\sqrt{2\pi} + 1)$, *then there exists a constant* $C > 0$ *such that every local minimizer* $\boldsymbol{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ *of* L *satisfies exactly one of the following possibilities (up to permutation of the indices of* $\{\boldsymbol{\beta}_i\}$ *and that of* $\{\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_3^*\}$*):*

1. $\left\|\boldsymbol{\beta}_1 - \frac{1}{2}\left(\boldsymbol{\theta}_1^* + \boldsymbol{\theta}_2^*\right)\right\| \le \sigma \cdot e^{-C\Delta^2/\sigma^2}$, $\left\|\boldsymbol{\beta}_2 - \boldsymbol{\theta}_3^*\right\| \le \sigma \cdot e^{-C\Delta^2/\sigma^2}$, *and* $\mathbb{E}_*[\Psi_3] \le e^{-C\Delta^2/\sigma^2}$;

2. $\left\|\boldsymbol{\beta}_1 - \frac{1}{2}\left(\boldsymbol{\theta}_1^* + \boldsymbol{\theta}_2^*\right)\right\| \le \sigma \cdot e^{-C\Delta^2/\sigma^2}$ *and* $\left\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_3^*\right\| \le \sigma \cdot e^{-C\Delta^2/\sigma^2}, i \in \{2, 3\}$;

3. $\left\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_i^*\right\| \le \sigma \cdot e^{-C\Delta^2/\sigma^2}, i \in \{1, 2, 3\}$.

Theorem 3, whose proof is presented in Section 8, provides tighter bounds on the approximation error and the association coefficient, $\mathbb{E}_*[\Psi_3]$, both of which decrease exponentially fast as a function of the SNR Δ/σ. As a result, as the SNR increases, each fitted enters is either exponentially close to a true center (or to the mean of two), or its association coefficient (and hence its Voronoi cell) becomes exponentially small. In fact, since L has no other local minima near the true centers $\boldsymbol{\theta}^*$ due to existing local results on GMM [3], the approximation errors in Possibility 3 above are actually zero, in which case $\boldsymbol{B} = \boldsymbol{\Theta}^*$ is the exact global minimizer.

Compared to Theorem 2, Theorem 3 provides more refined information on the possible configurations that a local minimizer can admit. For instance, Theorem 3 rules out the possibility of having one center $\boldsymbol{\beta}_1 \approx \frac{1}{2}(\boldsymbol{\theta}_{1,}^* + \boldsymbol{\theta}_3^*)$ fitting two non-adjacent true centers, thanks to the one-dimension assumption $d = 1$. Theorem 3 also eliminates the possibility that one estimate fits all three true centers, i.e., $\boldsymbol{\beta}_1 \approx \frac{1}{3}(\boldsymbol{\theta}_1^* + \boldsymbol{\theta}_2^* + \boldsymbol{\theta}_3^*)$, while the other two estimates $\boldsymbol{\beta}_2, \boldsymbol{\beta}_3$ are far away from the true centers and have non-association. Such a one-fit-all configuration fails to capture the mixture structure of the data, but we note that formally excluding

this possibility in a general setting is not easy due to the issue of "local minimum at infinity".[2] Theorem 3 addresses this issue when $d = 1$ and shows that any *finite* $\boldsymbol{\beta}$ of this form cannot be a local minimum. It remains open whether one can exclude Scenario 1 in Theorem 3 that involves a non-associated center.

The proof of Theorem 3 builds upon the coarse characterization of local minima state in Theorem 2 and its general form Theorem 4. The coarse characterization restricts the local minimizers of L into a small neighborhood of a few ideal solutions, such as $(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_3^*)$ and $((\boldsymbol{\theta}_1^* + \boldsymbol{\theta}_2^*)/2, \boldsymbol{\theta}_3^*, \boldsymbol{\theta}_3^*)$. By further exploiting the first-order stationary condition (12), we show that the local minimizers must be exponentially close to these ideal solutions within this neighborhood. We believe that this strategy could be extended beyond the one-dimensional three-component setting, and we leave this to future work.

The case of a mis-specified model The techniques used in the proof of Theorem 3 also apply to the case where the number of mixture components is under-specified. Recall that Corollary 4 addresses the scenario where a single Gaussian is used to fit a mixture of three Gaussians, in which case the overall mean $\boldsymbol{\beta}_1 = \frac{1}{3} (\boldsymbol{\theta}_1^* + \boldsymbol{\theta}_2^* + \boldsymbol{\theta}_3^*)$ is the only stationary point. The corollary below deals with fitting a two-component GMM to a three-component GMM.

Corollary 5 (Tight bounds for underfitting 3-component GMM). *Let $d = 1$, $k = 2$, $k_* = 3$, $\sigma > 0$ and $\boldsymbol{\Theta}^* = (\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_3^*) = (-\Delta, 0, \Delta)$ for some $\Delta > 0$. If $\Delta/\sigma > 2^{16} \cdot 3^{10} \cdot (\sqrt{2\pi} + 1)$, then there exists a constant $C > 0$ for which every local minimizer $\boldsymbol{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ of L satisfies (up to permutation of the indices of $\{\boldsymbol{\beta}_i\}$ and that of $\{\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_3^*\}$)*

$$\left\| \boldsymbol{\beta}_1 - \frac{1}{2} (\boldsymbol{\theta}_1^* + \boldsymbol{\theta}_2^*) \right\| \le \sigma \cdot e^{-C\Delta^2/\sigma^2} \quad and$$

$$\|\boldsymbol{\beta}_2 - \boldsymbol{\theta}_3^*\| \le \sigma \cdot e^{-C\Delta^2/\sigma^2}. \tag{26}$$

We prove Corollary 5 in Section 8.3 using intermediate results from the proof of Theorem 3.

We remark that Corollary 5 is related to the recent work by Dwivedi *et al.* [19], who study a similar under-specification setting. They consider fitting a *symmetric* 2-mixture $\frac{1}{2}\mathcal{N}(\boldsymbol{\beta}_1, \sigma^2) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\beta}_1, \sigma^2)$ to a 3-mixture of the form $\frac{1}{4}\mathcal{N}(\boldsymbol{\theta}_1^*(1 + \alpha), \sigma^2) + \frac{1}{4}\mathcal{N}(\boldsymbol{\theta}_1^*(1 - \alpha), \sigma^2) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\theta}_1^*, \sigma^2)$. They provide finite-sample convergence rates for the EM algorithm, assuming that the EM starts from an initial solution sufficiently close to the global minimizer. Note that in the setting of Corollary 5, we effectively establish that there is no other local minimizer besides the global minimizer $\boldsymbol{B} \approx \left(\frac{1}{2} (\boldsymbol{\theta}_1^* + \boldsymbol{\theta}_2^*), \boldsymbol{\theta}_3^* \right)$, and thus the EM converges to this solution from any initialization.

5 Discussion

In this paper, we study the population negative log-likelihood of GMMs with a general number of components, and show that all local minimizers have the common structure that partially reveals the locations of the true components in the global minimum. Our findings have many algorithmic implications and point to a variety of ways of improving iterative methods for learning GMMs—we mention a few here.

(i) Better initialization schemes, such as those that pick k initial centers that are far away from each other, are useful for finding a local minimizer free from the one-fits-many sub-configuration and therefore facilitate the retrieval of a global minimimizer. This ideas underlies the k-means++ and several other clustering algorithms [4, 2, 5, 6].

(ii) Given any local minimizer (e.g., obtained by the EM algorithm), it may be possible to refine the solution iteratively and recover the remaining components by deflating the components already recovered.

(iii) Another natural idea is to resolve the one-fits-many and many-fit-one sub-configurations by adding more center estimates or combining redundant centers. In fact, several improved versions of the EM are based on this "split-and-merge" operation [36, 37, 49].

---

[2]When $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ go to infinity, and hence, move away from the data, the negative log-likelihood L approaches a constant function of $(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$, and the function $\boldsymbol{\beta}_1 \mapsto L((\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3))$ approaches the negative log-likelihood of fitting a single Gaussian to a mixture of three. In this limiting case, $\boldsymbol{\beta}_1 = \frac{1}{3}(\boldsymbol{\theta}_1^* + \boldsymbol{\theta}_2^* + \boldsymbol{\theta}_3^*)$ is indeed a local minimum (see Corollary 4).

A particularly promising algorithmic idea suggested by our results is *over-parameterization*, which helps avoid spurious local minima. In particular, when the number of fitted centers k is sufficiently larger than the number of true components k_*, one expects that the one-fits-many sub-configuration is unlikely to occur. Therefore, an overparmeterized iterative method is likely to converge to a local minimizers that involves only many-fit-one and near-empty subconfigurations, which can then be pruned to identify most (if not all) true components and hence the global minimizer. A similar idea was proposed in the work by Dasgupta and Schulman [13], who propose to over-specify k in the EM algorithm and, after convergence, merge fitted centers that are close to each other (corresponding to many-fit-one) and remove fitted centers with low mixing weights (corresponding to non-association). Recent work on non-parametric maximum likelihood estimation [31, 47], which can be interpreted as an extreme form of overpameterization with $k \to \infty$, demonstrates great potential of this approach.

There are several avenues for future research to improve upon the analysis presented in this paper. First, it is of immediate interest to obtain sharper results regarding the SNR requirement, approximation error bounds, and the strength of association across different sub-configurations (see Remarks on Theorem 2 in Section 4.1). Achieving these goals may require developing analytical techniques that do not rely on Voronoi-cells, as their use in our analysis essentially impedes the propagation of influence from a component beyond its own cell, which possibly have led to suboptimal guarantees. Another direction for future research is to extend our population-level results to the finite-sample setting, for which the uniform concentration and localization techniques developed in [28, 18] could be useful. In addition, it would be interesting to explore the low-SNR regime, where the mixture components have small or even no separation, as the structures of the local minima may become more complicated in such cases. Finally, it would be valuable to understand whether the phenomenon of structured local minima holds more generally in other mixture and latent variable models; for example, empirical evidences in a recent study [8] suggest that this phenomenon may be universal.

6 Deferred Proofs from Section 3

6.1 Proofs of Lemma 1 and Lemma 2

Proof of Lemma 1. Recall that $f_i(\boldsymbol{x}) := \phi(\boldsymbol{x} \mid \boldsymbol{\beta}_j, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\|\boldsymbol{x} - \boldsymbol{\beta}_i\|^2 / 2\sigma^2\right)$, and hence,

$$\frac{\partial f_i(\boldsymbol{x})}{\partial \boldsymbol{\beta}_j} = \begin{cases} \frac{1}{\sigma^2} \cdot f_i(\boldsymbol{x}) \cdot (\boldsymbol{x} - \boldsymbol{\beta}_i) & \text{if } j = i, \\ 0 & \text{if } j \neq i. \end{cases} \tag{27}$$

Since $f := \frac{1}{k} \sum_{i \in [k]} f_i$, it follows that

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}_j} L(\boldsymbol{B}) &= -\frac{\partial}{\partial \boldsymbol{\beta}_j} \mathbb{E}_* \left[\log f(\mathsf{x}) \right] \\ &\overset{(a)}{=} -\mathbb{E}_* \left[\frac{\frac{1}{k} \sum_{i \in [k]} \frac{\partial}{\partial \boldsymbol{\beta}_j} f_i(\mathsf{x})}{f(\mathsf{x})} \right] \\ &\overset{(b)}{=} -\mathbb{E}_* \left[\frac{\frac{1}{k} \cdot \frac{1}{\sigma^2} \cdot f_j(\mathsf{x}) \cdot (\mathsf{x} - \boldsymbol{\beta}_\mathsf{j})}{f(\mathsf{x})} \right] \\ &\overset{(c)}{=} \frac{1}{\sigma^2} \cdot \mathbb{E}_* \left[\Psi_j \cdot (\boldsymbol{\beta}_j - \mathsf{x}) \right], \end{aligned}$$

where (a) follows from exchanging the expectation and differentiation using the dominated convergence theorem[3], (b) follows from (27), and (c) follows from the definition of Ψ_j in (9) (see Definition 1 as well). □

---

[3] We omit explicitly mentioning this argument in the rest of this paper.

Proof of Lemma 2. First of all, we compute the partial derivative of $\psi_i(\boldsymbol{x})$. For each $j \in [k]$, we observe that

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}_j} \psi_i(\boldsymbol{x}) &= \frac{\partial}{\partial \boldsymbol{\beta}_j} \frac{\frac{1}{k} f_i(\boldsymbol{x})}{f(\boldsymbol{x})} \\
&= \frac{1}{k} \frac{\frac{\partial}{\partial \boldsymbol{\beta}_j} f_i(\boldsymbol{x}) \cdot f(\boldsymbol{x}) - f_i(\boldsymbol{x}) \cdot \frac{\partial}{\partial \boldsymbol{\beta}_j} f(\boldsymbol{x})}{f(\boldsymbol{x})^2} \\
&\overset{(a)}{=} \frac{1}{k} \left[\frac{\frac{1}{\sigma^2} \cdot f_i(\boldsymbol{x}) \cdot (\boldsymbol{x} - \boldsymbol{\beta}_i) \cdot \delta_{ij}}{f(\boldsymbol{x})} \right. \\
&\qquad\qquad \left. - \frac{f_i(\boldsymbol{x}) \cdot \frac{1}{k} \cdot \frac{1}{\sigma^2} \cdot f_j(\boldsymbol{x}) \cdot (\boldsymbol{x} - \boldsymbol{\beta}_j)}{f^2(\boldsymbol{x})} \right] \\
&= \frac{1}{\sigma^2} \big[\psi_i(\boldsymbol{x}) \cdot \delta_{ij} - \psi_i(\boldsymbol{x}) \psi_j(\boldsymbol{x}) \big] \cdot (\boldsymbol{x} - \boldsymbol{\beta}_j).
\end{aligned}
\tag{28}
$$

where (a) follows from (27). Therefore, we obtain that

$$
\begin{aligned}
\sigma^2 \cdot \frac{\partial^2}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_j} L(\boldsymbol{B}) &\overset{(a)}{=} \sigma^2 \cdot \frac{\partial}{\partial \boldsymbol{\beta}_j} \left(\frac{1}{\sigma^2} \cdot \mathbb{E}_* \big[\Psi_i \cdot (\boldsymbol{\beta}_i - \mathsf{x}) \big] \right) \\
&\overset{(b)}{=} \mathbb{E}_* \left[\left(\frac{\partial}{\partial \boldsymbol{\beta}_j} \Psi_i \right) \cdot (\boldsymbol{\beta}_i - \mathsf{x})^\top + \Psi_i \cdot \left(\frac{\partial}{\partial \boldsymbol{\beta}_j} \boldsymbol{\beta}_i^\top \right) \right] \\
&\overset{(b)}{=} \mathbb{E}_* \left[\left(\frac{1}{\sigma^2} \big[\Psi_i \cdot \delta_{ij} - \Psi_i \Psi_j \big] \cdot (\mathsf{x} - \boldsymbol{\beta}_j) \right) \cdot (\boldsymbol{\beta}_i - \mathsf{x})^\top \right] + \mathbb{E}_* \left[\Psi_i \cdot \left(\frac{\partial}{\partial \boldsymbol{\beta}_j} \boldsymbol{\beta}_i^\top \right) \right] \\
&= \frac{1}{\sigma^2} \cdot \mathbb{E}_* \left[\Psi_i \Psi_j \cdot (\boldsymbol{\beta}_j - \mathsf{x})(\boldsymbol{\beta}_i - \mathsf{x})^\top \right] + \delta_{ij} \cdot \mathbb{E}_* \left[\Psi_i \cdot \left\{ \boldsymbol{I}_d - \frac{1}{\sigma^2} \cdot (\boldsymbol{\beta}_i - \mathsf{x})(\boldsymbol{\beta}_i - \mathsf{x})^\top \right\} \right],
\end{aligned}
$$

where (a) follows from Lemma 1, (b) is by the chain rule, and (c) is due to (28). The proof is complete after rearranging terms. $\qquad \square$

6.2   Proof of Theorem 1

We apply the Stein's identity [35] to prove Theorem 1. Specifically, we make use of the following multivariate version of the Stein's identity specialized to the identity covariance setting.

Lemma 3 (Stein's identity)**.** *Suppose* $\mathsf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d)$ *and* $g : \mathbb{R}^d \to \mathbb{R}$ *is a differentiable function. Then*

$$
\mathbb{E}\left[g(\mathsf{x}) \cdot (\mathsf{x} - \boldsymbol{\mu}) \right] = \sigma^2 \cdot \mathbb{E}\left[\nabla g(\mathsf{x}) \right].
$$

Next, we present a proof of Theorem 1.

Proof of Theorem 1. Fix an arbitrary $i \in [k]$ and compute the derivative $\nabla \psi_i(\boldsymbol{x})$ with respect to $\boldsymbol{x}$. Recall from (8) that

$$
\begin{aligned}
\psi_i(\boldsymbol{x}) &:= \frac{e^{-\|\boldsymbol{x} - \boldsymbol{\beta}_i\|^2 / (2\sigma^2)}}{\sum_{j \in [k]} e^{-\|\boldsymbol{x} - \boldsymbol{\beta}_j\|^2 / (2\sigma)^2}} \\
&= \frac{1}{\sum_{j \in [k]} e^{\left(\|\boldsymbol{x} - \boldsymbol{\beta}_i\|^2 - \|\boldsymbol{x} - \boldsymbol{\beta}_j\|^2 \right) / (2\sigma^2)}}.
\end{aligned}
$$

Taking the derivative, we have

$$\nabla \psi_i(\boldsymbol{x}) = - \sum_{j \in [k]} \frac{e^{\left(\|\boldsymbol{x}-\boldsymbol{\beta}_i\|^2 - \|\boldsymbol{x}-\boldsymbol{\beta}_j\|^2\right)/(2\sigma)^2} \cdot (\boldsymbol{\beta}_j - \boldsymbol{\beta}_i) \cdot \sigma^{-2}}{\left(\sum_{\ell \in [k]} e^{\left(\|\boldsymbol{x}-\boldsymbol{\beta}_i\|^2 - \|\boldsymbol{x}-\boldsymbol{\beta}_\ell\|^2\right)/(2\sigma)^2}\right)^2}$$

$$= \sum_{j \in [k]} \frac{e^{-\|\boldsymbol{x}-\boldsymbol{\beta}_i\|^2/(2\sigma)^2} e^{-\|\boldsymbol{x}-\boldsymbol{\beta}_j\|^2/(2\sigma)^2}}{\left(\sum_{\ell \in [k]} e^{-\|\boldsymbol{x}-\boldsymbol{\beta}_\ell\|^2/(2\sigma)^2}\right)^2} \cdot (\boldsymbol{\beta}_i - \boldsymbol{\beta}_j) \cdot \sigma^{-2}$$

$$= \sigma^{-2} \sum_{j \in [k]} \psi_i(\boldsymbol{x}) \psi_j(\boldsymbol{x}) \cdot (\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)$$

$$= \sigma^{-2} \cdot \left(\psi_i(\boldsymbol{x}) \cdot \boldsymbol{\beta}_i - \sum_{j \in [k]} \psi_i(\boldsymbol{x}) \psi_j(\boldsymbol{x}) \cdot \boldsymbol{\beta}_j \right),$$

where the last step follows from the fact that $\sum_{j \in [k]} \psi_j(\boldsymbol{x}) = 1$. Noting that $\mathsf{x} \sim \mathcal{N}(\boldsymbol{\theta}_s^*, \sigma^2 \boldsymbol{I}_d)$ under the density f_s^*, we apply the Stein's identity (Lemma 3) to obtain that for all $(s, i) \in [k_*] \times [k]$,

$$\mathbb{E}_s \left[\Psi_i \cdot \mathsf{x} \right] = \boldsymbol{\theta}_s^* \cdot \mathbb{E}_s \left[\Psi_i \right] + \sigma^2 \cdot \mathbb{E}_s \left[\nabla \psi_i(\mathsf{x}) \right]$$

$$= \boldsymbol{\theta}_s^* \cdot \mathbb{E}_s \left[\Psi_i \right] + \boldsymbol{\beta}_i \cdot \mathbb{E}_s \left[\Psi_i \right] - \sum_{j \in [k]} \boldsymbol{\beta}_j \cdot \mathbb{E}_s \left[\Psi_i \Psi_j \right]. \tag{29}$$

Recall from the stationarity condition (12) that $\boldsymbol{\beta}$ is a stationary point of L if and only if

$$\boldsymbol{\beta}_i = \frac{\sum_{s \in [k_*]} \mathbb{E}_s \left[\Psi_i \mathsf{x} \right]}{\sum_{s \in [k_*]} \mathbb{E}_s \left[\Psi_i \right]}, \qquad \forall i \in [k].$$

Combining this condition with the above expression for $\mathbb{E}_s \left[\Psi_i \cdot \mathsf{x} \right]$ in (29), we obtain that for all $i \in [k]$,

$$\sum_{s \in [k_*]} \mathbb{E}_s \left[\Psi_i \right] \cdot \boldsymbol{\beta}_i = \sum_{s \in [k_*]} \left(\boldsymbol{\theta}_s^* \cdot \mathbb{E}_s \left[\Psi_i \right] + \boldsymbol{\beta}_i \cdot \mathbb{E}_s \left[\Psi_i \right] - \sum_{j \in [k]} \boldsymbol{\beta}_j \cdot \mathbb{E}_s \left[\Psi_i \Psi_j \right] \right).$$

Rearranging terms yields

$$\sum_{s \in [k_*]} \sum_{j \in [k]} \boldsymbol{\beta}_j \cdot \mathbb{E}_s \left[\Psi_i \Psi_j \right] = \sum_{s \in [k_*]} \boldsymbol{\theta}_s^* \cdot \mathbb{E}_s \left[\Psi_i \right], \quad \forall i \in [k],$$

which is equivalent to the condition in (16). Thus, the proof is complete. $\qquad \square$

6.3   Proof of Corollary 2

Proof of Corollary 2. Given $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) \in \mathbb{R}^{d \times k}$, let $k' \in \mathbb{N}_+$ denote the number of distinct elements (=columns) in $\boldsymbol{B}$. Note that $k' \leq k$, and that we may assume $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{k'}$ are all distinct without loss of generality (by permuting the order). For each $i \in [k']$, let $\mathcal{I}_i := \{i' \in [k] : \boldsymbol{\beta}_{i'} = \boldsymbol{\beta}_i\}$, $m_i := |\mathcal{I}_i|$, and $\Psi_i' := \sum_{i' \in \mathcal{I}_i} \Psi_{i'} = m_i \cdot \Psi_i$.

Then we observe that for all $i \in [k']$,

$$\sum_{s \in [k_*]} \boldsymbol{\theta}_s^* \cdot \mathbb{E}_s \left[\Psi_i' \right] = \sum_{i' \in \mathcal{I}_i} \sum_{s \in [k_*]} \boldsymbol{\theta}_s^* \cdot \mathbb{E}_s \left[\Psi_{i'} \right]$$

$$= \sum_{i' \in \mathcal{I}_i} \sum_{s \in [k_*]} \sum_{j \in [k]} \boldsymbol{\beta}_j \cdot \mathbb{E}_s \left[\Psi_{i'} \Psi_j \right] \qquad \because (16)$$

$$= \sum_{i' \in \mathcal{I}_i} \sum_{s \in [k_*]} \sum_{j \in [k']} \sum_{j' \in \mathcal{I}_j} \boldsymbol{\beta}_j \cdot \mathbb{E}_s \left[\Psi_{i'} \Psi_{j'} \right]$$

$$= \sum_{s \in [k_*]} \sum_{j \in [k']} \boldsymbol{\beta}_j \cdot \mathbb{E}_s \left[\Psi_i' \Psi_j' \right]. \tag{30}$$

20

Letting $\vec{\Psi'} := (\Psi'_1, \ldots, \Psi'_{k'})^\top \in \mathbb{R}^{k' \times 1}$, we can rewrite the condition (30) as the following matrix equation:

$$\boldsymbol{\Theta}^* \cdot \boldsymbol{\Phi'} = \boldsymbol{B'} \cdot \mathbb{E}_* \left[\vec{\Psi'} \cdot \vec{\Psi'}^\top \right] \tag{31}$$

where

$$\boldsymbol{\Phi'} = \left[\mathbb{E}_1[\vec{\Psi'}] \quad \ldots \quad \mathbb{E}_{k_*}[\vec{\Psi'}] \right]^\top \in \mathbb{R}^{k_* \times k'}, \quad \text{and}$$
$$\boldsymbol{B'} = \boldsymbol{B}([k']) \in \mathbb{R}^{d \times k'}.$$

Next, we claim that the k'-by-k' matrix $\mathbb{E}_* \left[\vec{\Psi'} \cdot \vec{\Psi'}^\top \right]$ is invertible. If we assume otherwise, there exists a nonzero vector $\boldsymbol{u} \in \mathbb{R}^{k'}$ such that

$$\mathbb{E}_* \left[(\boldsymbol{u}^T \vec{\Psi'})^2 \right] = \boldsymbol{u}^T \mathbb{E}_* \left[\vec{\Psi'} \cdot \vec{\Psi'}^\top \right] \boldsymbol{u} = 0,$$

which implies that

$$\boldsymbol{u}^T \vec{\Psi'} = \sum_{i \in [k']} u_i \cdot m_i \cdot \Psi_i = 0$$

almost surely with respect to f^*. This is equivalent to the condition that

$$\sum_{i \in [k']} u_i \cdot m_i \cdot \psi_i(\boldsymbol{x}) = 0$$

almost everywhere (with respect to the Lebesgue measure) because $f^*(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$. Because $m_i \geq 1 > 0$ for all $i \in [k']$ and $\{\boldsymbol{\beta}_i : i \in [k']\}$ are distinct, the k' functions $m_i \cdot \psi_i : \boldsymbol{x} \mapsto m_i \cdot e^{-\|\boldsymbol{x} - \boldsymbol{\beta}_i\|^2/(2\sigma^2)}$, $i \in [k']$ are linearly independent. This renders a contradiction.

Finally, we obtain from (31) using the invertibility of $\mathbb{E}_* \left[\vec{\Psi'} \cdot \vec{\Psi'}^\top \right]$ that

$$\boldsymbol{B'} = \boldsymbol{\Theta}^* \cdot \boldsymbol{\Phi'} \cdot \mathbb{E}_* \left[\vec{\Psi'} \cdot \vec{\Psi'}^\top \right]^{-1}.$$

Therefore, $\boldsymbol{\beta}_i$ belongs to the column space of $\boldsymbol{\Theta}^*$ for all $i \in [k']$, and the proof of Corollary 2 is complete. $\qquad\square$

7 Proof of Theorem 2 (the Main Theorem)

This section is dedicated to proving the main theorem of this paper, namely, Theorem 2. To this end, we start in Section 7.1 by presenting Theorem 4, which serves as the master theorem of this work and from which Theorem 2 is derived as a special instance. In Section 7.2, we prove Theorem 4 by relying on five propositions that we introduce in Section 7.2.1. We then proceed in Section 7.3 to derive Theorem 2 from Theorem 4 by choosing specific parameters (Section 7.3.1) and making slight adjustments (Section 7.3.2) when necessary.

Throughout this section, we assume that $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) \in \mathbb{R}^{d \times k}$ is an arbitrary, yet fixed, local minimizer of L, unless stated otherwise. To avoid clutter, we may omit the explicit reference to $\boldsymbol{B}$ in our notation, such as when referring to the coefficient of association $\psi_i(\boldsymbol{x})$ or the Voronoi cell $\mathcal{V}_i$, when it is clear from the context.

7.1 A Complete Version of Theorem 2

To state the master theorem, we define a relaxed notion of a partition of a finite set. To better understand of this relaxation, it is helpful to first recall the formal definition of a partition.

Definition 3. *A family of sets* $\mathbb{T}$ *is a* partition *of* $\mathcal{X}$ *if and only if all of the following conditions hold:*

- $\mathbb{T}$ *does not contain the empty set, i.e.,* $\emptyset \notin \mathbb{T}$.

- *The sets in $\mathbb{T}$ cover $\mathcal{X}$, i.e., $\bigcup_{A \in \mathbb{T}} A = \mathcal{X}$.*

- *The elements of $\mathbb{T}$ are pairwise disjoint, i.e., $\forall A, B \in \mathbb{T}$, if $A \neq B$ then $A \cap B = \emptyset$.*

We now introduce the notion of a quasi-partition, which relaxes the third requirement above.

Definition 4 (Quasi-partition of a set)**.** *Let $\mathbb{T}$ be a family of sets and $\mathcal{I}, \mathcal{J} \subset \mathbb{T}$ such that $\mathcal{I} \cap \mathcal{J} = \emptyset$. $\mathbb{T}$ is an $(\mathcal{I}, \mathcal{J})$-quasi-partition of $\mathcal{X}$ if and only if all of the following conditions hold:*

- *$\mathbb{T}$ does not contain the empty set, i.e., $\emptyset \notin \mathbb{T}$.*

- *The sets in $\mathbb{T}$ cover $\mathcal{X}$, i.e., $\bigcup_{A \in \mathbb{T}} A = \mathcal{X}$.*

- *The elements of $\mathbb{T}$ are pairwise disjoint except for pairs across $\mathcal{I}$ and $\mathcal{J}$, i.e., $\forall A, B \in \mathbb{T}$ such that $A \neq B$, if $A \cap B \neq \emptyset$, then either (1) $A \in \mathcal{I}$ and $B \in \mathcal{J}$ or (2) $A \in \mathcal{J}$ and $B \in \mathcal{I}$.*

Remark 2. Suppose that $\mathbb{T}$ is an $(\mathcal{I}, \mathcal{J})$-quasi-partition of $\mathcal{X}$ for some $\mathcal{I}, \mathcal{J} \subset \mathbb{T}$. Then for each $x \in \mathcal{X}$, there exist either one or two sets in $\mathbb{T}$ that contain x. Moreover, if there are two sets that contain x, then one should belong to $\mathcal{I}$ with the other being a member of $\mathcal{J}$.

7.1.1 Statement of the Master Theorem

Theorem 4 (Master theorem)**.** *Let $k, k_* \in \mathbb{N}$, $\boldsymbol{\Theta}^* \in \mathbb{R}^{d \times k_*}$, and $\boldsymbol{B} \in \mathbb{R}^{d \times k}$ be a local minimum of $L(\cdot \mid \boldsymbol{\Theta}^*)$. If $\delta \in \mathbb{R}_+$ satisfies $\frac{4k_* \cdot \sigma}{\Delta_{\min}} < \delta \leq \frac{1}{18(\sqrt{2\pi}+1) \cdot k^4}$, then there exist $q \in \mathbb{N}$, and two collections of sets $\mathbb{T}^\delta := \left\{ \mathcal{T}_a^\delta \in [k] : a \in [q]_0 \right\}$ and $\mathbb{S}^\delta := \left\{ \mathcal{S}_a^\delta \in [k_*] : a \in [q] \right\}$, for which the following properties hold.*

1. *(Simple quasi-partitions) There exist q_0 with $0 \leq q_0 \leq q$ such that the following properties hold:*

    (a) *$\mathbb{T}^\delta \setminus \{\mathcal{T}_0^\delta\}$ is a $\left(\{\mathcal{T}_a^\delta\}_{a=1}^{q_0}, \{\mathcal{T}_a^\delta\}_{a=q_0+1}^q \right)$-quasi-partition of $[k] \setminus \mathcal{T}_0^\delta$.*

    (b) *$\mathbb{S}^\delta$ is a partition of $[k_*]$.*

    (c) *$|\mathcal{T}_a^\delta| = 1$ for all $a \in [q_0]$.*

    (d) *$|\mathcal{T}_a| \geq 2$ and $|\mathcal{S}_a| = 1$ for all $a \in [q] \setminus [q_0]$.*

2. *(Mutual exclusiveness) Suppose that $i \in \mathcal{T}_a^\delta$ and $j \in \mathcal{T}_b^\delta$ for $a, b \in [q]_0$ with $a \neq b$. If $i \neq j$, then $\boldsymbol{\beta}_i \neq \boldsymbol{\beta}_j$.*

3. *(Approximation error)*

    (a) *(one-fits-many) Let*
    $$D_{\mathrm{cell}}(\boldsymbol{B}) := \max_{j \in [k]} \max_{s \in [k_*] : \boldsymbol{\theta}_s^* \in \mathcal{V}_j} \|\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_j\|.$$

    For $a \in [q_0]$,

    $$
    \begin{aligned}
    \frac{1}{\sigma} \left\| \boldsymbol{\beta}_{i_a} - \frac{1}{|\mathcal{S}_a^\delta|} \sum_{s \in \mathcal{S}_a^\delta} \boldsymbol{\theta}_s^* \right\| &\leq 18 \left(\sqrt{2\pi} + 1 \right) \cdot k^3 \cdot (5k_* + 2k) \cdot \frac{D_{\mathrm{cell}}(\boldsymbol{B})}{\sigma} \cdot \delta \\
    &\quad + \frac{1}{|\mathcal{S}_a^\delta|} \cdot \left\{ \frac{8k_*}{3} \cdot \frac{1}{\delta} + 2k_* \cdot (\sqrt{d} + 4) + \frac{4}{3\sqrt{2\pi}} \right\} \\
    &\quad + 3.
    \end{aligned}
    \tag{32}
    $$

    where i_a denotes the unique element in $\mathcal{T}_a^\delta$.

    (b) *(many-fit-one) For $a \in [q] \setminus [q_0]$,*

    $$\frac{1}{\sigma} \left\| \boldsymbol{\beta}_i - \boldsymbol{\theta}_{s_a}^* \right\| \leq \frac{2k_*}{\delta}, \qquad \forall i \in \mathcal{T}_a^\delta, \tag{33}$$

    where s_a denotes the unique element in $\mathcal{S}_a^\delta$.

4. *(Near-empty cross-association)* Let $(a, b) \in [q] \times [q]_0$ such that $a \neq b$. If $s \in \mathcal{S}_a^\delta$ and $i \in \mathcal{T}_b^\delta$, then

$$
\begin{aligned}
\mathbb{E}_s\left[\Psi_i\right] &\leq 9\left(\sqrt{2\pi} + 1\right) \cdot \tilde{\kappa}(a) \cdot \delta, \quad and \\
\mathbb{P}_s\left(\mathcal{V}_i\right) &\leq 3\left(\sqrt{2\pi} + 1\right) \cdot \tilde{\kappa}(a) \cdot \delta
\end{aligned}
\tag{34}
$$

where

$$
\tilde{\kappa}(a) = \begin{cases} k^4 & if\ a \in [q_0], \\ k^3 & if\ a \in [q] \setminus [q_0]. \end{cases}
$$

7.1.2 Construction of the Collections of Sets $\mathbb{T}^\delta$ and $\mathbb{S}^\delta$ in Theorem 4

Here, we describe an algorithm that produces the collections of sets, $\mathbb{T}^\delta$ and $\mathbb{S}^\delta$, in the statement of Theorem 4. For the convenience of presentation, we define two types of sets. For any $i \in [k]$ and for any $\delta \geq 0$, we let

$$
\mathcal{A}_i^\delta := \left\{ s \in [k_*] \ \middle| \ \iota_{\boldsymbol{B}}(\boldsymbol{\theta}_s^*) = i \quad and \quad \max_{j, j' \in [k]} \frac{\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'}\|}{\sigma} \cdot \mathbb{E}_s\left[\Psi_j \Psi_{j'}\right] < \delta \right\}.
\tag{35}
$$

For any $s \in [k_*]$ and for any $\delta > 0$, let

$$
\mathcal{E}_s^\delta := \left\{ i \in [k] \ \middle| \ \max_{j \in [k] \setminus \{i\}} \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \cdot \mathbb{E}_s\left[\Psi_i \Psi_j\right] \geq \delta \right\}.
\tag{36}
$$

Remark 3. Intuitively, $\mathcal{A}_i^\delta$ is the set of true cluster indices that are exclusively associated to $\boldsymbol{\beta}_i$ at level δ. Similarly, we can interpret $\mathcal{E}_s^\delta$ as the set of indices of estimated centers that contend for the possesion of $\boldsymbol{\theta}_s^*$.

Then we describe a procedure to construct covers of $[k_*]$ and $[k]$, whose pseudocode can be found in Algorithm 1. For any inputs $\delta > 0$ and $\boldsymbol{B} \in \mathbb{R}^{d \times k}$, Algorithm 1 outputs $q \in \mathbb{N}$ and $q \in \mathbb{N}$ along with two collections of sets, namely, $\mathbb{T}^\delta = \left\{\mathcal{T}_a^\delta\right\}_{a=0}^q$ and $\mathbb{S}^\delta = \left\{\mathcal{S}_a^\delta\right\}_{a=1}^q$.

Algorithm 1 Construction of partitions of $[k_*]$ and $[k]$

Input: $\delta \in \mathbb{R}_+$, $\boldsymbol{B} \in \mathbb{R}^{d \times k}$
Output: q_0, q, $\mathbb{T}^\delta = \{\mathcal{T}_a^\delta\}_{a=0}^q$, $\mathbb{S}^\delta = \{\mathcal{S}_a^\delta\}_{a=1}^q$

1: Initialize $a \leftarrow 0$
2: **for** $i = 1, 2, \ldots, k$ **do**
3:    **if** $\mathcal{A}_i^\delta \neq \emptyset$ **then**
4:       $a \leftarrow a + 1$
5:       $\mathcal{T}_a^\delta \leftarrow \{i\}$
6:       $\mathcal{S}_a^\delta \leftarrow \mathcal{A}_i^\delta$ {See (35) for definition of $\mathcal{A}_i^\delta$}
7:    **end if**
8: **end for**
9: $q_0 \leftarrow a$
10: $\mathcal{R}^\delta \leftarrow [k_*] \setminus \left(\bigcup_{a'=1}^{q_0} \mathcal{S}_{a'}^\delta\right)$ {'Remaining'}
11: **for** $s = 1, 2, \ldots, k_*$ **do**
12:    **if** $s \in \mathcal{R}^\delta$ **then**
13:       $a \leftarrow a + 1$
14:       $\mathcal{T}_a^\delta \leftarrow \mathcal{E}_s^\delta$ {See (36) for definition of $\mathcal{E}_s^\delta$}
15:       $\mathcal{S}_a^\delta \leftarrow \{s\}$
16:    **end if**
17: **end for**
18: $q \leftarrow a$
19: $\mathcal{T}_0^\delta \leftarrow [k] \setminus \bigcup_{a=1}^q \mathcal{T}_a^\delta$

7.2 Proof of Theorem 4

7.2.1 Five Preparatory Propositions

Here we present five key propositions that establish useful properties of association coefficients (Propositions 1 and 2), geometry of local minima (Propositions 3 and 4), and combinatorial properties of the covers $\mathbb{T}^\delta, \mathbb{S}^\delta$ constructed by Algorithm 1 (Proposition 5). We will use these propositions to prove Theorem 4 in Section 7.2.2. Proofs of Propositions 1 through 5 are deferred to the Appendix (Appendices A through E).

Strength of associations First of all, we present a proposition that controls the strength of association between the estimate $\boldsymbol{\beta}_i$ and the true mean $\boldsymbol{\theta}_s^*$, measured in two quantities $\mathbb{P}_s(\mathcal{V}_i)$ and $\mathbb{E}_s[\Psi_i]$.

Proposition 1 (Association bounds). *Let $\boldsymbol{B} \in \mathbb{R}^{d \times k}$ be an arbitrary ordered set of vectors. For any $(s, i) \in [k_*] \times [k]$ and any sequence $(\alpha_j \in \mathbb{R}_+ : j \in [k] \setminus \{i\})$, if $\boldsymbol{\theta}_s^* \notin \operatorname{int} \mathcal{V}_i$, then*

1. $\mathbb{P}_s(\mathcal{V}_i) \le M_{si}(\boldsymbol{\alpha}, \boldsymbol{B})$, and

2. $\mathbb{E}_s[\Psi_i] \le 3M_{si}(\boldsymbol{\alpha}, \boldsymbol{B})$.

where

$$M_{si}(\boldsymbol{\alpha}, \boldsymbol{B}) := \left(\sqrt{2\pi} + 1\right) \cdot k^2 \sum_{j \in [k] \setminus \{i\}} \max\left\{1, \frac{1}{\alpha_j}\right\} \cdot \mathbb{E}_s[\Psi_i \Psi_j] \cdot \exp\left(3\alpha_j \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right).$$

Proposition 1 states that if $\boldsymbol{\theta}_s^* \notin \operatorname{int} \mathcal{V}_i$ and $\mathbb{E}_s[\Psi_i \Psi_j]$ is small for *all* $j \in [k] \setminus \{i\}$, then the Voronoi cell associated with $\boldsymbol{\beta}_i$ must have a small size (under ν_s^*) as well as a small coefficient of association to ν_s^*. We present a proof of Proposition 1 in Appendix A.

Corollary 6. *Let $\boldsymbol{B} \in \mathbb{R}^{d \times k}$ be an arbitrary ordered set of vectors and let $(s, i) \in [k_*] \times [k]$. If $\boldsymbol{\theta}_s^* \notin \operatorname{int} \mathcal{V}_i$, then*

1. $\mathbb{P}_s(\mathcal{V}_i) \le 3\left(\sqrt{2\pi} + 1\right) \cdot k^2 \sum_{j \in [k] \setminus \{i\}} \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \cdot \mathbb{E}_s[\Psi_i \Psi_j]$; and

2. $\mathbb{E}_s[\Psi_i] \le 9\left(\sqrt{2\pi} + 1\right) \cdot k^2 \sum_{j \in [k] \setminus \{i\}} \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \cdot \mathbb{E}_s[\Psi_i \Psi_j]$.

Proof of Corollary 6. Let $z \in \mathbb{R}_+$ and Let $f_z : \mathbb{R}_+ \to \mathbb{R}_+$ be a function such that $f_z(\alpha) = \max\left\{1, \frac{1}{\alpha}\right\} \cdot \exp(\alpha \cdot z)$. Observe that

$$
\begin{aligned}
\inf_{\alpha \in \mathbb{R}_+} f_z(\alpha) &= \inf\left\{\inf_{\alpha \in (0,1)}\left\{\frac{1}{\alpha} \cdot e^{\alpha \cdot z}\right\}, \inf_{\alpha \in [1, \infty)]}\left\{e^{\alpha \cdot z}\right\}\right\} \\
&= \begin{cases} e \cdot z & \text{if } z > 1 \text{ (minimum attained at } \alpha = \frac{1}{z}), \\ e^z & \text{if } z \le 1, \end{cases} \\
&\le e \cdot z.
\end{aligned}
$$

Then the conclusion of this corollary immediately follows from Proposition 1. $\qquad\square$

Exclusion principle for associations Recall from (35) that for any $i \in [k]$ and any $\delta \ge 0$,

$$\mathcal{A}_i^\delta := \left\{s \in [k_*] \;\middle|\; \iota_{\boldsymbol{B}}(\boldsymbol{\theta}_s^*) = i \quad \text{and} \quad \max_{j, j' \in [k]} \frac{\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'}\|}{\sigma} \cdot \mathbb{E}_s[\Psi_j \Psi_{j'}] < \delta\right\}.$$

Proposition 2. *Let $\boldsymbol{B} \in \mathbb{R}^{d \times k}$ be an arbitrary ordered set of vectors, let $i \in [k]$, and let $\delta \in \mathbb{R}_+$. If $\mathcal{A}_i^\delta \ne \emptyset$ and $\delta \le \frac{1}{18(\sqrt{2\pi}+1) \cdot k^4}$, the following three statements hold:*

1. $\boldsymbol{\beta}_i \ne \boldsymbol{\beta}_j$ for all $j \in [k] \setminus \{i\}$;

2. $1 - \mathbb{E}_s[\Psi_i] \le 9\left(\sqrt{2\pi} + 1\right) \cdot k^4 \cdot \delta$ for all $s \in \mathcal{A}_i^\delta$;

3. $\mathbb{P}_s(\mathcal{V}_i^c) \le 3\left(\sqrt{2\pi} + 1\right) \cdot k^4 \cdot \delta$ for all $s \in \mathcal{A}_i^\delta$.

Proposition 2 states that when the parameter δ is sufficiently small, the mixture components with $s \in \mathcal{A}_i^\delta$ are almost exclusively associated to $\boldsymbol{\beta}_i$. This claim is quantified in two different measures, namely, $\mathbb{E}_s[\Psi_i]$ and $\mathbb{P}_s[\mathcal{V}_i]$. We present a proof of Proposition 2 in Appendix B.

Proximity of $\boldsymbol{B}$ controlled by product of association coefficients   Next, we discuss several useful properties of $\boldsymbol{B}$, which is a local minimum of L. First, we present a proposition that establishes several upper bounds on the proximity of the estimates, $\boldsymbol{\beta}_i$, to the true centers, $\boldsymbol{\theta}_s^*$.

Proposition 3 (Proximity bound, I). *Let $\boldsymbol{B}$ be a local minimum of L. For all $i,j \in [k]$ such that $i \neq j$, the following three statements hold:*

1. $\dfrac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|^2}{\sigma^2} \leq \dfrac{\mathbb{E}_*[\Psi_i + \Psi_j]}{\mathbb{E}_*[\Psi_i \cdot \Psi_j]}$,

2. $\min_{s \in [k_*]} \dfrac{\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\|^2}{\sigma^2} \leq \dfrac{2(\mathbb{E}_*[\Psi_i] + 1)}{\mathbb{E}_*[\Psi_i \cdot \Psi_j]}$, *and*

3. $\dfrac{\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\|^2}{\sigma^2} \leq 2k_* \cdot \dfrac{\mathbb{E}_*[\Psi_i] + 1}{\mathbb{E}_s[\Psi_i \cdot \Psi_j]}$ *for all $s \in [k_*]$.*

Proposition 3 states that if $\mathbb{E}_*[\Psi_i \Psi_j]$ is large for some $i,j \in [k]$ with $i \neq j$, then $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ must be both close to each other (Claim 1). Moreover, there exists a true center $\boldsymbol{\theta}_s^*$ that is close to both $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ (Claim 2). Claim 3 of Proposition 3 provides a weaker, yet still useful, upper bound for our analysis. We postpone our proof of Proposition 3 to Appendix C.

Corollary 7. *Let $\boldsymbol{B}$ be a local minimum of L, let $\delta \in \mathbb{R}_+$, and $s \in [k_*]$. If $\mathcal{E}_s^\delta \neq \emptyset$, then for all $i \in \mathcal{E}_s^\delta$,*

$$\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\|}{\sigma} \leq \frac{2k_*}{\delta}.$$

Proof of Corollary 7. Observe that if $\mathcal{E}_s^\delta \neq \emptyset$, then $|\mathcal{E}_s^\delta| \geq 2$ by definition. Choose any $i \in \mathcal{E}_s^\delta$ and any $j \in \mathcal{E}_s^\delta \setminus \{i\}$. Then

$$
\begin{aligned}
\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} &\overset{(a)}{\leq} \frac{\mathbb{E}_*[\Psi_i + \Psi_j]}{\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \cdot \mathbb{E}_*[\Psi_i \cdot \Psi_j]} \\
&\overset{(b)}{\leq} \frac{k_* \cdot \mathbb{E}_*[\Psi_i + \Psi_j]}{\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \cdot \mathbb{E}_s[\Psi_i \cdot \Psi_j]} \\
&\overset{(c)}{\leq} \frac{k_*}{\delta}
\end{aligned}
$$

where (a) follows from Proposition 3, Claim 1; (b) follows from $\mathbb{E}_*[\Psi_i \cdot \Psi_j] \geq \frac{1}{k_*} \mathbb{E}_s[\Psi_i \cdot \Psi_j]$; and (c) is from $i,j \in \mathcal{E}_s^\delta$. Next, we obtain by Proposition 3 that

$$
\begin{aligned}
\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\|^2}{\sigma^2} &\leq 2k_* \cdot \frac{\mathbb{E}_*[\Psi_i] + 1}{\mathbb{E}_s[\Psi_i \cdot \Psi_j]} \\
&= 2k_* \cdot \frac{\mathbb{E}_*[\Psi_i] + 1}{\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \cdot \mathbb{E}_s[\Psi_i \cdot \Psi_j]} \cdot \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \\
&\leq 2k_* \cdot \frac{2}{\delta} \cdot \frac{k_*}{\delta} \\
&= \left(\frac{2k_*}{\delta}\right)^2.
\end{aligned}
$$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Approximation error bound for local minima of L   Here, we argue that the mean estimates $\boldsymbol{\beta}_i$ (for $i \in [k]$) in the so-called 'one-fits-many' configuration are well approximated by the barycenter of several true component means, $\boldsymbol{\theta}_s^*$. We recall the definition of $\mathcal{A}_i^\delta$ from (35). With aid of the set $\mathcal{A}_i^\delta$, we state an approximation error bound as follows.

Proposition 4 (Proximity bound, II). *Let $\boldsymbol{B} \in \mathbb{R}^{d \times k}$ be a local minimum of L. If $\delta \in \mathbb{R}_+$ satisfies that $\frac{4k_* \cdot \sigma}{\Delta_{\min}} < \delta \leq \frac{1}{18(\sqrt{2\pi}+1) \cdot k^4}$, then for all $i \in [k]$ with $\mathcal{A}_i^\delta \neq \emptyset$,*

$$
\frac{1}{\sigma} \left\| \boldsymbol{\beta}_i - \frac{1}{|\mathcal{A}_i^\delta|} \sum_{s \in \mathcal{A}_i^\delta} \boldsymbol{\theta}_s^* \right\| \leq 18 \left(\sqrt{2\pi} + 1 \right) \cdot \left(\frac{5k_*}{|\mathcal{A}_i^\delta|} \cdot \frac{D_{\text{cell}}(\boldsymbol{B})}{\sigma} + k \cdot \frac{\widetilde{\Delta}_{\text{cell}}^{i,\delta}(\boldsymbol{B})}{\sigma} \right) \cdot \delta + \frac{4k_*}{|\mathcal{A}_i^\delta|} \cdot \frac{1}{\delta}
$$

$$
+ \left\{ \frac{1}{|\mathcal{A}_i^\delta|} \cdot \left(2k_* \cdot (\sqrt{d} + 4) + \sqrt{\frac{2}{\pi}} \right) + 3 \right\}. \tag{37}
$$

where $D_{\text{cell}}(\boldsymbol{B}) := \max_{j \in [k]} \max_{s \in [k_*]: \boldsymbol{\theta}_s^* \in \mathcal{V}_j} \| \boldsymbol{\theta}_s^* - \boldsymbol{\beta}_j \|$ and $\widetilde{\Delta}_{\text{cell}}^{i,\delta}(\boldsymbol{B}) := \min_{s \in \mathcal{A}_i^\delta} \max_{s' \in \mathcal{A}_i^\delta} \| \boldsymbol{\theta}_s^* - \boldsymbol{\theta}_{s'}^* \|$.

We remark that $\widetilde{\Delta}_{\text{cell}}^{i,\delta}(\boldsymbol{B}) \leq 2D_{\text{cell}}(\boldsymbol{B})$. Also, we conjecture that the $\sqrt{d}$ term in (37) is an artifact of our analysis, and thus, is removable. Our proof of Proposition 4 is deferred to Appendix D.

Properties of the covers constructed by Algorithm 1 Lastly, we discuss some properties of the collections $\{\mathcal{T}_a^\delta\}_{a=0}^q$ and $\{\mathcal{S}_a^\delta\}_{a=1}^q$ produced by Algorithm 1. Specifically, we argue that these are collections of non-empty sets that cover $[k]$ and $[k_*]$, respectively, which possess useful combinatorial properties.

Proposition 5. *Let $\delta \in \mathbb{R}_+$ and $\boldsymbol{B} \in \mathbb{R}^{d \times k}$ be a local minimum of L. Let $\mathbb{T}^\delta = \{\mathcal{T}_a^\delta\}_{a=0}^q$ and $\mathbb{S}^\delta = \{\mathcal{S}_a^\delta\}_{a=1}^q$ be the collections of sets produced by Algorithm 1. If $\frac{4k_* \cdot \sigma}{\Delta_{\min}} < \delta \leq \frac{1}{18(\sqrt{2\pi}+1) \cdot k^4}$, then $\mathbb{T}^\delta$ and $\mathbb{S}^\delta$ possess the following properties.*

1. *For all $a \in [q]$, both $\mathcal{T}_a^\delta$ and $\mathcal{S}_a^\delta$ are not empty. Moreover,*

   (a) *$|\mathcal{T}_a^\delta| = 1$ for all $a \in [q_0]$, and*

   (b) *$|\mathcal{T}_a^\delta| \geq 2$ and $|\mathcal{S}_a^\delta| = 1$ for all $a \in [q] \setminus [q_0]$.*

2. *$\{\mathcal{T}_a^\delta\}_{a=0}^q$ covers $[k]$ and $\{\mathcal{S}_a^\delta\}_{a=1}^q$ covers $[k_*]$, i.e.,*

$$
\bigcup_{a=0}^q \mathcal{T}_a^\delta = [k] \qquad and \qquad \bigcup_{a=1}^q \mathcal{S}_a^\delta = [k_*].
$$

3. *Let $a \in [q]$ and $s \in [k_*]$. If $s \in \mathcal{S}_a^\delta$, then $\iota_{\boldsymbol{B}}(\boldsymbol{\theta}_s^*) \in \mathcal{T}_a^\delta$.*

4. *$\mathbb{S}^\delta$ is a collection of disjoint sets, i.e., $\mathcal{S}_a^\delta \cap \mathcal{S}_b^\delta = \emptyset$ for all $a, b \in [q]$ such that $a \neq b$.*

5. *$\mathbb{T}^\delta$ is a collection of partially disjoint sets. That is,*

   (a) *$\mathcal{T}_a^\delta \cap \mathcal{T}_0^\delta = \emptyset$ for all $a \in [q]$;*

   (b) *$\mathcal{T}_a^\delta \cap \mathcal{T}_b^\delta = \emptyset$ for all $a, b \in [q_0]$ such that $a \neq b$, and*

   (c) *$\mathcal{T}_a^\delta \cap \mathcal{T}_b^\delta = \emptyset$ for all $a, b \in [q] \setminus [q_0]$ such that $a \neq b$.*

We defer the proof of Proposition 5 to Appendix E.

Remark 4. Proposition 5 states that $\mathbb{S}^\delta$ is a partition of $[k_*]$. Nevertheless, $\mathbb{T}^\delta$ falls short of being a partition of $[k]$ for two reasons: (i) $\mathcal{T}_0^\delta$ may or may not be empty; and (ii) we do not have disjointness for index pairs $(a, b) \in [q_0] \times ([q] \setminus [q_0])$. Indeed, this motivated us to define the notion of quasi-partitions as in Definition 4. While our current analysis cannot resolve these issues, we conjecture that $\mathcal{T}_0^\delta = \emptyset$ and $\mathcal{T}_a^\delta \cap \mathcal{T}_b^\delta = \emptyset$ for all $(a, b) \in [q_0] \times ([q] \setminus [q_0])$, i.e., $\mathbb{T}^\delta \setminus \{\mathcal{T}_0^\delta\}$ forms a partition of $[k]$.

7.2.2 Completing the Proof of Theorem 4

Proof of Theorem 4. We prove the four claims one by one, using the propositions stated in Section 7.2.1.

    Proof of Claim 1 (Simple quasi-partition). This is straightforward from Proposition 5.

    Proof of Claim 2 (Mutual exclusiveness). First, suppose that $i \in \mathcal{T}_a^\delta$ for some $a \in [q_0]$. Then $\mathcal{A}_i^\delta \neq \emptyset$, and thus, $\boldsymbol{\beta}_j \neq \boldsymbol{\beta}_i$ for all $j \in [k] \setminus \{i\}$ by Proposition 2, Claim 1.

    Second, suppose that $i \in \mathcal{T}_a^\delta$ for some $a \in [q] \setminus [q_0]$. Then $\mathcal{S}_a^\delta = \{s_a\}$ for some $s_a \in [k_*]$. Let $b \in [q]_0 \setminus \{a\}$ and assume that there exists $j \in \mathcal{T}_b^\delta \setminus \{i\}$ such that $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$. Then $\psi_i(\boldsymbol{x}) = \psi_j(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$, and thus, $j \in \mathcal{T}_a^\delta$ because $i \in \mathcal{E}_{s_a}^\delta$ implies $j \in \mathcal{E}_{s_a}^\delta$. It follows that $j \in \mathcal{T}_a^\delta \cap \mathcal{T}_b^\delta$, however, this contradicts the disjointness property of $\mathbb{T}^\delta$ proved above (Claim 1 of this theorem).

    Proof of Claim 3 (Approximation error bounds).

(a) For $a \in [q_0]$, notice that $\mathcal{T}_a^\delta = \{i_a\}$ and $\mathcal{S}_a^\delta = \mathcal{A}_{i_a}^\delta$. Then the inequality (32) immediately follows from Proposition 4, cf. (37).

(b) For $a \in [q] \setminus [q_0]$, note that $\mathcal{S}_a^\delta = \{s_a\}$ and $\mathcal{T}_a^\delta = \mathcal{E}_{s_a}^\delta = \left\{ i \in [k] : \max_{j \in [k] \setminus \{i\}} \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \cdot \mathbb{E}_{s_a} [\Psi_i \Psi_j] \geq \delta \right\}$. Thus, we obtain the inequality (33) by applying Corollary 7.

    Proof of Claim 4 (Weak cross-associations). First, suppose that $s \in \mathcal{S}_a^\delta$ for some $a \in [q_0]$, and let i_a denote the unique element in $\mathcal{T}_a^\delta$. We observe that $s \in \mathcal{A}_{i_a}^\delta$ by construction of the sets $\mathcal{T}_a^\delta$ and $\mathcal{S}_a^\delta$ in Algorithm 1. Then it follows from Proposition 2 that for all $i \in [k] \setminus \{i_a\}$,

$$\mathbb{E}_s [\Psi_i] \leq 1 - \mathbb{E}_s [\Psi_{i_a}] \leq 9 \left(\sqrt{2\pi} + 1 \right) \cdot k^4 \cdot \delta, \quad \text{and}$$

$$\mathbb{P}_s (\mathcal{V}_i) \leq \mathbb{P}_s (\mathcal{V}_{i_a}^c) \leq 3 \left(\sqrt{2\pi} + 1 \right) \cdot k^4 \cdot \delta.$$

    Second, suppose that $s \in \mathcal{S}_a^\delta$ for some $a \in [q] \setminus [q_0]$. Let s_a denote the unique element in $\mathcal{S}_a^\delta$. By Claim 3 of Proposition 5, if $\iota_{\boldsymbol{B}}(\boldsymbol{\theta}_s^*) = i$, then $i \in \mathcal{T}_a^\delta$. Then we observe that if $i \in [k] \setminus \mathcal{T}_a^\delta$, then (1) $\iota_{\boldsymbol{B}}(\boldsymbol{\theta}_s^*) \neq i$, and thus, $\boldsymbol{\theta}_{s_a}^* \notin \mathrm{int}\, \mathcal{V}_i$; and (2) $\max_{j \in [k] \setminus \{i\}} \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \cdot \mathbb{E}_{s_a} [\Psi_i \cdot \Psi_j] < \delta$ by definition of $\mathcal{E}_{s_a}^\delta$. Therefore, it follows from Proposition 1 (Corollary 6) that for all $i \in [k] \setminus \mathcal{T}_a^\delta$,

$$\mathbb{E}_{s_a} [\Psi_i] \leq 9 \left(\sqrt{2\pi} + 1 \right) \cdot k^3 \cdot \delta, \quad \text{and}$$

$$\mathbb{P}_{s_a} (\mathcal{V}_i) \leq 3 \left(\sqrt{2\pi} + 1 \right) \cdot k^3 \cdot \delta.$$

$\square$

7.3 Obtaining Theorem 2 from Theorem 4

7.3.1 Remarks that Simplify Theorem 4

In this section, we make three remarks on Theorem 4, which will be used to simplify the conditions and conclusions of the theorem. First of all, we remark on the minimum separation between the components in the true Gaussian mixture model required to apply Theorem 4. Second, we discuss the reduction of dimensionality from the dimension of ambient space, d, to the effective dimension (Definition 5) of the given problem instance, d_{eff}, which leads to a significant improvement of approximation error bound (32) when $k_* \ll d$. Lastly, we determine the optimal value for the auxiliary parameter $\delta \in \mathbb{R}_+$ that approximately minimizes the approximation error bound (32).

Minimum requirement on the signal-to-noise ratio Note that Theorem 4 is vacuous unless there exists $\delta \in \mathbb{R}_+$ satisfies $\frac{4k_* \cdot \sigma}{\Delta_{\min}} < \delta \leq \frac{1}{18(\sqrt{2\pi}+1)\cdot k^4}$. This requires the minimum signal-to-noise ratio, cf. (4), to exceed a certain threshold. Specifically, in order to apply Theorem 4, we need the minimum SNR requirement (18) to be satisfied:

$$\frac{\Delta_{\min}}{\sigma} > 72(\sqrt{2\pi}+1) \cdot k_* \cdot k^4.$$

Effective dimensionality For any given instance $\mathtt{GMM}(k, \boldsymbol{\Theta}^*)$ of Gaussian mixture learning problem, we define its effective dimension as follows.

Definition 5. *Given a Gaussian mixture learning problem instance* $\mathtt{GMM}(k, \boldsymbol{\Theta}^*)$, *its effective dimension, denoted by* d_{eff}, *is defined as*

$$d_{\mathrm{eff}} := \dim \boldsymbol{\Theta}^* = \dim \operatorname{span}\{\boldsymbol{\theta}_s^* : \ s \in [k_*]\}. \tag{38}$$

Note that $d_{\mathrm{eff}} \leq d$ by definition. It follows from Corollary 2 that if $\boldsymbol{B} \in \mathbb{R}^{d\times k}$ is a stationary point of $L(\cdot \mid \boldsymbol{\Theta}^*)$, then $\boldsymbol{\beta}_i \in V_{\boldsymbol{\Theta}^*}$ for all $i \in [k]$, where $V_{\boldsymbol{\Theta}^*} := \operatorname{span}\{\boldsymbol{\theta}_s^* : \ s \in [k_*]\}$. If $d_{\mathrm{eff}} < d$, then we may choose an (ordered) orthonormal basis of $\mathbb{R}^d$ whose first d_{eff} basis vectors span the subspace $V_{\boldsymbol{\Theta}^*}$. Under the coordinate system specified by this basis, we can represent $\boldsymbol{\theta}_s^* = (\boldsymbol{\theta}_s^{*\prime}, 0)$ with $\boldsymbol{\theta}_s^{*\prime} \in \mathbb{R}^{d_{\mathrm{eff}}}$ for all $s \in [k_*]$. Likewise, for a stationary point $\boldsymbol{B} \in \mathbb{R}^{d\times k}$ of L, we can represent its elements as $\boldsymbol{\beta}_i = (\boldsymbol{\beta}_i', 0)$ with $\boldsymbol{\beta}_i' \in \mathbb{R}^{d_{\mathrm{eff}}}$, for all $i \in [k]$.

Then we observe that the probability density function of an isotropic Gaussian distribution is coordinate-wisely decomposable. That is, for any $\boldsymbol{u}, \boldsymbol{x} \in \mathbb{R}^d$ and any $\sigma \in \mathbb{R}_+$,

$$\phi(\boldsymbol{x} \mid \boldsymbol{u}, \sigma^2 \boldsymbol{I}_d) = \prod_{i=1}^{d} \phi(x_1 \mid u_1, \sigma^2).$$

Under the coordinate system specified above,

$$\begin{aligned}
f_s^*(\boldsymbol{x}) &= \phi(\boldsymbol{x} \mid \boldsymbol{\theta}_s^*, \sigma^2 \boldsymbol{I}_d) \\
&= \phi(\boldsymbol{x}_{1:d_{\mathrm{eff}}} \mid \boldsymbol{\theta}_s^{*\prime}, \sigma^2 \boldsymbol{I}_{d_{\mathrm{eff}}}) \cdot \phi(\boldsymbol{x}_{d_{\mathrm{eff}}+1:d} \mid 0, \sigma^2 \boldsymbol{I}_{d-d_{\mathrm{eff}}}), \\
f_i(\boldsymbol{x}) &= \phi(\boldsymbol{x} \mid \boldsymbol{\beta}_i, \sigma^2 \boldsymbol{I}_d) \\
&= \phi(\boldsymbol{x}_{1:d_{\mathrm{eff}}} \mid \boldsymbol{\beta}_i', \sigma^2 \boldsymbol{I}_{d_{\mathrm{eff}}}) \cdot \phi(\boldsymbol{x}_{d_{\mathrm{eff}}+1:d} \mid 0, \sigma^2 \boldsymbol{I}_{d-d_{\mathrm{eff}}}),
\end{aligned}$$

for all $s \in [k_*]$ and for all $i \in [k]$, respectively. Therefore, for all $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) \in V_{\boldsymbol{\Theta}^*}^k \cong \mathbb{R}^{d_{\mathrm{eff}}\times k}$,

$$L(\boldsymbol{X} \mid \boldsymbol{\Theta}^*) = L(\boldsymbol{X}' \mid \boldsymbol{\Theta}^{*\prime})$$

where $\boldsymbol{\Theta}^{*\prime} = (\boldsymbol{\theta}_1^{*\prime} \ldots, \boldsymbol{\theta}_{k_*}^{*\prime}) \in \mathbb{R}^{d_{\mathrm{eff}}\times k_*}$. Thus, a local minimum of $L(\cdot \mid \boldsymbol{\Theta}^*)$ is a local minimum of $L(\cdot \mid \boldsymbol{\Theta}^{*\prime})$ (with a change of coordinates), and vice versa. Consequently, we may replace the dimension d in (32) of Theorem 4 by the effective dimension d_{eff}, which satisfies $d_{\mathrm{eff}} \leq \min\{d, k_*\}$.

Remark 5. If $d_{\mathrm{eff}} < d$ and $\boldsymbol{B} \in \mathbb{R}^{d\times k}$ is a stationary point of $L(\cdot \mid \boldsymbol{\Theta}^*)$, then for every $i \in [k]$, the i-th Voronoi cell of B takes the form $\mathcal{V}_i \cong \mathcal{V}_i' \times \mathbb{R}^{d-d_{\mathrm{eff}}}$ for some $\mathcal{V}_i' \subseteq V_{\boldsymbol{\Theta}^*}$.

Approximately optimal choice of parameter δ Lastly, we optimize the value of δ to minimize the approximation error bounds (32) and (33). If $q_0 = 0$, there does not exists "one-fits-many." Thus, we may simply choose the largest value $\delta = \frac{1}{18(\sqrt{2\pi}+1)\cdot k^4}$ to optimize the approximation error bound (33).

If $q_0 \neq 0$, we choose a particular value of δ that minimizes the upper bound (32). Observe, e.g., by the AM-GM inequality that for any $A, B > 0$, $\min_{\delta>0}\left\{A\delta + \frac{B}{\delta}\right\} = 2\sqrt{AB}$, which is achieved by $\delta^\star := \arg\min_{\delta>0}\left\{A\delta + \frac{B}{\delta}\right\} = \sqrt{B/A}$. With the choice of $A = 18\left(\sqrt{2\pi}+1\right)\cdot k^3 \cdot (5k_*+2k)\cdot \frac{D_{\mathrm{cell}}(\boldsymbol{B})}{\sigma}$ and $B = \frac{4k_*}{|\mathcal{S}_a^\delta|}$, we obtain

$$\begin{aligned}
\delta^\star &= \sqrt{\frac{B}{A}} \\
&= \min_{a\in[q_0]}\left\{\left(\frac{2k_*}{9\left(\sqrt{2\pi}+1\right)\cdot k^3 \cdot (5k_*+2k)}\right)^{\frac{1}{2}} \cdot \left(\frac{1}{|\mathcal{S}_a^\delta|} \cdot \frac{\sigma}{D_{\mathrm{cell}}(\boldsymbol{B})}\right)^{\frac{1}{2}}\right\}. \tag{39}
\end{aligned}$$

Next, we ensure $\delta = \delta^\star$ satisfies the requirement $\frac{4k_* \cdot \sigma}{\Delta_{\min}} < \delta \leq \frac{1}{18(\sqrt{2\pi}+1)\cdot k^4}$ in Theorem 4. Firstly, since $|\mathcal{S}_a^\delta| \leq k_*$ for all $a \in [q_0]$, the lower bound $\frac{4k_* \cdot \sigma}{\Delta_{\min}} < \delta^\star$ is satisfied if

$$\frac{4k_* \cdot \sigma}{\Delta_{\min}} < \sqrt{\frac{2}{9\left(\sqrt{2\pi}+1\right)\cdot k^3 \cdot (5k_* + 2k)} \cdot \frac{\sigma}{D_{\mathrm{cell}}(\boldsymbol{B})}},$$

or equivalently, if the inequality in (19) holds:

$$\frac{D_{\mathrm{cell}}(\boldsymbol{B})}{\sigma} < \frac{1}{72\left(\sqrt{2\pi}+1\right)\cdot k_*^2 \cdot k^3 \cdot (5k_* + 2k)} \cdot \left(\frac{\Delta_{\min}}{\sigma}\right)^2.$$

Secondly, the upper bound $\delta^\star \leq \frac{1}{18(\sqrt{2\pi}+1)\cdot k^4}$ is automatically satisfied when $\frac{\Delta_{\min}}{\sigma} > 72(\sqrt{2\pi}+1)\cdot k_* \cdot k^4$ as in (18). To see this, it suffices to notice that $D_{\mathrm{cell}}(\boldsymbol{B}) \geq \frac{1}{2}\Delta_{\min}$ when $q_0 \geq 1$. Then we observe that

$$\delta^\star = \left(\frac{2k_*}{9\left(\sqrt{2\pi}+1\right)\cdot k^3 \cdot (5k_* + 2k)} \cdot \frac{1}{|\mathcal{S}_a^\delta|} \cdot \frac{\sigma}{D_{\mathrm{cell}}(\boldsymbol{B})}\right)^{\frac{1}{2}}$$

$$\overset{(a)}{\leq} \left(\frac{4k_*}{9\left(\sqrt{2\pi}+1\right)\cdot k^3 \cdot (5k_* + 2k)} \cdot \frac{\sigma}{\Delta_{\min}}\right)^{\frac{1}{2}}$$

$$\overset{(b)}{\leq} \frac{1}{18\left(\sqrt{2\pi}+1\right)\cdot k^4} \cdot \left(\frac{2k}{5k_* + 2k}\right)^{\frac{1}{2}}$$

$$\leq \frac{1}{18\left(\sqrt{2\pi}+1\right)\cdot k^4},$$

where (a) follows from $|\mathcal{S}_a^\delta| \geq 1$ and $D_{\mathrm{cell}}(\boldsymbol{B}) \geq \frac{1}{2}\Delta_{\min}$, and (b) is due to (18).

All in all, if the conditions (18) and (19) are met, then the approximation error bounds (32) and (33) in Theorem 4 reduce to the following: $\forall a \in [q_0]$ and $\forall a \in [q] \setminus [q_0]$ respectively,

$$\frac{1}{\sigma}\left\|\boldsymbol{\beta}_{i_a} - \frac{1}{|\mathcal{S}_a^\delta|}\sum_{s \in \mathcal{S}_a^\delta} \boldsymbol{\theta}_s^*\right\| \lesssim \left(\frac{D_{\mathrm{cell}}(\boldsymbol{B})}{\sigma}\cdot k_*^2 \cdot k^3 \cdot (k_* + k)\right)^{1/2} + k_* \cdot d^{1/2}$$

$$\lesssim \frac{\Delta_{\min}}{\sigma} + k_* \cdot d^{1/2},$$

$$\frac{1}{\sigma}\left\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_{s_a}^*\right\| \lesssim \left(\frac{D_{\mathrm{cell}}(\boldsymbol{B})}{\sigma}\cdot k_*^2 \cdot k^3 \cdot (k_* + k)\right)^{1/2}$$

$$\lesssim \frac{\Delta_{\min}}{\sigma}.$$

Likewise, the cross-association bounds in (34) are bounded from above by

$$k^4 \cdot \delta^\star \lesssim \left(\frac{k^5 \cdot k_*}{k_* + k} \cdot \frac{\sigma}{D_{\mathrm{cell}}(\boldsymbol{B})}\right)^{1/2} \lesssim \left(\frac{k^5 \cdot k_*}{k_* + k} \cdot \frac{\sigma}{\Delta_{\min}}\right)^{1/2}.$$

Here, the last inequality follows from the observation that if $q_0 \neq 0$, then $D_{\mathrm{cell}}(\boldsymbol{B}) \geq \frac{1}{2}\Delta_{\min}$.

7.3.2 Distillation of Partitions from Quasi-partitions

We extract partitions of $[k]$ and $[k_*]$ by slightly altering the quasi-partitions $\mathbb{T}^\delta$ and $\mathbb{S}^\delta$ in Theorem 4.

Alteration process to distill decomposable partitions Let $\boldsymbol{B}$ be a local minimum of $L(\,\cdot\,|\,\boldsymbol{\Theta}^*)$, and suppose that we are given $q, q_0, \mathbb{T}, \mathbb{S}$ as stated in Theorem 4. We define the set of 'problematic indices' by letting

$$\mathcal{A}^{\mathrm{prob}} := \left\{ a \in [q] \setminus [q_0] : \mathcal{B}_a^{\mathrm{conf}} \neq \emptyset \right\} \tag{40}$$

where

$$\mathcal{B}_a^{\mathrm{conf}} := \left\{ b \in [q_0] : \mathcal{T}_a \cap \mathcal{T}_b \neq \emptyset \right\}.$$

We resolve the conflict in $\mathcal{T}_b$ for each of $a \in \mathcal{A}^{\mathrm{prob}}$ sequentially with a simple 'surgery' (Algorithm 2).

Algorithm 2 Alteration process to distill fully decomposable partitions

---

Input: $q_0, q, \mathbb{T}, \mathbb{S}$ that satisfy the conclusions of Theorem 4
Output: $q_0, q, \mathbb{T}, \mathbb{S}$; these may be different from the inputs

1: **for** $a \in \mathcal{A}^{\mathrm{prob}}$ **do**
2:      **if** $\mathcal{T}_a \setminus \bigcup_{b \in \mathcal{B}_a^{\mathrm{conf}}} \mathcal{T}_b \neq \emptyset$ **then**
3:          $\mathcal{T}_a \leftarrow \mathcal{T}_a \setminus \bigcup_{b \in \mathcal{B}_a^{\mathrm{conf}}} \mathcal{T}_b$ {$\mathbb{T}$ is also updated accordingly}
4:      **else**
5:          $i \leftarrow \arg\min \mathcal{T}_a$
6:          $b_0 \leftarrow$ the unique element $b \in \mathcal{B}_a^{\mathrm{conf}}$ such that $\mathcal{T}_b = \{i\}$
7:          $\mathcal{S}_{b_0} \leftarrow \mathcal{S}_{b_0} \cup \mathcal{S}_a$ {Merge $\mathcal{S}_a$ to $\mathcal{S}_{b_0}$}
8:          $\mathbb{T} \leftarrow \mathbb{T} \setminus \{\mathcal{T}_a\}, \mathbb{S} \leftarrow \mathbb{S} \setminus \{\mathcal{S}_a\}$
9:          **for** $a' \in [q] \setminus [a]$ **do**
10:            $\mathcal{T}_{a'-1} \leftarrow \mathcal{T}_{a'}, \mathcal{S}_{a'-1} \leftarrow \mathcal{S}_{a'}$ {Update the group index for $a' > a$}
11:          **end for**
12:          $q \leftarrow q - 1$
13:      **end if**
14: **end for**

Properties of the output of Algorithm 2 Let $(q, q_0, \mathbb{T}, \mathbb{S})$ be any tuple that satisfies the conclusions of Theorem 4, and let $(q', q_0', \mathbb{T}', \mathbb{S}') = \texttt{Alg}(q, q_0, \mathbb{T}, \mathbb{S})$ where $\texttt{Alg}$ is Algorithm 2. We observe that $(q', q_0', \mathbb{T}', \mathbb{S}')$ satisfy the conclusions in Claims 1 through 3 of Theorem 4, and moreover, $\mathbb{T}'$ is a partition of $[k]$. This observation is formally stated in the following proposition, whose proof is deferred to Appendix F.

Proposition 6. *Let $k, k_* \in \mathbb{N}$, $\boldsymbol{\Theta}^* \in \mathbb{R}^{d \times k_*}$, and $\boldsymbol{B} \in \mathbb{R}^{d \times k}$ be a local minimum of $L(\,\cdot\,|\,\boldsymbol{\Theta}^*)$. Let $q, q_0 \in \mathbb{N}$, and $\mathbb{T}, \mathbb{S}$ be collections of subsets of $[k], [k_*]$, respectively. If $q, q_0, \mathbb{T}, \mathbb{S}$ satisfy the conclusions of Theorem 4, then $(q', q_0', \mathbb{T}', \mathbb{S}') = \texttt{Alg}(q, q_0, \mathbb{T}, \mathbb{S})$ where $\texttt{Alg}$ is Algorithm 2 has the following properties:*

1. *Claim 1 in Theorem 4 holds for $(q', q_0', \mathbb{T}', \mathbb{S}')$. Moreover, $\mathbb{T}'$ is a partition of $[k]$.*

2. *Claim 2 in Theorem 4 holds for $(q', q_0', \mathbb{T}', \mathbb{S}')$.*

3. *Claim 3 in Theorem 4 holds for $(q', q_0', \mathbb{T}', \mathbb{S}')$.*

7.3.3   Derivation of Theorem 2 from Theorem 4

Proof of Theorem 2. Given $\boldsymbol{B}$, we choose $\delta = \delta^\star$ per (39). Then, we let $\tilde{q}, \tilde{q}_0, \tilde{\mathbb{T}}, \tilde{\mathbb{S}}$ be the outputs of the construction algorithm (Algorithm 1), and let $q, q_0, \mathbb{T}, \mathbb{S} = \texttt{Alg}\left(\tilde{q}, \tilde{q}_0, \tilde{\mathbb{T}}, \tilde{\mathbb{S}}\right)$ where $\texttt{Alg}$ is the distillation algorithm (Algorithm 2). With these, Theorem 2 immediately follows from Theorem 4 and Proposition 6. $\qquad\square$

8   A Fine-grained Analysis for One-dimensional GMM

In this section, we prove Theorem 3 (and Corollary 5). Our proof consists of four steps outlined here:

1. **(Step 1)** Enumerate all possible configurations of tuples $(q, q_0, |\mathcal{T}_0|)$ that satisfy the simple quasi-partition property for local minima; see Claim 1 of Theorem 2.

2. **(Step 2)** For each $(q, q_0, |\mathcal{T}_0|)$, identify all possible compositions of $\mathbb{T}$ and $\mathbb{S}$ that satisfy the near-empty cross-association property established in Claim 4 of Theorem 2.

3. **(Step 3)** For each composition, establish refined approximation error bounds individually.

4. **(Step 4)** Regroup the compositions for the convenience of presentation.

In Section 8.1, we state and prove useful technical lemmas. Thereafter, in Section 8.2, we complete the proof of Theorem 3 by detailing Steps 1 through 3. Lastly, in Section 8.3, we prove Corollary 5 following essentially the same, yet simpler, argument in the proof of Theorem 3.

8.1 Useful Technical Lemmas

8.1.1 Additional Notation and Basic Facts about Gaussians

Let $\phi : \mathbb{R} \to \mathbb{R}$ denote the probability density of the standard Gaussian, i.e.,

$$\phi(x) := \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \tag{41}$$

Let Φ and Q denote the Gaussian cumulative distribution function and the Gaussian Q-function, respectively:

$$\Phi(t) := \int_{-\infty}^{t} \phi(z) \, \mathrm{d}z \quad \text{and} \quad Q(t) := \int_{t}^{\infty} \phi(z) \, \mathrm{d}z. \tag{42}$$

It is well known [7] that

$$\frac{t}{t^2 + 1} \cdot \phi(t) \leq Q(t) \leq \frac{1}{t} \cdot \phi(t), \qquad \forall t > 0. \tag{43}$$

Lemma 4 (Gaussian tail bounds). *For any $t \geq 0$,*

$$\frac{1}{t+1}\phi(t) \leq \frac{1}{t+\sqrt{t^2+4}}\phi(t) \leq Q(t) \leq \sqrt{2\pi} \cdot \phi(t).$$

Proof of Lemma 4. The upper bound is standard. The lower bounds can be found in [1, Formula 7.1.13]. $\square$

Also, it can be verified that for all $t \in \mathbb{R}$,

$$\int_{-\infty}^{t} z \cdot \phi(z) \, \mathrm{d}z = -\phi(t) \quad \text{and} \quad \int_{t}^{\infty} z \cdot \phi(z) \, \mathrm{d}z = \phi(t). \tag{44}$$

A more extensive list of Gaussian integrals can be found in [30] for example.

Moreover, we observe a simple equation, namely, for any $x, y, \delta \in \mathbb{R}$,

$$\frac{\phi(x+\delta)}{\phi(y+\delta)} \cdot \frac{\phi(y)}{\phi(x)} = e^{\delta(y-x)}. \tag{45}$$

Lastly, we make a simple observation that for any $\alpha, x_0 \in \mathbb{R}$,

$$e^{-\frac{x^2}{2}} \cdot e^{\alpha(x-x_0)} = e^{-\frac{(x-\alpha)^2}{2}} \cdot e^{\frac{\alpha}{2}(\alpha-2x_0)}. \tag{46}$$

This readily implies that $\phi(x) \cdot e^{\alpha(x-x_0)} = \phi(x-\alpha) \cdot e^{\frac{\alpha}{2}(\alpha-2x_0)}$.

8.1.2 Useful Technical Lemmas

For $\alpha, \beta \in \mathbb{R}$, we define $\psi_{\alpha,\beta} : \mathbb{R} \to \mathbb{R}$ be a function such that

$$\psi_{\alpha,\beta}(x) = \frac{e^{-(x-\beta)^2/2}}{e^{-(x-\alpha)^2/2} + e^{-(x-\beta)^2/2}}. \tag{47}$$

Lemma 5. *Let $\alpha, \beta \in \mathbb{R}$ such that $\alpha < 0 < \beta$ and $\beta \geq |\alpha|$. Then*

$$\mathbb{E}_{x \sim \mathcal{N}(0,1)}\left[\psi_{\alpha,\beta}(x) \cdot (\beta - x)\right] \geq \beta \cdot \mathbb{E}_0\left[\psi_{\alpha,\beta}(x) \cdot \left(2\psi_{\alpha,\beta}(x) - 1\right)\right]$$

$$\geq \frac{\beta}{8} \cdot Q\left(\frac{\alpha + \beta}{2} - \frac{1}{\alpha}\right).$$

Remark 6. We note that it is possible to obtain a tighter lower bound by refining the proof of Lemma 5. Specifically, we can derive a tighter lower bound for $\mathbb{E}_0\left[\psi_{\alpha,\beta}(x) \cdot \left(2\psi_{\alpha,\beta}(x) - 1\right)\right]$ by continuing from the expression in (145) as follows[4]:

$$\mathbb{E}_0\left[\psi_{\alpha,\beta}(x) \cdot \left(2\psi_{\alpha,\beta}(x) - 1\right)\right] \geq \frac{1}{\sqrt{2\pi}} \int_{1/\delta}^{\infty} g(z+c) \cdot \left[h_1(z) - h_2(z)\right] \, \mathrm{d}z$$

$$\geq \frac{1}{\sqrt{2\pi}} \int_{1/\delta}^{\infty} \frac{1}{4} e^{(z-\delta)^2/2} \cdot \left[1 - e^{2\alpha z}\right] \cdot h_1(z) \, \mathrm{d}z$$

$$\geq \frac{1}{4\sqrt{2\pi}} \int_{1/\delta}^{\infty} e^{-(z-c)^2/2} \, \mathrm{d}z - \frac{1}{4\sqrt{2\pi}} \cdot \frac{\phi(c)}{\phi(c+2\alpha)} \int_{1/\delta}^{\infty} e^{-(z-c-2\alpha)^2/2} \, \mathrm{d}z$$

$$= \frac{1}{4}\left[Q\left(\frac{1}{\delta} - c\right) - \frac{\phi(c)}{\phi(c+2\alpha)} \cdot Q\left(\frac{1}{\delta} - c - 2\alpha\right)\right]. \tag{48}$$

Lemma 6 (Variance lower bound). *Let $\alpha, \beta \in \mathbb{R}$ such that $\alpha \neq \beta$, and let $\psi_{\alpha,\beta}$ be the function as defined in (47). Then*

$$\mathrm{Var}_{x \sim \mathcal{N}(0,1)}\left(\psi_{\alpha,\beta}(x)\right) \geq \frac{e^{-4\left(\frac{|\alpha+\beta|}{2} + |\beta - \alpha|\right)^2}}{32\sqrt{2\pi}}(\alpha - \beta)^2 \cdot \left(\frac{|\alpha+\beta|}{2} + |\beta - \alpha|\right)^3.$$

Lemma 7 (Exponential association). *Let $\sigma = 1$, $d = 1$, $s \in [k_*]$ and $i_s := \arg\min_{i \in [k]} |\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*|$. For $j \in [k] \setminus \{i_s\}$,*

$$\mathbb{E}_s\left[\Psi_j\right] \leq \left(1 + \frac{4}{\sqrt{2\pi} \cdot \delta_j}\right) e^{-\frac{\delta_j^2}{32}}. \tag{49}$$

where $\delta_j := |\boldsymbol{\beta}_j - \boldsymbol{\theta}_s^*| - |\boldsymbol{\beta}_{i_s} - \boldsymbol{\theta}_s^*|$.

Lemma 8 (Exponential accuracy). *Let $\sigma = 1$ and $d = 1$. Suppose that $\boldsymbol{B}$ be a stationary point of L. Let $i \in [k]$, $\mathcal{S} \subseteq [k_*]$, and $\delta \in \mathbb{R}_+$ such that $\delta \geq \max\left\{\frac{4}{\sqrt{2\pi}}, 8\sqrt{\log(2\sqrt{2} \cdot kk_*)}\right\}$. If*

$$\begin{aligned}
\min_{j \in [k] \setminus \{i\}} \|\boldsymbol{\beta}_j - \boldsymbol{\theta}_s^*\| - \|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\| &\geq \delta, & \forall s \in \mathcal{S}, \\
\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\| - \min_{j \in [k] \setminus \{i\}} \|\boldsymbol{\beta}_j - \boldsymbol{\theta}_s^*\| &\geq \delta, & \forall s \in [k_*] \setminus \mathcal{S},
\end{aligned} \tag{50}$$

then

$$\left\|\boldsymbol{\beta}_i - \frac{1}{|\mathcal{S}|}\sum_{s \in \mathcal{S}} \boldsymbol{\theta}_s^*\right\| \leq 4kk_* \Delta_{\max} \cdot e^{-\frac{\delta^2}{64}}.$$

---

[4]See (144) and (145) in Appendix G.1 (the proof of Lemma 5) for the definition of the functions g, h_1, h_2 and more details.

8.2 Completing the Proof of Theorem 3

This entire subsection is dedicated to the proof of Theorem 3, which follows the four-step strategy outlined in the preamble of Section 8. Here we begin by making several preliminary observations. Thereafter, we prove Theorem 3 by detailing Step 1 (Section 8.2.1), Step 2 (Section 8.2.2), and Step 3 (Section 8.2.3) of the outlined argument.

Note that we may assume $\sigma = 1$ by treating Δ/σ as new Δ; thus, in the rest of the proof, we assume $\sigma = 1$ and write Δ in place of Δ/σ. If $\Delta > 2^{16} \cdot 3^{10} \cdot (\sqrt{2\pi} + 1)$, then the approximation error bounds in Claim 3 of Theorem 2, cf. (21), reduce to the following.

- For $a \in [q_0]$,

$$
\left\| \boldsymbol{\beta}_{i_a} - \frac{1}{|\mathcal{S}_a|} \sum_{s \in \mathcal{S}_a} \boldsymbol{\theta}_s^* \right\| \leq \left(2^8 \cdot 3^8 \cdot (\sqrt{2\pi} + 1) \cdot \Delta \right)^{1/2} + 2
$$
$$
< \frac{\Delta}{32}
$$

(51)

  where i_a denotes the unique element in $\mathcal{T}_a$.

- For $a \in [q] \setminus [q_0]$, and for all $i \in \mathcal{T}_a$,

$$
\left\| \boldsymbol{\beta}_i - \boldsymbol{\theta}_{s_a}^* \right\| \leq \left(2^4 \cdot 3^8 \cdot (\sqrt{2\pi} + 1) \cdot \Delta \right)^{1/2}
$$
$$
< \frac{\Delta}{192},
$$

(52)

  where s_a denotes the unique element in $\mathcal{S}_a$.

Moreover, the cross-association bound in Claim 4 of Theorem 2, cf. (22), reduces to the following.

Let $(a, b) \in [q] \times [q]_0$ such that $a \neq b$. If $s \in \mathcal{S}_a$ and $i \in \mathcal{T}_b$, then

$$
\mathbb{E}_s \left[\Psi_i \right] < \frac{1}{2^9 \cdot 3^2} \quad \text{and} \quad \mathbb{P}_s \left(\mathcal{V}_i \right) < \frac{1}{2^9 \cdot 3^3}.
$$

(53)

8.2.1 Step 1. Preliminary Screening

First of all, we enumerate all possible configurations of parameter tuples $(q, q_0, |\mathcal{T}_0|)$ that are allowed by the quasi-partition property stated in Theorem 4. To this end, we first observe that (i) $1 \leq q \leq \min\{k, k_*\} = 3$, (ii) $0 \leq q_0 \leq q$, and (3) $0 \leq |\mathcal{T}_0| \leq k = 3$. Next, we exclude some of these parameter combinations based on the simple quasi-partition property of the collections $\mathbb{T}$ and $\mathbb{S}$, cf. claim 1 of Theorem 4.

- Suppose that $(q, q_0) = (1, 0)$. This is not allowed as $|\mathcal{S}_1| = 1$ contradicts $\mathcal{S}_1 = \bigcup_{a=1}^q \mathcal{S}_a = [k_*] = \{1, 2, 3\}$.

- Suppose that $(q, q_0) = (1, 1)$. Then $|\mathcal{T}_1| = 1$. Thus, $|\mathcal{T}_0| = 2$ because $\mathcal{T}_0 \cap \mathcal{T}_1 = \emptyset$ and $\mathcal{T}_0 \cup \mathcal{T}_1 = [k]$.

- Suppose that $(q, q_0) = (2, 0)$. This is not allowed as $|\mathcal{S}_1| = |\mathcal{S}_2| = 1$ contradicts $\mathcal{S}_1 \cup \mathcal{S}_2 = \{1, 2, 3\}$.

- Suppose that $(q, q_0) = (2, 1)$. Then $|\mathcal{T}_1| = 1$ and $|\mathcal{T}_2| \geq 2$. Observe that $\mathcal{T}_0 \cap \mathcal{T}_2 = \emptyset$ and $\bigcup_{a=0}^2 \mathcal{T}_a = [k]$. Thus, $|\mathcal{T}_0| \leq 1$.

- Suppose that $(q, q_0) = (2, 2)$. Then $|\mathcal{T}_1| = |\mathcal{T}_2| = 1$. Observe that $\mathcal{T}_a \cap \mathcal{T}_b = \emptyset$ for all $a, b \in \{0, 1, 2\}$ such that $a \neq b$, and that $\bigcup_{a=0}^2 \mathcal{T}_a = [k] = \{1, 2, 3\}$. Therefore, $|\mathcal{T}_0| = 1$.

- Suppose that $q = 3$ and $q_0 \in \{0, 1\}$. Then $|\mathcal{T}_2| \geq 2$, $|\mathcal{T}_3| \geq 2$, and $\mathcal{T}_2 \cap \mathcal{T}_3 = \emptyset$. This contradicts $\bigcup_{a=0}^3 \mathcal{T}_a = [k]$ because $|\mathcal{T}_2 \cup \mathcal{T}_3| \geq 4$. Thus, these configurations are forbidden.

- Suppose that $(q, q_0) = (3, 2)$. Then $|\mathcal{T}_1| = |\mathcal{T}_2| = 1$ and $\mathcal{T}_1 \cap \mathcal{T}_2 = \emptyset$. Since $\mathcal{T}_0 \cap \mathcal{T}_a = \emptyset$ for $a \in \{1, 2\}$ and $\bigcup_{a=0}^3 \mathcal{T}_a = [k] = \{1, 2, 3\}$, it must hold that $|\mathcal{T}_0| \leq 1$.

33

Table 1: All possible configurations of parameter tuples $(q, q_0, |\mathcal{T}_0|)$ for the collections $\mathbb{T}, \mathbb{S}$ that satisfy the simple quasi-partition property stated in Theorem 4.

q	q_0	$|\mathcal{T}_0|$	
1	0	0, 1, 2 or 3	impossible
	1	0 or 1	impossible
		2	-
		3	impossible
2	0	0, 1, 2 or 3	impossible
	1	0 or 1	-
		2 or 3	impossible
	2	0	impossible
		1	-
		2 or 3	impossible
3	0 or 1	0, 1, 2 or 3	impossible
	2	0 or 1	-
		2 or 3	impossible
	3	0	-
		1, 2 or 3	impossible

- Suppose that $(q, q_0) = (3, 3)$. Then $|\mathcal{T}_1| = |\mathcal{T}_2| = |\mathcal{T}_3| = 1$ and $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$ are mutually disjoint. Since $\mathcal{T}_0 \cap \mathcal{T}_a = \emptyset$ for $a \in \{1, 2, 3\}$ and $\bigcup_{a=0}^{3} \mathcal{T}_a = [k] = \{1, 2, 3\}$, it must hold that $|\mathcal{T}_0| = 0$.

These are summarized in Table 1. In the next step, we consider each of these configurations $(q, q_0, |\mathcal{T}_0|)$ and identify all possible compositions of $\mathbb{T}$ and $\mathbb{S}$.

8.2.2 Step 2. Identification of All Possible Compositions of $\mathbb{T}$ and $\mathbb{S}$ for Local Minima

Next, we list all possible compositions of $\mathbb{T}$ and $\mathbb{S}$ for the remaining configurations of $(q, q_0, |\mathcal{T}_0|)$. Note that we may freely relabel the indices of estimates $\{\boldsymbol{\beta}_i : i \in [k]\}$ by a permutation of $[k]$. Also, we can relabel the indices of the true component means $\{\boldsymbol{\theta}_s^* : s \in [k_*]\}$ by flipping the order: $(1, 2, 3) \mapsto (3, 2, 1)$. Thus, the possible compositions of $\mathbb{T}$ and $\mathbb{S}$ are presented up to permutation of $[k]$ and the flipping of the order in $[k_*]$.

1. $(q, q_0, |\mathcal{T}_0|) = (1, 1, 2)$. The only possibility is $\mathcal{T}_0 = \{2, 3\}$; $\mathcal{T}_1 = \{1\}$; $\mathcal{S}_1 = \{1, 2, 3\}$.

2. $(q, q_0, |\mathcal{T}_0|) = (2, 1, 0)$. There are two possibilities: (a) $|\mathcal{T}_1| = 1$ and $|\mathcal{T}_2| = 2$; or (b) $|\mathcal{T}_1| = 1$ and $|\mathcal{T}_2| = 3$. In both cases, $\mathcal{T}_0 = \emptyset$.

   (a) $\mathcal{T}_1 = \{1\}$, $\mathcal{T}_2 = \{2, 3\}$.

       i. $\mathcal{S}_1 = \{1, 2\}$, $\mathcal{S}_2 = \{3\}$.
       ii. $\mathcal{S}_1 = \{1, 3\}$, $\mathcal{S}_2 = \{2\}$.
   (b) $\mathcal{T}_1 = \{1\}$, $\mathcal{T}_2 = \{1, 2, 3\}$.

       i. $\mathcal{S}_1 = \{1, 2\}$, $\mathcal{S}_2 = \{3\}$.
       ii. $\mathcal{S}_1 = \{1, 3\}$, $\mathcal{S}_2 = \{2\}$.

3. $(q, q_0, |\mathcal{T}_0|) = (2, 1, 1)$. There are two possibilities: (a) $|\mathcal{T}_1| = 1$ and $|\mathcal{T}_2| = 1$; or (b) $|\mathcal{T}_1| = 1$ and $|\mathcal{T}_2| = 2$. However, (a) is forbidden by the requirement $|\mathcal{T}_a| \geq 2$ for $a \in [q] \setminus [q_0]$. Thus, we may write $\mathcal{T}_0 = \{3\}$, $\mathcal{T}_1 = \{1\}$, $\mathcal{T}_2 = \{1, 2\}$.

   (a) $\mathcal{S}_1 = \{1, 2\}$, $\mathcal{S}_2 = \{3\}$.
   (b) $\mathcal{S}_1 = \{1, 3\}$, $\mathcal{S}_2 = \{2\}$.

4. $(q, q_0, |\mathcal{T}_0|) = (2, 2, 1)$. We may write $\mathcal{T}_0 = \{3\}$; $\mathcal{T}_1 = \{1\}$, $\mathcal{T}_2 = \{2\}$.

34

(a) $\mathcal{S}_1 = \{1,2\}$, $\mathcal{S}_2 = \{3\}$.

(b) $\mathcal{S}_1 = \{1,3\}$, $\mathcal{S}_2 = \{2\}$.

5. $(q, q_0, |\mathcal{T}_0|) = (3,2,0)$. $\mathcal{T}_0 = \emptyset$. There are two possibilities: (a) $|\mathcal{T}_3| = 2$; or (b) $|\mathcal{T}_3| = 3$.

(a) $\mathcal{T}_1 = \{1\}$, $\mathcal{T}_2 = \{2\}$, $\mathcal{T}_3 = \{1,3\}$.

   i. $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{2\}$, $\mathcal{S}_3 = \{3\}$.
   ii. $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{3\}$, $\mathcal{S}_3 = \{2\}$.
   iii. $\mathcal{S}_1 = \{2\}$, $\mathcal{S}_2 = \{1\}$, $\mathcal{S}_3 = \{3\}$.

(b) $\mathcal{T}_1 = \{1\}$, $\mathcal{T}_2 = \{2\}$, $\mathcal{T}_3 = \{1,2,3\}$;

   i. $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{2\}$, $\mathcal{S}_3 = \{3\}$.
   ii. $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{3\}$, $\mathcal{S}_3 = \{2\}$.

6. $(q, q_0, |\mathcal{T}_0|) = (3,2,1)$. $\mathcal{T}_0 = \{3\}$; $\mathcal{T}_1 = \{1\}$, $\mathcal{T}_2 = \{2\}$, $\mathcal{T}_3 = \{1,2\}$.

(a) $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{2\}$, $\mathcal{S}_3 = \{3\}$.

(b) $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{3\}$, $\mathcal{S}_3 = \{2\}$.

7. $(q, q_0, |\mathcal{T}_0|) = (3,3,0)$. The only possibility is $\mathcal{T}_0 = \emptyset$; $\mathcal{T}_1 = \{1\}$, $\mathcal{T}_2 = \{2\}$, $\mathcal{T}_3 = \{3\}$; $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{2\}$, $\mathcal{S}_3 = \{3\}$.

Now we rule out some of these compositions via geometric considerations.

- **The compositions with $\mathcal{T}_1 = \{1\}$ and $\mathcal{S}_1 = \{1,3\}$.** We claim that it is impossible for a single estimate $\boldsymbol{\beta}_1$ to fit two non-adjacent centers $\{\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_3^*\}$. These include the compositions 2-(a)-ii, 2-(b)-ii, 3-(b) and 4-(b). Without loss of generality, we may assume $\boldsymbol{\beta}_1 \leq \boldsymbol{\beta}_2$; also, we may assume $\boldsymbol{\beta}_2 \leq \boldsymbol{\beta}_3$ in 2-(a)-ii, 2-(b)-ii. Moreover, $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$ by Theorem 2, Claim 2, and thus, $\boldsymbol{\beta}_1 < \boldsymbol{\beta}_2$. Let $b = \frac{\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2}{2}$. Then $b > \boldsymbol{\theta}_3^*$ because $\mathcal{V}_1$ is a convex set and $\iota_{\boldsymbol{B}}(\boldsymbol{\theta}_3^*) = 1$ (which implies $\boldsymbol{\theta}_3^* \in \mathcal{V}_1$; see Definition 2 for the definition of index function $\iota_{\boldsymbol{B}}$) by construction; recall Algorithm 1 and the definition of $\mathcal{A}_i^\delta$ in (35). Thus, $\mathcal{V}_2 \subseteq [b, \infty) \subseteq (\boldsymbol{\theta}_3^*, \infty)$. Therefore, $\mathbb{P}_2(\mathcal{V}_2) \leq \mathbb{P}_2((\boldsymbol{\theta}_3^*, \infty)) = Q(\Delta)$ where Q is the Gaussian Q-function, cf. (42). Using the well-known upper bound for the Q-function, cf. (43), it follows that $\mathbb{P}_2(\mathcal{V}_2) \leq Q(\Delta) \leq \frac{1}{\Delta} \frac{1}{\sqrt{2\pi}} e^{-\frac{\Delta^2}{2}} < \frac{1}{4}$. Observe that $\mathbb{P}_2(\mathcal{V}_1) < \frac{1}{2^9 \cdot 3^3} \leq \frac{1}{4}$ by (53), which follows from Theorem 2, Claim 4. Also, we can see that $\mathbb{P}_2(\mathcal{V}_3) \leq \frac{1}{4}$, either by using $\boldsymbol{\beta}_2 \leq \boldsymbol{\beta}_3$ (when $3 \in \mathcal{T}_2$), or by using (53) (when $3 \in \mathcal{T}_0$). Then we have

$$1 = \mathbb{P}_2(\mathbb{R}) \leq \sum_{i=1}^{3} \mathbb{P}_2(\mathcal{V}_i) < \frac{3}{4},$$

which is a contradiction. Therefore, these compositions are forbidden.

- **The compositions with overlapping $\mathcal{T}_a$'s, i.e., $\mathcal{T}_1 \cap \mathcal{T}_3 = \{1\}$.** We claim that when Δ is sufficiently large, it is impossible for two groups of estimates, namely, $\mathcal{T}_1$ and $\mathcal{T}_a$ ($a = 2$ or 3), to share a common estimate $\boldsymbol{\beta}_1$. These include the conpositions 2-(b) and 3 ($a = 2$ and $q_0 = 1$) as well as cases 5 and 6 ($a = 3$ and $q_0 = 2$) with all of their subcases. Note that $1 \in \mathcal{T}_1 \cap \mathcal{T}_a$. Let $\boldsymbol{\theta}_s^*$ be the unique element of $\mathcal{S}_1$ and choose an arbitrary $\boldsymbol{\theta}_t^* \in \mathcal{S}_a$. Then we observe that $|\boldsymbol{\beta}_1 - \boldsymbol{\theta}_s^*| < \Delta/32$ by (51) and $|\boldsymbol{\beta}_1 - \boldsymbol{\theta}_t^*| < \Delta/192$ by (52). This yields $\Delta \leq |\boldsymbol{\theta}_s^* - \boldsymbol{\theta}_t^*| \leq |\boldsymbol{\beta}_1 - \boldsymbol{\theta}_s^*| + |\boldsymbol{\beta}_1 - \boldsymbol{\theta}_t^*| < 7\Delta/192$, which is a contradiction. Consequently, these compositions are excluded from our consideration.

In the end, we are left with only four cases as summarized in Table 2. In Step 3, we investigate each of these four cases and establish more refined approximation error bounds.

8.2.3 Step 3. Refined Approximation Error Analysis for the Four Remaining Cases

In this step, we analyze the four remaining cases (Case A through Case D listed in Table 2) one by one.

Table 2: All possible compositions (up to permutation symmetry) of the collections $\mathbb{T}, \mathbb{S}$ that are remaining after Step 2 of the proof of Theorem 3.

Case number	$(q, q_0, |\mathcal{T}_0|)$	$\mathcal{T}_0$	$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$
Case A / comp 1	$(1,1,2)$	$\{2,3\}$	$\{1\}$	-	-	$\{1,2,3\}$	-	-
Case B / comp 2-(a)-i	$(2,1,0)$	$\emptyset$	$\{1\}$	$\{2,3\}$	-	$\{1,2\}$	$\{3\}$	-
Case C / comp 4-(a)	$(2,2,1)$	$\{3\}$	$\{1\}$	$\{2\}$	-	$\{1,2\}$	$\{3\}$	-
Case D / comp 7	$(3,3,0)$	$\emptyset$	$\{1\}$	$\{2\}$	$\{3\}$	$\{1\}$	$\{2\}$	$\{3\}$

Step 3 - Case A   We draw a contradiction to rule argue that Case A is impossible to happen.

To this end, we begin by observing that $|\boldsymbol{\beta}_1 - \boldsymbol{\theta}_s^*| < |\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*|$ for all $i \in \{2,3\}$ and all $s \in [k_*]$. If we assume otherwise, there exists $(i,s) \in \{2,3\} \times [k_*]$ such that $|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*| \le |\boldsymbol{\beta}_1 - \boldsymbol{\theta}_s^*|$, which implies $\mathbb{P}_s[\mathcal{V}_1] \le 1/2$. Then it follows that $\max_{i \in \{2,3\}}\{\mathbb{P}_s[\mathcal{V}_i]\} \ge 1/4$, which contradicts (53). Note that $\boldsymbol{\theta}_2^* = \frac{1}{3}\sum_{s \in [3]} \boldsymbol{\theta}_s^* = \mathbf{0}$, and thus, $|\boldsymbol{\beta}_1| < \Delta/32$, cf. (51). Thus, $\boldsymbol{\beta}_2, \boldsymbol{\beta}_3 \notin [\boldsymbol{\theta}_1^* - 31\Delta/32, \boldsymbol{\theta}_3^* + 31\Delta/32]$. In what follows, we assume that $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ are on the same side of this interval, namely, $\boldsymbol{\theta}_3^* < \boldsymbol{\beta}_2 \le \boldsymbol{\beta}_3$. This is illustrated in Figure 4. However,
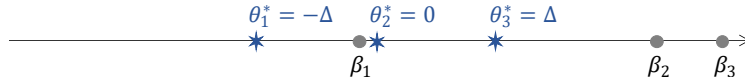
Figure 4: Illustration of Case A in Step 3.

the opposite-side case, i.e., the case where $\boldsymbol{\beta}_2 < \boldsymbol{\theta}_1^* < \boldsymbol{\theta}_3^* < \boldsymbol{\beta}_3$, can be analyzed in a similar manner with a minor modification to the argument that follows.

From the first-order stationarity condition, cf. (12), it follows that $\boldsymbol{\beta}_i = \frac{\mathbb{E}_*[\Psi_i \mathsf{x}]}{\mathbb{E}_*[\Psi_i]}$, $\forall i \in [k]$, and therefore,

$$0 = \sum_{s \in [k_*]} \mathbb{E}_s\left[\Psi_i \cdot (\boldsymbol{\beta}_i - \mathsf{x})\right], \qquad \forall i \in [k]. \tag{54}$$

Summing up the equations (54) for $i \in \{2,3\}$ yields

$$0 = \sum_{i \in \{2,3\}} \mathbb{E}_1\left[\Psi_i \cdot (\boldsymbol{\beta}_i - \mathsf{x})\right] + \sum_{s \in \{2,3\}} \left(\mathbb{E}_s\left[\Psi_2 \cdot (\boldsymbol{\beta}_2 - \mathsf{x})\right] + \mathbb{E}_s\left[\Psi_3 \cdot (\boldsymbol{\beta}_3 - \mathsf{x})\right]\right)$$
$$\ge \sum_{i \in \{2,3\}} \mathbb{E}_1\left[\Psi_i \cdot (\boldsymbol{\beta}_i - \mathsf{x})\right] + \sum_{s \in \{2,3\}} \mathbb{E}_s\left[\left(\Psi_2 + \Psi_3\right) \cdot (\boldsymbol{\beta}_2 - \mathsf{x})\right], \tag{55}$$

where the last inequality follows from $\boldsymbol{\beta}_2 \le \boldsymbol{\beta}_3$. Rewriting the first term in (55) using the Stein's identity (Lemma 3), cf. (29), we obtain that

$$\sum_{i \in \{2,3\}} \mathbb{E}_1\left[\Psi_i \cdot (\boldsymbol{\beta}_i - \mathsf{x})\right] = \sum_{i \in \{2,3\}} \left(-\boldsymbol{\theta}_1^* \cdot \mathbb{E}_1\left[\Psi_i\right] + \sum_{j \in [k]} \boldsymbol{\beta}_j \cdot \mathbb{E}_1\left[\Psi_i \Psi_j\right]\right)$$
$$\stackrel{(a)}{=} \sum_{i \in \{2,3\}} \sum_{j \in [k]} (\boldsymbol{\beta}_j - \boldsymbol{\theta}_1^*) \cdot \mathbb{E}_1\left[\Psi_i \Psi_j\right]$$
$$\stackrel{(b)}{\ge} 0,$$

because (a) $\sum_{j \in [k]} \Psi_j = 1$ with probability 1 and (b) $\boldsymbol{\beta}_j > \boldsymbol{\theta}_1^*$, $\forall j \in [k]$. Therefore, it follows from (55) that

$$0 \ge \sum_{s \in \{2,3\}} \mathbb{E}_s\left[\left(\Psi_2 + \Psi_3\right) \cdot (\boldsymbol{\beta}_2 - \mathsf{x})\right]$$
$$= \sum_{s \in \{2,3\}} \mathbb{E}_s\left[\left(1 - \Psi_1\right) \cdot (\boldsymbol{\beta}_2 - \mathsf{x})\right]. \tag{56}$$

In the rest, we will argue that $\mathbb{E}_s\left[\left(1 - \Psi_1\right)\cdot(\boldsymbol{\beta}_2 - \mathsf{x})\right] > 0$ for $s \in \{2,3\}$ to draw a contradiction.

We define a function $\psi^{\mathrm{prox}} : \mathbb{R} \to \mathbb{R}$ so that

$$\psi^{\mathrm{prox}}(x) = \frac{e^{-\|x - \boldsymbol{\beta}_2\|^2/2}}{e^{-\|x - \boldsymbol{\beta}_1\|^2/2} + e^{-\|x - \boldsymbol{\beta}_2\|^2/2}},$$

and define a random variable $\Psi^{\mathrm{prox}} = \psi^{\mathrm{prox}}(\mathsf{x})$. Note that $\psi^{\mathrm{prox}}(x)$ is equal to the association coefficient of $\boldsymbol{\beta}_2$ at x if there were only $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ (with $\boldsymbol{\beta}_3$ removed). We use Ψ^{prox} as a proxy of $1 - \Psi_1$ to facilitate our subsequent analysis of Case A. Specifically, we observe that for each $s \in \{2,3\}$,

$$\mathbb{E}_s\left[\left(1 - \Psi_1\right)\cdot(\boldsymbol{\beta}_2 - \mathsf{x})\right] = \mathbb{E}_s\left[\Psi^{\mathrm{prox}}\cdot(\boldsymbol{\beta}_2 - \mathsf{x})\right] + \mathbb{E}_s\left[\left(1 - \Psi_1 - \Psi^{\mathrm{prox}}\right)\cdot(\boldsymbol{\beta}_2 - \mathsf{x})\right]. \tag{57}$$

Then we can easily verify that

$$\mathbb{E}_s\left[\Psi^{\mathrm{prox}}\cdot(\boldsymbol{\beta}_2 - \mathsf{x})\right] \geq \frac{\boldsymbol{\beta}_2 - \boldsymbol{\theta}_s^*}{8}\cdot Q\left(\frac{\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2}{2} - \boldsymbol{\theta}_s^* - \frac{1}{\boldsymbol{\beta}_1 - \boldsymbol{\theta}_s^*}\right) \tag{58}$$

by Lemma 5. Then we also observe that

$$\begin{aligned}
\mathbb{E}_s\left[\left(1 - \Psi_1 - \Psi^{\mathrm{prox}}\right)\cdot(\boldsymbol{\beta}_2 - \mathsf{x})\right] &\overset{(a)}{\geq} \mathbb{E}_s\left[\left(1 - \Psi_1 - \Psi^{\mathrm{prox}}\right)\cdot(\boldsymbol{\beta}_2 - \mathsf{x})\cdot\mathbb{1}\{\mathsf{x} \geq \boldsymbol{\beta}_2\}\right] \\
&\overset{(b)}{\geq} -\mathbb{E}_s\left[(\mathsf{x} - \boldsymbol{\beta}_2)\cdot\mathbb{1}\{\mathsf{x} \geq \boldsymbol{\beta}_2\}\right] \\
&\overset{(c)}{\geq} -\mathbb{E}_s\left[(\mathsf{x} - \boldsymbol{\theta}_s^*)\cdot\mathbb{1}\{\mathsf{x} \geq \boldsymbol{\beta}_2\}\right] \\
&= -\int_{\boldsymbol{\beta}_2 - \boldsymbol{\theta}_s^*}^{\infty} z\cdot\phi(z)\,\mathrm{d}z \\
&\overset{(d)}{=} -\phi(\boldsymbol{\beta}_2 - \boldsymbol{\theta}_s^*)
\end{aligned} \tag{59}$$

because (a) $1 - \Psi_1 - \Psi^{\mathrm{prox}} \geq 0$; (b) $\left|1 - \Psi_1 - \Psi^{\mathrm{prox}}\right| \leq 1$; (c) $\boldsymbol{\beta}_2 > \boldsymbol{\theta}_s^*$; and (d) follows from (44).

Letting $\tau_s := \frac{\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2}{2} - \boldsymbol{\theta}_s^* - \frac{1}{\boldsymbol{\beta}_1 - \boldsymbol{\theta}_s^*}$ for $s \in \{2,3\}$, we observe that $\tau_s \geq 0$, and thus,

$$\begin{aligned}
Q(\tau_s) &\overset{(a)}{\geq} \frac{1}{\tau_s + 1}\phi(\tau_s) \overset{(b)}{\geq} e^{-\tau_s}\cdot\phi(\tau_s) \\
&\geq \phi(\tau_s + 1),
\end{aligned} \tag{60}$$

where (a) follows from Lemma 4, (b) is due to $e^t \geq t + 1$, $\forall t \geq 0$, and (c) is from the observation $t^2/2 + t \leq (t + 1)^2$, $\forall t \in \mathbb{R}$. Collecting (57), (58), (59), and (60) together with (56), we have

$$\begin{aligned}
0 &\geq \sum_{s \in \{2,3\}} \mathbb{E}_s\left[\left(1 - \Psi_1\right)\cdot(\boldsymbol{\beta}_2 - \mathsf{x})\right] \\
&\geq \sum_{s \in \{2,3\}} \left(\frac{\boldsymbol{\beta}_2 - \boldsymbol{\theta}_s^*}{8}\cdot\phi(\tau_s + 1) - \phi(\boldsymbol{\beta}_2 - \boldsymbol{\theta}_s^*)\right) \\
&\overset{(a)}{\geq} \frac{\boldsymbol{\beta}_2 - \boldsymbol{\theta}_3^*}{8}\cdot\phi(\tau_3 + 1) - \sum_{s \in \{2,3\}} \phi(\boldsymbol{\beta}_2 - \boldsymbol{\theta}_s^*) \\
&\overset{(b)}{>} 0
\end{aligned} \tag{61}$$

where (a) is due to $(\boldsymbol{\beta}_2 - \boldsymbol{\theta}_s^*)\cdot\phi(\tau_s + 1) > 0$. The last inequality (b) follows from the observations that $\phi(t)$ is strictly monotone decreasing for $t \geq 0$ and that

$$\begin{aligned}
\tau_3 + 1 &= \frac{\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2}{2} - \boldsymbol{\theta}_3^* - \frac{1}{\boldsymbol{\beta}_1 - \boldsymbol{\theta}_3^*} + 1 \\
&\leq (\boldsymbol{\beta}_2 - \boldsymbol{\theta}_3^*) - \frac{31\Delta}{32} + \frac{32}{31\Delta} + 1 \\
&< \boldsymbol{\beta}_2 - \boldsymbol{\theta}_3^*, \\
\frac{\boldsymbol{\beta}_2 - \boldsymbol{\theta}_3^*}{8} &> \frac{31\Delta}{256} > 2,
\end{aligned}$$

because $\boldsymbol{\beta}_2 - \boldsymbol{\theta}_3^* > \boldsymbol{\theta}_3^* - \boldsymbol{\beta}_1 > \frac{31\Delta}{32}$ and Δ is assumed to be sufficiently large (to be precise, $\Delta > 2^{16} \cdot 3^{10} \cdot (\sqrt{2\pi} + 1)$).

The inequality $0 > 0$ in (61) is a contradiction, and we conclude that case A cannot happen.

Step 3 - Case B Recall from (51) and (52) in Step 0 of this proof that

$$\left\| \boldsymbol{\beta}_1 - \frac{\boldsymbol{\theta}_1^* + \boldsymbol{\theta}_2^*}{2} \right\| < \frac{\Delta}{32},$$

$$\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_3^*\| < \frac{\Delta}{192}, \qquad \forall i \in \{2,3\}.$$

This implies that (1) $\|\boldsymbol{\beta}_1 - \boldsymbol{\theta}_s^*\| < \frac{17\Delta}{32}$, $\forall s \in \{1,2\}$; (2) $\|\boldsymbol{\beta}_1 - \boldsymbol{\theta}_3^*\| \geq \left\| \frac{\boldsymbol{\theta}_1^* + \boldsymbol{\theta}_2^*}{2} - \boldsymbol{\theta}_3^* \right\| - \left\| \boldsymbol{\beta}_1 - \frac{\boldsymbol{\theta}_1^* + \boldsymbol{\theta}_2^*}{2} \right\| > \frac{3\Delta}{2} - \frac{\Delta}{32} = \frac{47\Delta}{32}$; and (3) $\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\| > \frac{191\Delta}{192}$ for all $(i,s) \in \{2,3\} \times \{1,2\}$.

Recall $\mathcal{S}_1 = \{1,2\}$ in this case and let $\delta_1 = 89\Delta/192$. Observe that $\delta_1 \geq \max \left\{ \frac{4}{\sqrt{2\pi}}, 8\sqrt{\log(2\sqrt{2} \cdot kk_*)} \right\}$ and that the conditions in (50) are satisfied for $i = 1$ with $\delta = \delta_1$ and $\mathcal{S} = \mathcal{S}_1$ because

$$
\begin{aligned}
\min_{j \in \{2,3\}} \|\boldsymbol{\beta}_j - \boldsymbol{\theta}_s^*\| - \|\boldsymbol{\beta}_1 - \boldsymbol{\theta}_s^*\| &\geq \frac{191}{192}\Delta - \frac{17}{32}\Delta \\
&= \frac{89}{192}\Delta, \quad \forall s \in \{1,2\}, \\
\|\boldsymbol{\beta}_1 - \boldsymbol{\theta}_3^*\| - \min_{j \in \{2,3\}} \|\boldsymbol{\beta}_j - \boldsymbol{\theta}_3^*\| &\geq \frac{47}{32}\Delta - \frac{1}{192}\Delta \\
&= \frac{281}{192}\Delta.
\end{aligned}
\tag{62}
$$

Applying Lemma 8, we obtain

$$
\begin{aligned}
\left\| \boldsymbol{\beta}_1 - \frac{\boldsymbol{\theta}_1^* + \boldsymbol{\theta}_2^*}{2} \right\| &\leq 4kk_* \Delta_{\max} \cdot e^{-\frac{\delta_1^2}{64}} \\
&= 8kk_* \Delta \cdot e^{-\frac{7921}{2359296}\Delta^2} \\
&\leq 8kk_* \Delta \cdot e^{-\frac{1}{298}\Delta^2}.
\end{aligned}
\tag{63}
$$

It remains to establish an upper bound for $\max_{i \in \{2,3\}} \|\boldsymbol{\beta}_i - \boldsymbol{\theta}_3^*\|$. To accomplish this, for each $i \in \{2,3\}$, we define a function $\psi_i^{\mathrm{prox}} : \mathbb{R} \to \mathbb{R}$ so that

$$\psi_i^{\mathrm{prox}}(x) = \frac{e^{-\|x - \boldsymbol{\beta}_i\|^2/2}}{e^{-\|x - \boldsymbol{\beta}_2\|^2/2} + e^{-\|x - \boldsymbol{\beta}_3\|^2/2}},$$

and define a random variable $\Psi_i^{\mathrm{prox}} = \psi_i^{\mathrm{prox}}(\mathsf{x})$. Note that $\psi_i^{\mathrm{prox}}(x)$ is equal to the association coefficient of $\boldsymbol{\beta}_i$ at x if there were only $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ (with $\boldsymbol{\beta}_1$ removed). We use Ψ_i^{prox} as a proxy of Ψ_i to facilitate our subsequent analysis of Case B.

Next, we make three preparatory observations as follows.

(a) First, it follows from the first-order stationarity condition, cf. (12), that

$$\sum_{s \in [k_*]} \mathbb{E}_s \left[\Psi_i \cdot (\mathsf{x} - \boldsymbol{\beta}_i) \right] = 0, \qquad \forall i \in [k].$$

This leads to the following: for all $i \in [3]$,

$$\mathbb{E}_3 \left[\Psi_i \cdot (\mathsf{x} - \boldsymbol{\beta}_i) \right] = - \sum_{s \in \{1,2\}} \mathbb{E}_s \left[\Psi_i \cdot (\mathsf{x} - \boldsymbol{\beta}_i) \right]. \tag{64}$$

38

(b) Second, by the Stein's identity (Lemma 3), we have the following equation, cf. (29): for all $i \in \{2,3\}$,

$$\mathbb{E}_3\left[\Psi_i^{\text{prox}} \cdot (\mathsf{x} - \boldsymbol{\beta}_i)\right] = \boldsymbol{\theta}_s^* \cdot \mathbb{E}_3\left[\Psi_i^{\text{prox}}\right] - \sum_{j \in \{2,3\}} \boldsymbol{\beta}_j \cdot \mathbb{E}_3\left[\Psi_i^{\text{prox}}\Psi_j^{\text{prox}}\right].$$

Then it follows that

$$\mathbb{E}_3\left[\Psi_3^{\text{prox}}\right] \cdot \mathbb{E}_3\left[\Psi_2^{\text{prox}} \cdot (\mathsf{x} - \boldsymbol{\beta}_2)\right] - \mathbb{E}_3\left[\Psi_2^{\text{prox}}\right] \cdot \mathbb{E}_3\left[\Psi_3^{\text{prox}} \cdot (\mathsf{x} - \boldsymbol{\beta}_3)\right]$$
$$= - \sum_{j \in \{2,3\}} \left\{ \boldsymbol{\beta}_j \cdot \mathbb{E}_3\left[\Psi_3^{\text{prox}}\right] \cdot \mathbb{E}_3\left[\Psi_2^{\text{prox}}\Psi_j^{\text{prox}}\right] - \boldsymbol{\beta}_j \cdot \mathbb{E}_3\left[\Psi_2^{\text{prox}}\right] \cdot \mathbb{E}_3\left[\Psi_3^{\text{prox}}\Psi_j^{\text{prox}}\right] \right\}$$
$$= -(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_3) \cdot \text{Var}_3\left[\Psi_2^{\text{prox}}\right] \tag{65}$$

where $\text{Var}_3[\mathsf{z}] := \mathbb{E}_3[\mathsf{z}^2] - \mathbb{E}_3[\mathsf{z}]^2$ for a random variable z.

(c) Third, for each $i \in \{2,3\}$, the proxy $\psi_i^{\text{prox}}(x) \approx \psi_i(x)$ in the sense that

$$\mathbb{E}_3\left[(\Psi_i^{\text{prox}} - \Psi_i)^2\right] = \mathbb{E}_3\left[(\Psi_i^{\text{prox}} \cdot \Psi_1)^2\right]$$
$$\leq \mathbb{E}_3\left[\Psi_1\right].$$

As we observed $\|\boldsymbol{\beta}_1 - \boldsymbol{\theta}_3^*\| - \min_{i \in [k]} \|\boldsymbol{\beta}_i - \boldsymbol{\theta}_3^*\| > \frac{281}{192}\Delta$ in (62), we apply Lemma 7 to obtain

$$\mathbb{E}_3\left[(\Psi_i^{\text{prox}} - \Psi_i)^2\right] \leq \mathbb{E}_3\left[\Psi_1\right]$$
$$\leq \left(1 + \frac{768}{281\sqrt{2\pi} \cdot \Delta}\right) e^{-\frac{78961}{1179648}\Delta^2} \tag{66}$$
$$\leq 2e^{-\frac{1}{15}\Delta^2}.$$

Combining these observations together, we get

$$\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_3\| \cdot \text{Var}_3\left[\Psi_2^{\text{prox}}\right] = \left\|\mathbb{E}_3\left[\Psi_2^{\text{prox}}\right] \cdot \mathbb{E}_3\left[\Psi_3^{\text{prox}} \cdot (\mathsf{x} - \boldsymbol{\beta}_3)\right] - \mathbb{E}_3\left[\Psi_3^{\text{prox}}\right] \cdot \mathbb{E}_3\left[\Psi_2^{\text{prox}} \cdot (\mathsf{x} - \boldsymbol{\beta}_2)\right]\right\| \quad \because (65)$$
$$\leq \sum_{i \in \{2,3\}} \left\|\mathbb{E}_3\left[\Psi_i^{\text{prox}} \cdot (\mathsf{x} - \boldsymbol{\beta}_i)\right]\right\|$$
$$\leq \sum_{i \in \{2,3\}} \left\|\mathbb{E}_3\left[\Psi_i \cdot (\mathsf{x} - \boldsymbol{\beta}_i)\right]\right\| + \sum_{i \in \{2,3\}} \left\|\mathbb{E}_3\left[(\Psi_i^{\text{prox}} - \Psi_i) \cdot (\mathsf{x} - \boldsymbol{\beta}_i)\right]\right\|. \tag{67}$$

Then we obtain separate upper bounds for the two terms in (67) as the following.

- *First term in* (67). For each $i \in \{2,3\}$,

$$\left\|\mathbb{E}_3\left[\Psi_i \cdot (\mathsf{x} - \boldsymbol{\beta}_i)\right]\right\| \overset{(a)}{\leq} \sum_{s \in \{1,2\}} \left\|\mathbb{E}_s\left[\Psi_i \cdot (\mathsf{x} - \boldsymbol{\beta}_i)\right]\right\|$$
$$\overset{(b)}{\leq} \sum_{s \in \{1,2\}} \left(\mathbb{E}_s\left[\Psi_i^2\right] \cdot \mathbb{E}_s\left[\|\mathsf{x} - \boldsymbol{\beta}_i\|^2\right]\right)^{1/2}$$
$$\overset{(c)}{\leq} 12\Delta \cdot e^{-\frac{1}{298}\Delta^2}, \tag{68}$$

because (a) is due to (64); (b) follows from Cauchy-Schwarz inequality; and (c) for any $(i,s) \in \{2,3\} \times$

$\{1, 2\}$, it follows from (62) and Lemma 7 that

$$
\mathbb{E}_s\left[\Psi_i^2\right] \leq \mathbb{E}_s\left[\Psi_i\right]
$$

$$
\leq \left(1 + \frac{768}{89\sqrt{2\pi} \cdot \Delta}\right) e^{-\frac{7921}{1179648}\Delta^2}
$$

$$
\leq 2e^{-\frac{1}{149}\Delta^2},
$$

$$
\mathbb{E}_s\left\|\mathsf{x} - \boldsymbol{\beta}_i\right\|^2 \leq 2\left(\mathbb{E}_s\|\mathsf{x} - \boldsymbol{\theta}_s^*\|^2 + \mathbb{E}_s\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\|^2\right)
$$

$$
\leq 2\left(1 + \max_{i \in \{2,3\}}\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_1^*\|^2\right)
$$

$$
< 2\left(1 + \left(\frac{385}{192}\Delta\right)^2\right)
$$

$$
\leq 2 \cdot (3\Delta)^2.
$$

- *Second term in* (67). For each $i \in \{2, 3\}$,

$$
\mathbb{E}_3\left\|\mathsf{x} - \boldsymbol{\beta}_i\right\|^2 \leq 2\left(\mathbb{E}_3\|\mathsf{x} - \boldsymbol{\theta}_3^*\|^2 + \mathbb{E}_3\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_3^*\|^2\right)
$$

$$
\leq 2\left(1 + \max_{i \in \{2,3\}}\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_3^*\|^2\right)
$$

$$
< 2\left(1 + \left(\frac{1}{192}\Delta\right)^2\right)
$$

$$
\leq 2 \cdot \left(\frac{1}{96}\Delta\right)^2. \tag{69}
$$

and therefore,

$$
\left\|\mathbb{E}_3\left[\left(\Psi_i^{\mathrm{prox}} - \Psi_i\right) \cdot (\mathsf{x} - \boldsymbol{\beta}_i)\right]\right\| \leq \left(\mathbb{E}_3\left[\left(\Psi_i^{\mathrm{prox}} - \Psi_i\right)^2\right] \cdot \mathbb{E}_3\left[\|\mathsf{x} - \boldsymbol{\beta}_i)\|^2\right]\right)^{1/2}
$$

$$
\leq \frac{1}{48}\Delta \cdot e^{-\frac{1}{30}\Delta^2}. \qquad \because (66) \ \& \ (69) \tag{70}
$$

Combining (68) and (70) with (67), we have

$$
\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_3\| \cdot \mathrm{Var}_3\left[\Psi_2^{\mathrm{prox}}\right] \leq 24\Delta \cdot e^{-\frac{1}{298}\Delta^2} + \frac{1}{24}\Delta \cdot e^{-\frac{1}{30}\Delta^2}
$$

$$
\leq 25\Delta \cdot e^{-\frac{1}{298}\Delta^2}. \tag{71}
$$

Lastly, we apply Lemma 6 to obtain a variance lower bound

$$
\mathrm{Var}_3\left[\Psi_2^{\mathrm{prox}}\right] \geq \frac{1}{32\sqrt{2\pi}}\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_3\|^5 \cdot e^{-4\left(\frac{\Delta}{96}\right)^2} \tag{72}
$$

because $\|\frac{\boldsymbol{\beta}_2 + \boldsymbol{\beta}_3}{2} - \boldsymbol{\theta}_3^*\| + \|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_3\| \leq 2\max\{\|\boldsymbol{\beta}_2 - \boldsymbol{\theta}_3^*\|, \|\boldsymbol{\beta}_3 - \boldsymbol{\theta}_3^*\|\} < \frac{1}{96}\Delta$. Combining (71) and (72) yields

$$
\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_3\|^6 \leq 800\sqrt{2\pi}\Delta \cdot e^{\left(\frac{1}{2304} - \frac{1}{298}\right)\Delta^2}
$$

$$
\leq 800\sqrt{2\pi}\Delta \cdot e^{-\frac{1}{343}\Delta^2}.
$$

Consequently,

$$
\max_{i \in \{2,3\}}\|\boldsymbol{\beta}_i \boldsymbol{\theta}_3^*\| \leq \|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_3\|
$$

$$
\leq \left(800\sqrt{2\pi}\Delta \cdot e^{-\frac{1}{343}\Delta^2}\right)^{1/6}. \tag{73}
$$

We conclude this case with the two upper bounds (63) and (73).

Step 3 - Case C In this case, $q = q_0 = 2$, i.e., there are two 'one-fits-many' clusters. We observe that

$$\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^* \in \mathcal{V}_1 \quad \text{and} \quad \boldsymbol{\theta}_3^* \in \mathcal{V}_2, \tag{74}$$

which readily follows from the construction of these clusters; see Section 7.1.2, cf. Algorithm 1 and the definition of the set $\mathcal{A}_i^\delta$ in (35). Moreover, it follows from (51) that

$$\left\| \boldsymbol{\beta}_1 - \frac{\boldsymbol{\theta}_1^* + \boldsymbol{\theta}_2^*}{2} \right\| < \frac{\Delta}{32} \quad \text{and} \quad \|\boldsymbol{\beta}_2 - \boldsymbol{\theta}_3^*\| < \frac{\Delta}{32}. \tag{75}$$

We consider three possible subcases based on the location of $\boldsymbol{\beta}_3$, namely,

(i) $\boldsymbol{\beta}_3 < \boldsymbol{\beta}_1$;

(ii) $\boldsymbol{\beta}_3 > \boldsymbol{\beta}_2$;

(iii) $\boldsymbol{\beta}_3 \in (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$.

In the following, we discuss each of these three subcases individually.

Subcase (i): $\boldsymbol{\beta}_3 < \boldsymbol{\beta}_1$. Letting $\boldsymbol{c} = (\boldsymbol{\beta}_3 + \boldsymbol{\beta}_1)/2$, we observe that $\mathcal{V}_3 = (-\infty, \boldsymbol{c}]$ and that $\boldsymbol{c} \leq \boldsymbol{\theta}_1^*$ by observation 74. Thus, $\mathbb{P}_1(\mathcal{V}_3) = Q(\boldsymbol{\theta}_1^* - \boldsymbol{c})$. Because

$$\mathbb{P}_1(\mathcal{V}_3) = Q(\boldsymbol{\theta}_1^* - \boldsymbol{c}) \geq \phi(\boldsymbol{\theta}_1^* - \boldsymbol{c} + 1) \quad \text{by (60)} \quad \text{and}$$

$$\mathbb{P}_1(\mathcal{V}_3) < \frac{1}{2^9 \cdot 3^3} \quad \text{by (53)},$$

we have $\boldsymbol{\theta}_1^* - \boldsymbol{c} > \tau_0$ where

$$\tau_0 := \left(2 \log(2^9 \cdot 3^3) \right)^{1/2} - 1 \approx 3.367.$$

Next, we recall from the first-order stationary condition, cf. (12), that

$$\boldsymbol{\beta}_3 = \frac{\sum_{s=1}^3 \mathbb{E}_s[\Psi_3 \cdot \mathsf{x}]}{\sum_{s=1}^3 \mathbb{E}_s[\Psi_3]}. \tag{76}$$

We derive a lower bound for $\boldsymbol{\beta}_3$ to argue that $\boldsymbol{\beta}_3 \approx \boldsymbol{c}$. To this end, we begin by observing that for all $x \leq \frac{\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2}{2}$,

$$\psi_3(x) \geq \frac{\phi(x - \boldsymbol{\beta}_3)}{2\phi(x - \boldsymbol{\beta}_1) + \phi(x - \boldsymbol{\beta}_3)}$$

$$= \frac{1}{1 + 2e^{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3) \cdot (x - \boldsymbol{c})}}.$$

In particular, this implies that

$$\psi_3(x) \geq 1 - 2e^{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3) \cdot (x - \boldsymbol{c})}, \qquad \forall x \leq \boldsymbol{c}. \tag{77}$$

Then we consider the numerator of (76). Due to the translation invariance, we may assume $\boldsymbol{c} = 0$ without loss of generality, and therefore, it follows that

$$\sum_{s=1}^3 \mathbb{E}_s\left[\Psi_3 \cdot \mathsf{x}\right] = \sum_{s=1}^3 \mathbb{E}_s\left[\Psi_3 \cdot \mathsf{x} \cdot \mathbb{1}\{\mathsf{x} \leq \boldsymbol{c}\}\right] + \sum_{s=1}^3 \mathbb{E}_s\left[\Psi_3 \cdot \mathsf{x} \cdot \mathbb{1}\{\mathsf{x} > \boldsymbol{c}\}\right]$$

$$\overset{(a)}{\geq} \sum_{s=1}^3 \mathbb{E}_s\left[\Psi_3 \cdot \mathsf{x} \cdot \mathbb{1}\{\mathsf{x} \leq \boldsymbol{c}\}\right]$$

$$\overset{(b)}{\geq} \sum_{s=1}^3 \mathbb{E}_s\left[\mathsf{x} \cdot \mathbb{1}\{\mathsf{x} \leq \boldsymbol{c}\}\right]$$

$$\overset{(c)}{=} -\sum_{s=1}^3 \phi\left(\boldsymbol{\theta}_s^* - \boldsymbol{c}\right) + \sum_{s=1}^3 \left(\boldsymbol{\theta}_s^* - \boldsymbol{c}\right) \cdot Q\left(\boldsymbol{\theta}_s^* - \boldsymbol{c}\right), \tag{78}$$

41

because (a) $\psi_3(x) \cdot x \cdot \mathbb{1}\{x > c\} \geq 0$; (b) $\psi_3(x) \in [0,1]$ & $x \cdot \mathbb{1}\{x \leq c\} \leq 0$; and (c) follows from the Gaussian integral formula (44).

For each $s \in [3]$,

$$-\phi(\boldsymbol{\theta}_s^* - \boldsymbol{c}) + (\boldsymbol{\theta}_s^* - \boldsymbol{c}) \cdot Q(\boldsymbol{\theta}_s^* - \boldsymbol{c}) \geq \left(-1 + \frac{\boldsymbol{\theta}_s^* - \boldsymbol{c}}{\boldsymbol{\theta}_s^* - \boldsymbol{c} + 1}\right) \phi(\boldsymbol{\theta}_s^* - \boldsymbol{c})$$

$$= -\frac{1}{\boldsymbol{\theta}_s^* - \boldsymbol{c} + 1} \phi(\boldsymbol{\theta}_s^* - \boldsymbol{c})$$

due to the upper and lower bounds for the Gaussian Q-function, cf. (43) and Lemma 4. Plugging this into (78), we get

$$\sum_{s=1}^{3} \mathbb{E}_s[\Psi_3 \cdot \mathsf{x}] \geq -\sum_{s=1}^{3} \frac{1}{\boldsymbol{\theta}_s^* - \boldsymbol{c} + 1} \phi(\boldsymbol{\theta}_s^* - \boldsymbol{c})$$

$$\geq -\frac{3}{\boldsymbol{\theta}_1^* - \boldsymbol{c} + 1} \phi(\boldsymbol{\theta}_1^* - \boldsymbol{c}). \tag{79}$$

Thereafter, we consider the denominator of (76) and observe that

$$\sum_{s=1}^{3} \mathbb{E}_s[\Psi_3] \overset{(a)}{\geq} \mathbb{E}_1[\Psi_3 \cdot \mathbb{1}\{\mathsf{x} \leq \boldsymbol{c}\}]$$

$$\overset{(b)}{\geq} \mathbb{E}_1[\mathbb{1}\{\mathsf{x} \leq \boldsymbol{c}\}] - \mathbb{E}_1\left[2e^{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3) \cdot (\mathsf{x} - \boldsymbol{c})} \cdot \mathbb{1}\{\mathsf{x} \leq \boldsymbol{c}\}\right]$$

because (a) $\psi_3(x) \geq 0$ and (b) $\psi_3(x) \geq 1 - 2e^{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3) \cdot (x - \boldsymbol{c})}$, $\forall x \leq \boldsymbol{c}$ by (77). Observe that

$$\mathbb{E}_1[\mathbb{1}\{\mathsf{x} \leq \boldsymbol{c}\}] = Q(\boldsymbol{\theta}_1^* - \boldsymbol{c})$$

and that

$$\mathbb{E}_1\left[e^{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3) \cdot (\mathsf{x} - \boldsymbol{c})} \cdot \mathbb{1}\{\mathsf{x} \leq \boldsymbol{c}\}\right] = \mathbb{E}_1\left[e^{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3) \cdot (\mathsf{z} - (\boldsymbol{c} - \boldsymbol{\theta}_1^*))} \cdot \mathbb{1}\{\mathsf{z} \leq \boldsymbol{c} - \boldsymbol{\theta}_1^*\}\right]$$

$$= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\boldsymbol{c} - \boldsymbol{\theta}_1^*} e^{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3) \cdot (\mathsf{z} - (\boldsymbol{c} - \boldsymbol{\theta}_1^*))} \cdot e^{-\frac{z^2}{2}} \, dz$$

$$= \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3}{2} \cdot (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3 - 2(\boldsymbol{c} - \boldsymbol{\theta}_1^*))} \int_{-\infty}^{\boldsymbol{c} - \boldsymbol{\theta}_1^*} e^{-\frac{1}{2}(z - (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3))^2} \, dz \qquad \because (46)$$

$$= e^{-\frac{1}{2}(\boldsymbol{\theta}_1^* - \boldsymbol{c})^2} \cdot e^{\frac{1}{2}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3) + (\boldsymbol{\theta}_1^* - \boldsymbol{c}))^2} \cdot Q\big((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3) + (\boldsymbol{\theta}_1^* - \boldsymbol{c})\big).$$

Thus, we obtain

$$\sum_{s=1}^{3} \mathbb{E}_s[\Psi_3] = Q(\boldsymbol{\theta}_1^* - \boldsymbol{c}) - 2e^{-\frac{1}{2}(\boldsymbol{\theta}_1^* - \boldsymbol{c})^2} \cdot e^{\frac{1}{2}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3) + (\boldsymbol{\theta}_1^* - \boldsymbol{c}))^2} \cdot Q\big((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3) + (\boldsymbol{\theta}_1^* - \boldsymbol{c})\big) \tag{80}$$

Further proceeding with the lower bound in (80), we use the upper and lower bounds for the Gaussian Q-function, cf. (43) and Lemma 4, to obtain

$$\sum_{s=1}^{3} \mathbb{E}_s[\Psi_3] \geq \frac{1}{\boldsymbol{\theta}_1^* - \boldsymbol{c} + 1} \phi(\boldsymbol{\theta}_1^* - \boldsymbol{c}) - 2 \cdot \phi\big((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3) + (\boldsymbol{\theta}_1^* - \boldsymbol{c})\big) \cdot \frac{e^{-\frac{1}{2}(\boldsymbol{\theta}_1^* - \boldsymbol{c})^2} \cdot e^{\frac{1}{2}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3) + (\boldsymbol{\theta}_1^* - \boldsymbol{c}))^2}}{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3) + (\boldsymbol{\theta}_1^* - \boldsymbol{c})}$$

$$\geq \frac{1}{3} \cdot \frac{1}{\boldsymbol{\theta}_1^* - \boldsymbol{c} + 1} \cdot \phi(\boldsymbol{\theta}_1^* - \boldsymbol{c}) \tag{81}$$

with the last inequality following from $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_3 = 2(\boldsymbol{\beta}_1 - \boldsymbol{c}) > 2(\boldsymbol{\theta}_1^* + \frac{15}{32}\Delta - \boldsymbol{c}) \geq 2(\boldsymbol{\theta}_1^* - \boldsymbol{c} + 1)$.

In the end, we combine the lower bounds (79) and (81) with the stationary condition for $\boldsymbol{\beta}_3$ in (76) to obtain

$$\boldsymbol{\beta}_3 - \boldsymbol{c} \geq \frac{-\frac{3}{\boldsymbol{\theta}_1^* - \boldsymbol{c} + 1} \phi(\boldsymbol{\theta}_1^* - \boldsymbol{c})}{\frac{1}{3} \cdot \frac{1}{\boldsymbol{\theta}_1^* - \boldsymbol{c} + 1} \cdot \phi(\boldsymbol{\theta}_1^* - \boldsymbol{c})} = -9.$$

This implies that

$$9 \geq \|\boldsymbol{\beta}_3 - \boldsymbol{c}\| = \|\boldsymbol{\beta}_1 - \boldsymbol{c}\| \geq \|\boldsymbol{\beta}_1 - \boldsymbol{\theta}_1^*\| > \frac{15}{32}\Delta,$$

which is a contradiction. Consequently, we conclude that Subcase (i) of Case C cannot happen.

Subcase (ii): $\boldsymbol{\beta}_3 > \boldsymbol{\beta}_2$. We recall the equivalent stationary condition (Theorem 1), cf. (16):

$$\sum_{j \in [k]} \boldsymbol{\beta}_j \sum_{s \in [k_*]} \mathbb{E}_s\left[\Psi_i \Psi_j\right] = \sum_{s \in [k_*]} \boldsymbol{\theta}_s^* \mathbb{E}_s\left[\Psi_i\right], \qquad \forall i \in [k].$$

Letting $i = 3$ and rearranging the terms, we obtain

$$\sum_{j \in [k]} \boldsymbol{\beta}_j \sum_{s \in \{1,2\}} \mathbb{E}_s\left[\Psi_3 \Psi_j\right] - \sum_{s \in \{1,2\}} \boldsymbol{\theta}_s^* \mathbb{E}_s\left[\Psi_3\right] + \sum_{j \in [k]} (\boldsymbol{\beta}_j - \boldsymbol{\theta}_3^*) \cdot \mathbb{E}_3\left[\Psi_3 \Psi_j\right] = 0,$$

whence it follows that

$$(\boldsymbol{\beta}_3 - \boldsymbol{\theta}_3^*) \cdot \mathbb{E}_3\left[\Psi_3 \Psi_3\right] - (\boldsymbol{\beta}_2 - \boldsymbol{\theta}_3^*) \cdot \mathbb{E}_3\left[\Psi_3 \Psi_2\right] = -(\boldsymbol{\beta}_1 - \boldsymbol{\theta}_3^*) \cdot \mathbb{E}_3\left[\Psi_3 \Psi_1\right] + \sum_{s \in \{1,2\}} \sum_{j \in [k]} (\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_j) \cdot \mathbb{E}_s\left[\Psi_3 \Psi_j\right]$$

(82)

Then we derive an upper bound for the absolute value of the expression on the right-hand side of (82).

- Note that $\|\boldsymbol{\beta}_1 - \boldsymbol{\theta}_3^*\| \leq \frac{49}{32}\Delta \leq 2\Delta$, cf. (51). Moreover, $\|\boldsymbol{\beta}_1 - \boldsymbol{\theta}_3^*\| - \|\boldsymbol{\beta}_1 - \boldsymbol{\theta}_3^*\| \geq \frac{23}{16}\Delta \geq \sqrt{2}\Delta$. Applying Lemma 7, we have

$$\left| -(\boldsymbol{\beta}_1 - \boldsymbol{\theta}_3^*) \cdot \mathbb{E}_3\left[\Psi_3 \Psi_1\right] \right| \leq |\boldsymbol{\beta}_1 - \boldsymbol{\theta}_3^*| \cdot \mathbb{E}_3\left[\Psi_1\right]$$

$$\leq 2\Delta \cdot \left(1 + \frac{2}{\sqrt{\pi}\Delta}\right) \cdot e^{-\frac{\Delta^2}{16}}$$

(83)

$$\leq 3\Delta \cdot e^{-\frac{\Delta^2}{16}}.$$

- Similarly, for each $(s, j) \in [2] \times [3]$,

$$|(\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_j) \cdot \mathbb{E}_s\left[\Psi_3 \Psi_j\right]| \leq |\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_j| \cdot \mathbb{E}_s\left[\Psi_3\right]$$

$$\leq |(\boldsymbol{\beta}_3 - \boldsymbol{\theta}_3^*) + 2\Delta| \cdot \left(1 + \frac{128}{15\sqrt{2\pi} \cdot \Delta}\right) \cdot e^{-\frac{\left((\boldsymbol{\beta}_3 - \boldsymbol{\theta}_3^*) + \frac{15}{32}\Delta\right)^2}{32}}$$

(84)

$$\leq 3\Delta \cdot e^{-\frac{225}{32768}\Delta^2}.$$

Noticing that $\boldsymbol{\beta}_3 - \boldsymbol{\theta}_3^* \geq \boldsymbol{\theta}_3^* - \boldsymbol{\beta}_2$ (because $\boldsymbol{\theta}_2^* \in \mathcal{V}_2$), and gathering the upper bounds (83) and (84), we obtain from (82) that

$$(\boldsymbol{\beta}_3 - \boldsymbol{\theta}_3^*) \cdot \mathbb{E}_3\left[\Psi_3 \cdot (\Psi_3 - \Psi_2)\right] \leq 3\Delta \cdot e^{-\frac{\Delta^2}{16}} + 3\Delta \cdot e^{-\frac{225}{32768}\Delta^2}$$

$$\leq 6\Delta \cdot e^{-\frac{1}{146}\Delta^2}.$$

(85)

Next, we proceed to prove a lower bound for $\mathbb{E}_3\left[\Psi_3 \cdot (\Psi_3 - \Psi_2)\right]$. Defining a function $\psi^{\mathrm{prox}} : \mathbb{R} \to \mathbb{R}$ so that[5]

$$\psi^{\mathrm{prox}}(x) = \frac{e^{-\|x - \boldsymbol{\beta}_2\|^2/2}}{e^{-\|x - \boldsymbol{\beta}_2\|^2/2} + e^{-\|x - \boldsymbol{\beta}_3\|^2/2}},$$

Note that $\psi^{\mathrm{prox}}(x) \geq \psi_2(x)$ for all $x \in \mathbb{R}$. Letting $\Psi^{\mathrm{prox}} = \psi^{\mathrm{prox}}(\mathsf{x})$ and $\overline{\Psi}^{\mathrm{prox}} = 1 - \Psi^{\mathrm{prox}}$, we thus have

$$\mathbb{E}_3\left[\Psi_3 \cdot (\Psi_3 - \Psi_2)\right] \geq \mathbb{E}_3\left[\Psi_3 \cdot (\Psi_3 - \Psi^{\mathrm{prox}})\right]$$

$$= \mathbb{E}_3\left[\overline{\Psi}^{\mathrm{prox}} \cdot (\overline{\Psi}^{\mathrm{prox}} - \Psi^{\mathrm{prox}})\right] + \mathbb{E}_3\left[\Psi_3 \cdot (\Psi_3 - \Psi^{\mathrm{prox}})\right] - \mathbb{E}_3\left[\overline{\Psi}^{\mathrm{prox}} \cdot (\overline{\Psi}^{\mathrm{prox}} - \Psi^{\mathrm{prox}})\right]$$

$$= \mathbb{E}_3\left[\overline{\Psi}^{\mathrm{prox}} \cdot (\overline{\Psi}^{\mathrm{prox}} - \Psi^{\mathrm{prox}})\right] + \mathbb{E}_3\left[(\Psi_3 + \overline{\Psi}^{\mathrm{prox}} - \Psi^{\mathrm{prox}}) \cdot (\Psi_3 - \overline{\Psi}^{\mathrm{prox}})\right].$$

(86)

Now we analyze the two terms in the right-hand side of (86).

---

[5]Note that here we define ψ^{prox} differently from that used in Step 3 - Case A, although the underlying intuition is the same.

- First, we observe the following inequality by modifying the proof of Lemma 5, cf. (48):

$$\mathbb{E}_3\left[\overline{\Psi}^{\mathrm{prox}}\cdot(\overline{\Psi}^{\mathrm{prox}}-\Psi^{\mathrm{prox}})\right]=\mathbb{E}_3\left[\overline{\Psi}^{\mathrm{prox}}\cdot(2\overline{\Psi}^{\mathrm{prox}}-1)\right]$$
$$\geq\frac{1}{4}\left[Q\left(\frac{1}{\delta}-c\right)-\frac{\phi(c)}{\phi(c+2\alpha)}\cdot Q\left(\frac{1}{\delta}-c-2\alpha\right)\right] \tag{87}$$

where $\alpha=\boldsymbol{\beta}_2-\boldsymbol{\theta}_3^*$, $c=\frac{\boldsymbol{\beta}_2+\boldsymbol{\beta}_3}{2}-\boldsymbol{\theta}_3^*$, and $\delta=\frac{\boldsymbol{\beta}_3-\boldsymbol{\beta}_2}{2}$. Note that

$$\frac{\phi(c)}{\phi(c+2\alpha)}\cdot Q\left(\frac{1}{\delta}-c-2\alpha\right)=e^{\frac{2\alpha}{\delta}}\cdot\frac{\phi(\frac{1}{\delta}-c)}{\phi(\frac{1}{\delta}-c-2\alpha)}\cdot Q\left(\frac{1}{\delta}-c-2\alpha\right) \quad \because (45)$$
$$\leq e^{\frac{2\alpha}{\delta}}\cdot Q\left(\frac{1}{\delta}-c\right) \tag{88}$$

because $Q(x)/\phi(x)$ is monotone increasing over $x\in\mathbb{R}$.

- Second, since $|\Psi_3+\overline{\Psi}^{\mathrm{prox}}-\Psi^{\mathrm{prox}}|\leq 2$, we have

$$\left|\mathbb{E}_3\left[(\Psi_3+\overline{\Psi}^{\mathrm{prox}}-\Psi^{\mathrm{prox}})\cdot(\Psi_3-\overline{\Psi}^{\mathrm{prox}})\right]\right|$$
$$\leq 2\cdot\mathbb{E}_3\left[\overline{\Psi}^{\mathrm{prox}}-\Psi_3\right]$$
$$=2\cdot\mathbb{E}_3\left[\frac{\phi(\mathsf{x}-\boldsymbol{\beta}_1)\cdot\phi(\mathsf{x}-\boldsymbol{\beta}_3)}{[\phi(\mathsf{x}-\boldsymbol{\beta}_1)+\phi(\mathsf{x}-\boldsymbol{\beta}_2)+\phi(\mathsf{x}-\boldsymbol{\beta}_3)]}\cdot\frac{1}{[\phi(\mathsf{x}-\boldsymbol{\beta}_2)+\phi(\mathsf{x}-\boldsymbol{\beta}_3)]}\right]$$
$$\leq 2\cdot\mathbb{E}_3\left[\Psi_1\right]$$
$$\overset{(a)}{\leq} 2\cdot\left(1+\frac{2}{\sqrt{\pi}\Delta}\right)\cdot e^{-\frac{\Delta^2}{16}}$$
$$\leq 3\cdot e^{-\frac{\Delta^2}{16}}, \tag{89}$$

where (a) follows from Lemma 7; also see (83).

Combining (87), (88) and (89) with (86), we obtain

$$\mathbb{E}_3\left[\Psi_3\cdot(\Psi_3-\Psi_2)\right]\geq\frac{1}{4}\cdot\left(1-e^{4\frac{\boldsymbol{\beta}_2-\boldsymbol{\theta}_3^*}{\boldsymbol{\beta}_3-\boldsymbol{\beta}_2}}\right)\cdot Q\left(\frac{2}{\boldsymbol{\beta}_3-\boldsymbol{\beta}_2}-\frac{\boldsymbol{\beta}_2+\boldsymbol{\beta}_3}{2}+\boldsymbol{\theta}_3^*\right)-3\cdot e^{-\frac{\Delta^2}{16}}. \tag{90}$$

Inserting (90) to (85) and observing $\boldsymbol{\beta}_3-\boldsymbol{\theta}_3^*\geq\frac{1}{2}(\boldsymbol{\beta}_3-\boldsymbol{\beta}_2)$ yields

$$\frac{1}{8}\cdot(\boldsymbol{\beta}_3-\boldsymbol{\beta}_2)\cdot\left(1-e^{4\frac{\boldsymbol{\beta}_2-\boldsymbol{\theta}_3^*}{\boldsymbol{\beta}_3-\boldsymbol{\beta}_2}}\right)\cdot Q\left(\frac{2}{\boldsymbol{\beta}_3-\boldsymbol{\beta}_2}-\frac{\boldsymbol{\beta}_2+\boldsymbol{\beta}_3}{2}+\boldsymbol{\theta}_3^*\right)\leq 9\Delta\cdot e^{-\frac{1}{146}\Delta^2}. \tag{91}$$

If the inequality (91) holds, then either of the following must be true: (1) $\boldsymbol{\beta}_3-\boldsymbol{\beta}_2$ is very small, namely, $\boldsymbol{\beta}_3-\boldsymbol{\beta}_2\leq\frac{20}{\Delta}$, or (2) $\boldsymbol{\beta}_2-\boldsymbol{\theta}_3^*$ is positive, or very close to 0 so that $\boldsymbol{\theta}_3^*-\boldsymbol{\beta}_2\leq e^{-C\cdot\Delta^2}$ for some sufficiently small constant (e.g., $C=1/200$). Note that (1) cannot be the case, as it will violate the near-empty association condition $\mathbb{P}_3[\mathcal{V}_3]\approx 0$. Also, $\boldsymbol{\beta}_2-\boldsymbol{\theta}_3^*$ cannot be positive. Thus, the only possibility is to have $\boldsymbol{\theta}_3^*-\boldsymbol{\beta}_2\leq e^{-C\cdot\Delta^2}$, which implies that $\boldsymbol{\beta}_3-\boldsymbol{\theta}_3^*\geq C'\Delta$ (e.g., $C'=1/20$).

All in all, we conclude that this subcase may not be possible to happen, but if it occurs, then the following must be true: there exist some absolute constant $C>0$ such that

$$\left\|\boldsymbol{\beta}_1-\frac{1}{2}(\boldsymbol{\theta}_1^*+\boldsymbol{\theta}_2^*)\right\|\leq e^{-C\Delta^2},$$
$$\left\|\boldsymbol{\beta}_2-\boldsymbol{\theta}_3^*\right\|\leq e^{-C\Delta^2}, \quad\text{and}$$
$$\mathbb{E}_*[\Psi_3]\leq e^{-C\Delta^2}.$$

The approximation error bounds follow form Lemma 8.

Subcase (iii): $\boldsymbol{\beta}_3\in(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)$. We reach at a similar conclusion to Subcase (ii) using a similar argument. We omit the details.

Step 3 - Case D.   Recall from (51) in Step 0 of this proof that

$$\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_i^*\| < \frac{\Delta}{32}, \qquad \forall i \in [3].$$

Thus, $\min_{j \neq i} \|\boldsymbol{\beta}_j - \boldsymbol{\theta}_i^*\| - \|\boldsymbol{\beta}_i - \boldsymbol{\theta}_i^*\| > \frac{15\Delta}{16}$ for all $i \in [3]$. Observe that the premise of Lemma 8 is satisfied with $\delta = 15\Delta/16 \geq \max\left\{\frac{4}{\sqrt{2\pi}}, 8\sqrt{\log(2\sqrt{2} \cdot kk_*)}\right\}$. Therefore, it follows from Lemma 8 that for all $i \in [3]$,

$$\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_i^*\| \leq 4kk_* \Delta_{\max} \cdot e^{-\frac{\delta^2}{64}}$$
$$= 72\Delta \cdot e^{-\frac{225}{16384}\Delta^2}$$

because $\Delta_{\max} = 2\Delta$ and $\delta = \frac{15}{16}\Delta$.

8.3 Proof of Corollary 5

Proof of Corollary 5. The proof follows the same lines as in the proof of Theorem 3 in Section 8. By rescaling, we may assume unit variance $\sigma^2 = 1$. When $k_* = 3$ and $k = 2$, the value q and the sets $\{S_a\}$ and $\{S_a^*\}$ in Theorem 4 can only have, up to permutation of component labels, the following possibilities:

1. $(q, q_0, |\mathcal{A}_0|) = (1, 1, 1)$.

    (a) $S_0 = \{2\}$; $S_1 = \{1\}$, $S_1^* = \{1, 2, 3\}$.

2. $(q, q_0, |\mathcal{A}_0|) = (2, 2, 0)$.

    (a) $S_1 = \{1\}, S_1^* = \{1, 2\}$; $S_2 = \{2\}, S_2^* = \{3\}$;
    (b) $S_1 = \{1\}, S_1^* = \{1, 3\}$; $S_2 = \{2\}, S_2^* = \{2\}$.

We claim that Case 2b above, where $\boldsymbol{\beta}_1$ fits two non-adjacent centers $\{\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_3^*\}$, is impossible. If we assume it is possible, then we must have $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$ by Theorem 2; say $\boldsymbol{\beta}_1 < \boldsymbol{\beta}_2$. In this case, it holds that $\mathcal{V}_1 \subset (-\infty, \boldsymbol{\beta}_2) \subset (-\infty, \boldsymbol{\theta}_3^*]$, where the last inclusion holds since $|\boldsymbol{\beta}_2 - \boldsymbol{\theta}_2^*| \leq \Delta/192$ by (52) (cf. Theorem 4). It follows that $\mathbb{P}_3(\mathcal{V}_1) \leq \mathbb{P}_3\left((-\infty, \boldsymbol{\theta}_3^*)\right) = \frac{1}{2}$, contradicting the inequality $\mathbb{P}_3(\mathcal{V}_1) \geq 1 - 1/(2^9 \cdot 3^3)$ in (53).

In Case 1 above, $\boldsymbol{\beta}_1$ fits all three true centers and $\boldsymbol{\beta}_2$ has near-empty association. This case is impossible by an argument similar to Case A in the proof of Theorem 3 (see Section 8.2, Step 3 - Case A).

In Case 2a above, $\boldsymbol{\beta}_1$ fits $\{\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*\}$ and $\boldsymbol{\beta}_2$ fits $\boldsymbol{\theta}_3^*$. By an argument similar to Case B in the proof of Theorem 3 (see Section 8.2, Step 3 - Case B), we find that the exponential error bounds in equation (26) of Corollary 5 must hold. $\square$

References

[1] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55. US Government printing office, 1948.

[2] Morton M. Astrahan. Speech analysis by clustering, or the hyperphoneme method. Technical report, Stanford University Dept of Computer Science, CA, 1970.

[3] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

[4] Geoffrey H Ball and David J Hall. PROMENADE–An on-line pattern recognition system. Technical report, Stanford Research Inst Menlo Park CA, 1967.

[5] Ali Ridho Barakbah and Afrida 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.

[6] Ali Ridho Barakbah and Yasushi 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*, pages 61–68. IEEE, 2009.

[7] Per-Ola Borjesson and Carl-Erik W. Sundberg. Simple approximations of the error function q (x) for communications applications. *IEEE Transactions on Communications*, 27(3):639–643, 1979.

[8] Rares-Darius Buhai, Yoni Halpern, Yoon Kim, Andrej Risteski, and David Sontag. Empirical study of the benefits of overparameterization in learning latent variable models. In *International Conference on Machine Learning*, pages 1211–1219. PMLR, 2020.

[9] T. Tony Cai, Jing Ma, and Linjun Zhang. CHIME: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267, 2019.

[10] Jiahua Chen. Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, 23(1):221–233, 1995.

[11] Yen-Chi Chen. Statistical inference with local optima. *Journal of the American Statistical Association*, pages 1–13, 2022.

[12] Sanjoy Dasgupta. Learning mixtures of Gaussians. In *40th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 634–644. IEEE, 1999.

[13] Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learning Research*, 8(Feb):203–226, 2007.

[14] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of EM suffice for mixtures of two Gaussians. In *Conference on Learning Theory*, pages 704–710, 2017.

[15] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[16] Luke Dicks and DJ Wales. Elucidating the solution structure of the k-means cost function using energy landscape theory. *The Journal of Chemical Physics*, 156(5):054109, 2022.

[17] Raaz Dwivedi, Nhat Ho, Koulik Khamaru, Martin Wainwright, Michael Jordan, and Bin Yu. Sharp analysis of expectation-maximization for weakly identifiable models. In *International Conference on Artificial Intelligence and Statistics*, pages 1866–1876. PMLR, 2020.

[18] Raaz Dwivedi, Nhat Ho, Koulik Khamaru, Martin J. Wainwright, Michael I. Jordan, and Bin Yu. Singularity, misspecification and the convergence rate of EM. *The Annals of Statistics*, 48(6):3161 – 3182, 2020.

[19] Raaz Dwivedi, Koulik Khamaru, Martin J. Wainwright, and Michael I. Jordan. Theoretical guarantees for EM under misspecified Gaussian mixture models. In *Advances in Neural Information Processing Systems*, pages 9681–9689, 2018.

[20] Botao Hao, Will Wei Sun, Yufeng Liu, and Guang Cheng. Simultaneous clustering and estimation of heterogeneous graphical models. *The Journal of Machine Learning Research*, 18(1):7981–8038, 2017.

[21] Philippe Heinrich and Jonas Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6A):2844–2870, 2018.

[22] Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J. Wainwright, and Michael I. Jordan. Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems*, pages 4116–4124, 2016.

[23] Raunak Kumar and Mark Schmidt. Convergence rate of expectation-maximization. In *10th NIPS Workshop on Optimization for Machine Learning*, 2017.

[24] Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the minimax optimality of the EM algorithm for learning two-component mixed linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1405–1413. PMLR, 2021.

[25] Jeongyeol Kwon, Wei Qian, Constantine Caramanis, Yudong Chen, and Damek Davis. Global convergence of EM algorithm for mixtures of two component linear regression. In *Conference on Learning Theory (COLT)*, 2019.

[26] Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257. PMLR, 2016.

[27] Dong Li, Shuisheng Zhou, Tieyong Zeng, and Raymond H Chan. Multi-prototypes convex merging based k-means clustering algorithm. *arXiv preprint arXiv:2302.07045*, 2023.

[28] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

[29] XuanLong Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.

[30] Donald Bruce Owen. A table of normal integrals: A table. *Communications in Statistics-Simulation and Computation*, 9(4):389–419, 1980.

[31] Yury Polyanskiy and Yihong Wu. Self-regularizing property of nonparametric maximum likelihood estimator in mixture models. *arXiv preprint arXiv:2008.08244*, 2020.

[32] Wei Qian, Yuqian Zhang, and Yudong Chen. Global convergence of least squares EM for demixing two log-concave densities. In *Neural Information Processing Systems (NeurIPS)*, 2019.

[33] Wei Qian, Yuqian Zhang, and Yudong Chen. Structures of spurious local minima in k-means. *IEEE Transactions on Information Theory*, 68(1):395–422, 2021.

[34] Nathan Srebro. Are there local maxima in the infinite-sample likelihood of Gaussian mixture estimation? In *International Conference on Computational Learning Theory*, pages 628–629, 2007.

[35] Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 1135–1151, 1981.

[36] Naonori Ueda, Ryohei Nakano, Zoubin Ghahramani, and Geoffrey E Hinton. SMEM algorithm for mixture models. *Advances in Neural Information Processing Systems*, 11, 1998.

[37] Naonori Ueda, Ryohei Nakano, Zoubin Ghahramani, and Geoffrey E Hinton. Split and merge em algorithm for improving Gaussian mixture density estimates. *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, 26:133–140, 2000.

[38] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[39] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

[40] Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *arXiv preprint arXiv:1412.8729*, 2014.

[41] Chien-Fu Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[42] Yihong Wu and Harrison H Zhou. Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations. *Mathematical Statistics and Learning*, 4(3):143–220, 2022.

[43] Ji Xu, Daniel J. Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two Gaussians. In *Advances In Neural Information Processing Systems*, pages 2676–2684, 2016.

[44] Ji Xu, Daniel J. Hsu, and Arian Maleki. Benefits of over-parameterization with EM. In *Advances in Neural Information Processing Systems*, pages 10662–10672, 2018.

[45] Lei Xu and Michael I. Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8(1):129–151, 1996.

[46] Bowei Yan, Mingzhang Yin, and Purnamrita Sarkar. Convergence of gradient EM on multi-component mixture of Gaussians. In *Advances in Neural Information Processing Systems*, pages 6956–6966, 2017.

[47] Yuling Yan, Kaizheng Wang, and Philippe Rigollet. Learning gaussian mixtures using the Wasserstein-Fisher-Rao gradient flow. *arXiv preprint arXiv:2301.01766*, 2023.

[48] Xinyang Yi and Constantine Caramanis. Regularized EM algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 1567–1575, 2015.

[49] Zhihua Zhang, Chibiao Chen, Jian Sun, and Kap Luk Chan. EM algorithms for Gaussian mixtures with split-and-merge operation. *Pattern Recognition*, 36(9):1973–1983, 2003.

A   Deferred Proof of Proposition 1

A.1   Voronoi Cells and Their Geometry

In our analysis, the coefficients of association $\psi_i(\boldsymbol{x})$, $i \in [k]$ defined in Definition 1 play a key role in characterizing the gradient, the Hessian and optimality conditions of L. These quantities represent the strength of (soft-) associations between a data point $\boldsymbol{x} \in \mathbb{R}^d$ and the centers $\boldsymbol{\beta}_i$, which is quantified by the relative magnitudes of the squared distances between $\boldsymbol{x}$ and the k centers; see (8). To better understand the properties of $\psi_i(\boldsymbol{x})$, it is useful to study the hard-association analogue thereof (i.e., $\psi_i(\boldsymbol{x})$ in the limit $\sigma \to 0$), where a data point is associated only with the closest among the k centers. This hard association induces a partition of the space $\mathbb{R}^d$, which is the so-called Voronoi diagram of $\mathbb{R}^d$ generated by $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k\}$.

In this section, we take a closer look at this Voronoi diagram and elucidate its relationship with the association coefficients $\psi_i(\boldsymbol{x})$. Specifically, we recall the definition of Voronoi cells and introduce additional useful notions related to them. Thereafter, we state and prove Lemma 9 that will be used in the proof of Proposition 1 in Section A.3.

Useful notions related to Voronoi cells We begin by recalling the definition of Voronoi cells from Definition 2. Letting $\boldsymbol{B} = (\boldsymbol{\beta}_i)_{i=1}^k$, the i-th Voronoi cell associated with $\boldsymbol{B}$ has the following representations, cf. (17):

$$
\begin{aligned}
\mathcal{V}_i &= \mathcal{V}_i(\boldsymbol{B}) \\
&:= \left\{ \boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{\beta}_i\| \leq \|\boldsymbol{x} - \boldsymbol{\beta}_j\|, \forall j \in [k] \right\} \\
&= \left\{ \boldsymbol{x} \in \mathbb{R}^d : \psi_i(\boldsymbol{x}) \geq \psi_j(\boldsymbol{x}), \ \forall j \in [k] \right\}.
\end{aligned}
$$

For $i, j \in [k]$ with $i \neq j$, we let

$$
\bar{\boldsymbol{\beta}}_{ij} := \frac{\boldsymbol{\beta}_i + \boldsymbol{\beta}_j}{2} \tag{92}
$$

denote the mid point of $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$. Then the Voronoi cell can be equivalently represented as

$$
\mathcal{V}_i = \left\{ \boldsymbol{x} \in \mathbb{R}^d : \left\langle \boldsymbol{x} - \bar{\boldsymbol{\beta}}_{ij}, \boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_{ij} \right\rangle \geq 0, \qquad \forall j \in [k] \setminus \{i\} \right\}.
$$

This representation makes it clear that $\mathcal{V}_i$ is a polyhedron generated by at most $k - 1$ linear inequalities.

Next, we define the set of points that are equidistant from $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$: for $i, j \in [k]$ with $i \neq j$,

$$
\begin{aligned}
\partial_{ij} &:= \left\{ \boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{\beta}_i\| = \|\boldsymbol{x} - \boldsymbol{\beta}_j\| \right\} \\
&= \left\{ \boldsymbol{x} \in \mathbb{R}^d : \psi_i(\boldsymbol{x}) = \psi_j(\boldsymbol{x}) \right\} \\
&= \left\{ \boldsymbol{x} \in \mathbb{R}^d : \left\langle \boldsymbol{x} - \bar{\boldsymbol{\beta}}_{ij}, \boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_{ij} \right\rangle = 0 \right\}.
\end{aligned} \tag{93}
$$

It is clear from the last expression of ∂_{ij} that ∂_{ij} is an affine subspace of codimension at most 1; observe that $\partial_{ij} = \mathbb{R}^d$ if and only if $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$. Note that ∂_{ij} is the affine hull of $\mathcal{V}_i \cap \mathcal{V}_j$.

Remark 7. Note that if the $\boldsymbol{\beta}_i$'s in $\boldsymbol{B}$ are distinct, then so are their associated Voronoi cells. In this case, the Voronoi cells $\{\mathcal{V}_i\}_{i=1}^k$ form a partition of $\mathbb{R}^d$, up to the (Lebesgue) measure-zero boundaries $\mathcal{V}_i \cap \mathcal{V}_j \subseteq \partial_{ij}$. On the other hand, if $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$ for some pair $i, j \in [k]$, then $\mathcal{V}_i = \mathcal{V}_j$ and $\partial_{ij} = \mathbb{R}^d$.

For a parameter $\alpha \geq 0$, we define two parameterized families of sets

$$
\begin{aligned}
\widetilde{\mathcal{V}}_i^\alpha &:= \left\{ \boldsymbol{x} \in \mathbb{R}^d : \left\langle \boldsymbol{x} - \bar{\boldsymbol{\beta}}_{ij}, \frac{\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_{ij}}{\|\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_{ij}\|} \right\rangle \geq -\alpha\sigma, \quad \forall j \in [k] \setminus \{i\} \right\}, \\
\widetilde{\partial}_{ij}^\alpha &:= \left\{ \boldsymbol{x} \in \mathbb{R}^d : \left| \left\langle \boldsymbol{x} - \bar{\boldsymbol{\beta}}_{ij}, \frac{\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_{ij}}{\|\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_{ij}\|} \right\rangle \right| \leq \alpha\sigma \right\}.
\end{aligned} \tag{94}
$$

The sets $\widetilde{\mathcal{V}}_i^\alpha$ and $\widetilde{\partial}_{ij}^\alpha$ are the $\alpha \cdot \sigma$-enlargements of $\mathcal{V}_i$ and ∂_{ij}, respectively, with $\widetilde{\mathcal{V}}_i^0 = \mathcal{V}_i$ and $\widetilde{\partial}_{ij}^0 = \partial_{ij}$.

Remark 8. For any $\alpha \geq 0$, and any $i, j \in [k]$ with $i \neq j$, we have $\widetilde{\mathcal{V}}_i^\alpha \cap \widetilde{\mathcal{V}}_j^\alpha \subseteq \widetilde{\partial}_{ij}^\alpha$. Moreover, if $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$ for some pair $i, j \in [k]$, then $\widetilde{\mathcal{V}}_i^\alpha = \widetilde{\mathcal{V}}_j^\alpha$ and $\widetilde{\partial}_{ij}^\alpha = \mathbb{R}^d$.

Finally, for a parameter $\delta > 0$ and $i, j \in [k]$ with $i \neq j$, we let

$$
\mathcal{G}_{ij}^\delta := \left\{ \boldsymbol{x} \in \mathbb{R}^d : \psi_i(\boldsymbol{x})\psi_j(\boldsymbol{x}) \geq \delta \right\}, \tag{95}
$$

which denotes the set of points $\boldsymbol{x} \in \mathbb{R}^d$ that are simultaneously associated with both $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ (at level δ).

A useful lemma The following lemma establishes the relationship among (1) the association coefficient $\psi_i(\boldsymbol{x})$; (2) the α-enlarged Voronoi cells $\widetilde{\mathcal{V}}_i^\alpha, \widetilde{\partial}_{ij}^\alpha$; and (3) the set $\mathcal{G}_{ij}^\delta$ for some choices of $\alpha \geq 0$ and $\delta > 0$.

Lemma 9 (Soft Voronoi cells and boundaries)**.** *Let $\boldsymbol{B} \in \mathbb{R}^{d \times k}$ be an arbitrary ordered set of vectors. For any $\alpha \in \mathbb{R}$, let*

$$
\alpha' = \alpha'_{\alpha, \boldsymbol{B}, \sigma} := \alpha + \frac{\sigma}{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|} \log k,
$$

$$
\delta = \delta_{\alpha, \boldsymbol{B}, \sigma} := \frac{1}{k^2} \exp\left(-2\alpha \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \right),
$$

$$
\delta' = \delta'_{\alpha, \boldsymbol{B}, \sigma} := \frac{1}{k^2} \exp\left(-3\alpha \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \right).
$$

For each $i, j \in [k]$ such that $i \neq j$, and for any $\alpha \in \mathbb{R}$, the following inclusion relations hold.

49

1. *First, letting*

$$\mathcal{S} := \left\{ \boldsymbol{x} \in \mathbb{R}^d : \psi_i(\boldsymbol{x}) \geq \frac{1}{k} \exp\left(-\alpha \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right) \right\},$$

   we observe that

$$\widetilde{\mathcal{V}}_i^\alpha \subseteq \mathcal{S} \subseteq \widetilde{\mathcal{V}}_i^{\alpha'}; \tag{96}$$

2. *Second,*

$$\widetilde{\mathcal{V}}_i^\alpha \cap \widetilde{\mathcal{V}}_j^\alpha \subseteq \mathcal{G}_{ij}^\delta \subseteq \widetilde{\mathcal{V}}_i^{2\alpha'} \cap \widetilde{\mathcal{V}}_j^{2\alpha'}. \tag{97}$$

3. *Third,*

$$\widetilde{\partial}_{ij}^\alpha \cap \widetilde{\mathcal{V}}_j^\alpha \subseteq \mathcal{G}_{ij}^{\delta'}. \tag{98}$$

Proof of Lemma 9. We begin this proof by making a preparatory observation. Fix arbitrary $i, j \in [k]$ with $i \neq j$, and recall from (92) that $\bar{\boldsymbol{\beta}}_{ij} = (\boldsymbol{\beta}_i + \boldsymbol{\beta}_j)/2$. Then we observe the equivalence of the following expressions: for any $\alpha \in \mathbb{R}$,

$$\left\langle \boldsymbol{x} - \bar{\boldsymbol{\beta}}_{ij}, \frac{\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_{ij}}{\|\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_{ij}\|} \right\rangle \geq -\alpha\sigma$$

$$\Leftrightarrow \frac{1}{2\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|} \left\{ \|\boldsymbol{x} - \boldsymbol{\beta}_j\|^2 - \|\boldsymbol{x} - \boldsymbol{\beta}_i\|^2 \right\} \geq -\alpha\sigma$$

$$\Leftrightarrow -\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\beta}_i\|^2 \geq -\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\beta}_j\|^2 - \alpha\sigma\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|$$

$$\Leftrightarrow \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x} - \boldsymbol{\beta}_i\|^2\right)$$

$$\geq \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x} - \boldsymbol{\beta}_j\|^2\right) \cdot \exp\left(-\alpha\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right)$$

$$\Leftrightarrow f_i(\boldsymbol{x}) \geq f_j(\boldsymbol{x}) \cdot \exp\left(-\alpha\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right)$$

$$\Leftrightarrow \psi_i(\boldsymbol{x}) \geq \psi_j(\boldsymbol{x}) \cdot \exp\left(-\alpha\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right). \tag{99}$$

In the remainder of this proof, we prove each of the three claims in the lemma.

<u>*(1) Proof of Claim 1.*</u> First of all, we observe by (94) and (99) that for any $\alpha \in \mathbb{R}$,

$$\boldsymbol{x} \in \widetilde{\mathcal{V}}_i^\alpha \quad \implies \quad \psi_i(\boldsymbol{x}) \geq \psi_j(\boldsymbol{x}) \cdot \exp\left(-\alpha\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right), \quad \forall j \in [k] \setminus \{i\}.$$

Because $\sum_{i' \in [k]} \psi_{i'}(\boldsymbol{x}) \equiv 1$, we obtain

$$\psi_i(\boldsymbol{x}) \geq \frac{1}{1 + (k-1) \cdot \exp\left(\alpha\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right)}$$

$$\geq \frac{1}{k} \exp\left(-\alpha\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right).$$

To prove the second inclusion, we observe that for any $\alpha'' \in \mathbb{R}_+$,

$$\psi_i(\boldsymbol{x}) \geq \alpha'' \quad \implies \quad \psi_i(\boldsymbol{x}) = \frac{f_i(\boldsymbol{x})}{\sum_{j \in [k]} f_j(\boldsymbol{x})} \geq \alpha''$$

$$\implies \quad f_i(\boldsymbol{x}) \geq \alpha'' \sum_{j \in [k]} f_j(\boldsymbol{x})$$

$$\implies \quad f_i(\boldsymbol{x}) \geq \alpha'' \max_{\substack{j \in [k] \\ j \neq i}} f_j(\boldsymbol{x}).$$

50

Combining the last inequality with (99), we obtain

$$\psi_i(\boldsymbol{x}) \geq \alpha'' \qquad \Longrightarrow \qquad \boldsymbol{x} \in \widetilde{\mathcal{V}}_i^{\alpha'} \quad \text{for} \quad \alpha' = \frac{\sigma}{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|} \cdot \log\left(\frac{1}{\alpha''}\right). \tag{100}$$

Choosing $\alpha'' = \frac{1}{k}\exp\left(-\alpha\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right)$ yields $\alpha' = \alpha + \frac{\sigma}{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}\log k$.

(2) Proof of Claim 2. By the first claim of this lemma, i.e., (96), we observe that

$$\boldsymbol{x} \in \widetilde{\mathcal{V}}_i^{\alpha} \cap \widetilde{\mathcal{V}}_j^{\alpha}$$
$$\Longrightarrow \quad \min\{\psi_i(\boldsymbol{x}), \psi_j(\boldsymbol{x})\} \geq \frac{1}{k}\exp\left(-\alpha\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right)$$
$$\Longrightarrow \quad \psi_i(\boldsymbol{x})\psi_j(\boldsymbol{x}) \geq \frac{1}{k^2}\exp\left(-2\alpha\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right).$$

Thus, it follows from the definition of the set $\mathcal{G}_{ij}^{\delta}$ in (95) that

$$\widetilde{\mathcal{V}}_i^{\alpha} \cap \widetilde{\mathcal{V}}_j^{\alpha} \subseteq \mathcal{G}_{ij}^{\delta} \qquad \text{for} \quad \delta = \frac{1}{k^2}\exp\left(-2\alpha\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right).$$

As $\max\{\psi_i(\boldsymbol{x}), \psi_j(\boldsymbol{x})\} \leq 1$, $\psi_i(\boldsymbol{x})\psi_j(\boldsymbol{x}) \geq \delta$ implies $\min\{\psi_i(\boldsymbol{x}), \psi_j(\boldsymbol{x})\} \geq \delta$. Therefore, by the same argument as in (100), we get

$$\boldsymbol{x} \in \mathcal{G}_{ij}^{\delta} \quad \Longrightarrow \quad \boldsymbol{x} \in \widetilde{\mathcal{V}}_i^{\delta'} \cap \widetilde{\mathcal{V}}_j^{\delta'}$$

where

$$\delta' = \frac{\sigma}{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|} \cdot \log\left(\frac{1}{\delta}\right).$$

(3) Proof of Claim 3. Finally, we note that

$$\boldsymbol{x} \in \widetilde{\partial}_{ij}^{\alpha} \cap \widetilde{\mathcal{V}}_j^{\alpha}$$
$$\Longrightarrow \quad \boldsymbol{x} \in \widetilde{\mathcal{V}}_j^{\alpha} \quad \& \quad \left\langle \boldsymbol{x} - \bar{\boldsymbol{\beta}}_{ij}, \frac{\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_{ij}}{\|\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_{ij}\|} \right\rangle \geq -\alpha$$
$$\Longrightarrow \quad \psi_j(\boldsymbol{x}) \geq \frac{1}{k}\exp\left(-\alpha\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right) \quad \& \quad \psi_i(\boldsymbol{x}) \geq \psi_j(\boldsymbol{x}) \cdot \exp\left(-\alpha\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right)$$

by (96) and (99). Therefore,

$$\boldsymbol{x} \in \widetilde{\partial}_{ij}^{\alpha} \cap \widetilde{\mathcal{V}}_j^{\alpha} \qquad \Longrightarrow \qquad \psi_i(\boldsymbol{x})\psi_j(\boldsymbol{x}) \geq \frac{1}{k^2}\exp\left(-3\alpha\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right).$$

$\square$

A.2 Helper Lemmas for the Proof of Proposition 1

Gaussian lemmas Let $\phi_\sigma : \mathbb{R} \to \mathbb{R}$ denote the probability density function of the Gaussian distribution $\mathcal{N}(0, \sigma^2)$, i.e., $\phi_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-x^2/(2\sigma^2)}$ for all $x \in \mathbb{R}$. We denote by $\phi = \phi_1$, omitting the subscript when $\sigma = 1$. We observe that $\phi_\sigma(x) = \frac{1}{\sigma}\phi(x/\sigma)$ for all $\sigma \in \mathbb{R}_+$ and all $x \in \mathbb{R}$.

Here we state and prove a simple technical lemma that will be used later in our analysis.

Lemma 10. _For any $\sigma \in \mathbb{R}_+$, any $\tau \in \mathbb{R}_+$, and any $t_1, t_2 \in \mathbb{R}$ such that $0 \leq t_1 \leq t_2 \leq \infty$,_

$$\int_{t_1}^{t_2} \phi_\sigma(z)\mathrm{d}z \leq \left(\sqrt{2\pi} + 1\right) \cdot \max\left\{1, \frac{\sigma}{\tau}\right\} \int_{t_1}^{t_2} \mathbb{1}\{z - t_1 \leq \tau\}\phi_\sigma(z)\,\mathrm{d}z.$$

Proof of Lemma 10. If $\tau \geq t_2 - t_2$, then the conclusion trivially follows. In the rest of this proof, we assume $\tau < t_2 - t_1$ and consider two cases separately.

Case 1: $\tau \geq \sigma$. In this case, we observe that

$$
\begin{aligned}
\int_{t_1+\tau}^{t_2} \phi_\sigma(z)\mathrm{d}z &= \int_{\frac{t_1+\tau}{\sigma}}^{\frac{t_2}{\sigma}} \phi(z)\mathrm{d}z \\
&\overset{(a)}{\leq} \sqrt{2\pi} \cdot \phi\left(\frac{t_1+\tau}{\sigma}\right) \\
&\overset{(b)}{\leq} \frac{\tau}{\sigma} \cdot \sqrt{2\pi} \cdot \phi\left(\frac{t_1+\tau}{\sigma}\right) \\
&\overset{(c)}{\leq} \sqrt{2\pi} \int_{\frac{t_1}{\sigma}}^{\frac{t_1+\tau}{\sigma}} \phi(z)\mathrm{d}z \\
&= \sqrt{2\pi} \int_{t_1}^{t_1+\tau} \phi_\sigma(z)\mathrm{d}z,
\end{aligned}
$$

where (a) follows from Lemma 4, (b) is due to $\tau \geq \sigma$, and (c) holds because ϕ is non-increasing on $\left[\frac{t_1}{\sigma}, \frac{t_1+\tau}{\sigma}\right]$. It follows that

$$
\begin{aligned}
\int_{t_1}^{t_2} \phi(z)\mathrm{d}z &= \int_{t_1}^{t_1+\tau} \phi(z)\mathrm{d}z + \int_{t_1+\tau}^{t_2} \phi(z)\mathrm{d}z \\
&\leq \left(\sqrt{2\pi}+1\right) \int_{t_1}^{t_1+\tau} \phi(z)\mathrm{d}z \\
&= \left(\sqrt{2\pi}+1\right) \int_{t_1}^{t_2} \mathbb{1}\left\{z - t_1 \leq \tau\right\} \phi(z)\mathrm{d}z.
\end{aligned}
$$

Case 2: $\tau < \sigma$. In this case, we follow a similar argument as above, observing that

$$
\begin{aligned}
\int_{t_1+\tau}^{t_2} \phi_\sigma(z)\mathrm{d}z &= \int_{\frac{t_1+\tau}{\sigma}}^{\frac{t_2}{\sigma}} \phi(z)\mathrm{d}z \\
&\leq \sqrt{2\pi} \cdot \phi\left(\frac{t_1+\tau}{\sigma}\right) \\
&\leq \left(\sqrt{2\pi}+1-\frac{\tau}{\sigma}\right) \cdot \phi\left(\frac{t_1+\tau}{\sigma}\right) \\
&\leq \left(\sqrt{2\pi}+1-\frac{\tau}{\sigma}\right) \cdot \frac{\sigma}{\tau} \cdot \int_{\frac{t_1}{\sigma}}^{\frac{t_1+\tau}{\sigma}} \phi(z)\mathrm{d}z \\
&= \left(\left(\sqrt{2\pi}+1\right)\frac{\sigma}{\tau}-1\right) \cdot \int_{t_1}^{t_1+\tau} \phi_\sigma(z)\mathrm{d}z.
\end{aligned}
$$

Consequently, it follows that

$$
\begin{aligned}
\int_{t_1}^{t_2} \phi_\sigma(z)\mathrm{d}z &= \int_{t_1}^{t_1+\tau} \phi_\sigma(z)\mathrm{d}z + \int_{t_1+\tau}^{t_2} \phi_\sigma(z)\mathrm{d}z \\
&\leq \frac{\sigma}{\tau} \cdot \left(\sqrt{2\pi}+1\right) \cdot \int_{t_1}^{t_1+\tau} \phi_\sigma(z)\mathrm{d}z.
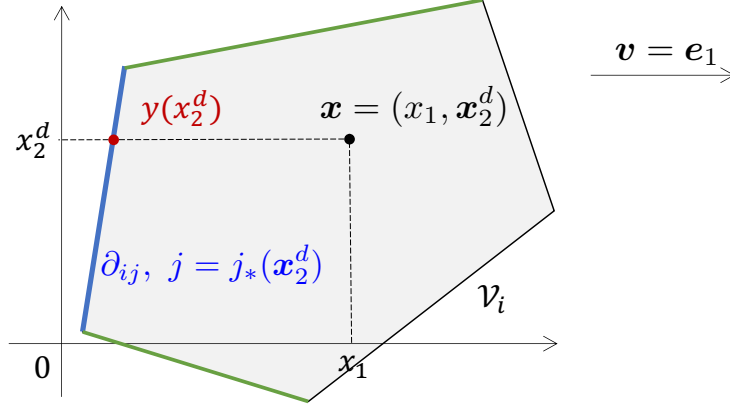\end{aligned}
$$

$\square$

Figure 5: Illustration for the proof of Lemma 11. The figure shows the polyhedral Voronoi cell $\mathcal{V}_i$ and the direction $\boldsymbol{v} = \boldsymbol{e}_1$ that defines the separating hyperplane between $\mathcal{V}_i$ and $\mathbf{0}$. For each point $\boldsymbol{x} = (x_1, \boldsymbol{x}_2^d)^\top \in \operatorname{int} \mathcal{V}_i$, the ray $\{\boldsymbol{x} - b\boldsymbol{e}_1, b \geq 0\}$ intersects a facet F of $\mathcal{V}_i$ at a unique point $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}_2^d)$, where $F \subseteq \partial_{ij}$ for some $j = j_*(\boldsymbol{x}_2^d) \in [k]$. The set $\mathcal{L}_i$ indexes the Voronoi boundaries colored in blue and green.

A useful geometric lemma   Our proof of Proposition 1 relies on the following geometric lemma.

Lemma 11 (Controlling volume by intersection). *Let $\boldsymbol{B} \in \mathbb{R}^{d \times k}$ be an arbitrary ordered set of vectors. For any $(s, i) \in [k_*] \times [k]$ and any $(\alpha_j \in \mathbb{R}_+ : j \in [k] \setminus \{i\})$, if $\boldsymbol{\theta}_s^* \notin \operatorname{int} \mathcal{V}_i$, then*

$$\mathbb{P}_s\left(\mathcal{V}_i\right) \leq \left(\sqrt{2\pi} + 1\right) \sum_{j \in [k] \setminus \{i\}} \max\left\{1, \ \frac{1}{\alpha_j}\right\} \cdot \mathbb{P}_s\left(\widetilde{\mathcal{V}}_i^{\alpha_j} \cap \widetilde{\partial}_{ij}^{\alpha_j}\right). \tag{101}$$

Proof of Lemma 11. We present this proof in four steps. In Step 1, we introduce some notation for the convenience of the later steps. In Step 2, we derive an useful expression for the probability $\mathbb{P}_s(\mathcal{V}_i)$; see (103). In Step 3, we establish a lower bound for $\mathbb{P}_s\left(\widetilde{\mathcal{V}}_i^{\alpha} \cap \widetilde{\partial}_{ij}^{\alpha}\right)$ for $j \in [k] \setminus \{i\}$, cf. (107). Finally, in Step 4, we conclude the proof.

  Step 1. We may assume that $\boldsymbol{\theta}_s^* = 0$ without loss of generality, due to the Euclidean invariance; see Remark 1. Since $\boldsymbol{\theta}_s^* \notin \operatorname{int} \mathcal{V}_i$ and $\mathcal{V}$ is a convex set, there exists a $\boldsymbol{v} \in \mathbb{R}^d$ such that $\langle \boldsymbol{v}, \boldsymbol{x} \rangle \geq 0$ for all $\boldsymbol{x} \in \mathcal{V}_j$ by the Separating Hyperplane Theorem. Again by the Euclidean invariance (rotational invariance, in particular), we may assume that $\boldsymbol{v} = \boldsymbol{e}_1$. For each point $\boldsymbol{x} \in \operatorname{int} \mathcal{V}_i$, the ray $\{\boldsymbol{x} - b\boldsymbol{e}_1 \in \mathbb{R}^d : b \geq 0\}$ intersects a facet F of the polyhedron $\mathcal{V}_i$ at a unique point, which we denote by $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$, where $F \subseteq \partial_{ij}$ for some $j \in [k]$ with $\boldsymbol{\beta}_i \neq \boldsymbol{\beta}_j$; we let $j_* = j_*(\boldsymbol{x})$ denote such j (if there exist multiple such j's, we pick the smallest).

  It is clear that $\boldsymbol{y}$ and j_* are independent of the first coordinate of $\boldsymbol{x}$, hence, we can write $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}_2^d)$ and $j_* = j_*(\boldsymbol{x}_2^d)$, where $\boldsymbol{x}_2^d := (x_2, \ldots, x_d) \in \mathbb{R}^{d-1}$. As a result, every $\boldsymbol{x} \in \operatorname{int} \mathcal{V}_i$ can be uniquely expressed as $\boldsymbol{x} = \boldsymbol{y}(\boldsymbol{x}_2^d) + b \cdot \boldsymbol{e}_1$ for some $b = b(\boldsymbol{x}) \geq 0$. Note that for all $\boldsymbol{x} \in \mathcal{V}_i$,

$$x_1 \geq y_1(\boldsymbol{x}_2^d) = \langle \boldsymbol{e}_1, \boldsymbol{y}(\boldsymbol{x}_2^d) \rangle \geq 0$$

by construction of $\boldsymbol{y}$ and by the separating hyperplane theorem. For each $i \in [k]$, we let

$$\mathcal{L}_i := \left\{ j \in [k] \setminus \{i\} : j = j_*(\boldsymbol{x}_2^d) \text{ for some } \boldsymbol{x} \in \operatorname{int} \mathcal{V}_i \right\}.$$

See Figure 5 for an illustration of these notations.

  Step 2. We overload notation and let ϕ_σ denote the probability density function of $\mathcal{N}(0, \sigma^2 \boldsymbol{I}_d)$ for any

$d \in \mathbb{N}_+$. As we assumed $\boldsymbol{\theta}_s^* = 0$, we can write

$$\mathbb{P}_s(\mathcal{V}_i) = \int_{\mathbb{R}^d} \mathbb{1}\left\{\boldsymbol{x} \in \mathcal{V}_i\right\} \cdot \phi_\sigma(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

$$= \sum_{j \in \mathcal{L}_i} \int_{\mathbb{R}^d} \mathbb{1}\left\{\boldsymbol{x} \in \mathcal{V}_i \ \& \ j_*(\boldsymbol{x}_2^d) = j\right\} \cdot \phi_\sigma(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

$$= \sum_{j \in \mathcal{L}_i} \int_{\mathbb{R}^{d-1}} \left[\int_{y_1(\boldsymbol{x}_2^d)}^{\infty} \mathbb{1}\left\{\boldsymbol{x} \in \mathcal{V}_i\right\} \cdot \phi_\sigma(x_1) \, \mathrm{d}x_1\right] \cdot \mathbb{1}\left\{j_*(\boldsymbol{x}_2^d) = j\right\} \cdot \phi_\sigma(\boldsymbol{x}_2^d) \, \mathrm{d}\boldsymbol{x}_2^d \quad (102)$$

by the definition of $j_*(\cdot)$ and $\mathcal{L}_i$, as well as the Fubini's theorem.

Define the quantity $u_i^*(\boldsymbol{x}_2^d) \coloneqq \sup\left\{x_1 \in \mathbb{R} : (x_1, \boldsymbol{x}_2^d)^\top \in \mathcal{V}_i\right\}$ with the convention that $y_1(\boldsymbol{x}_2^d) = u_i^*(\boldsymbol{x}_2^d) = 0$ when $\left\{(x_1, \boldsymbol{x}_2^d)^\top : x_1 \in \mathbb{R}\right\} \cap \mathcal{V}_i = \emptyset$. Then we observe that $0 \le y_1(\boldsymbol{x}_2^d) \le u_i^*(\boldsymbol{x}_2^d) \le \infty$. Continuing from the equation (102), we have

$$\mathbb{P}_s(\mathcal{V}_i) = \sum_{j \in \mathcal{L}_i} \int_{\mathbb{R}^{d-1}} \left[\int_{y_1(\boldsymbol{x}_2^d)}^{u_i^*(\boldsymbol{x}_2^d)} \phi_\sigma(x_1) \, \mathrm{d}x_1\right] \cdot \mathbb{1}\left\{j_*(\boldsymbol{x}_2^d) = j\right\} \cdot \phi_\sigma(\boldsymbol{x}_2^d) \, \mathrm{d}\boldsymbol{x}_2^d. \quad (103)$$

Step 3. For each $j \in \mathcal{L}_i$, we may write using a similar argument as in Step 2 that for any $\alpha \ge 0$,

$$\mathbb{P}_s\left(\widetilde{\mathcal{V}}_i^\alpha \cap \widetilde{\partial}_{ij}^\alpha\right) \overset{(a)}{\ge} \mathbb{P}_s\left(\mathcal{V}_i \cap \widetilde{\partial}_{ij}^\alpha\right)$$

$$\overset{(b)}{=} \sum_{j' \in \mathcal{L}_i} \int_{\mathbb{R}^{d-1}} \left[\int_{y_1(\boldsymbol{x}_2^d)}^{\infty} \mathbb{1}\left\{\boldsymbol{x} \in \mathcal{V}_i \cap \widetilde{\partial}_{ij}^\alpha\right\} \cdot \phi_\sigma(x_1) \, \mathrm{d}x_1\right] \cdot \mathbb{1}\left\{j_*(\boldsymbol{x}_2^d) = j'\right\} \cdot \phi_\sigma(\boldsymbol{x}_2^d) \, \mathrm{d}\boldsymbol{x}_2^d$$

$$= \sum_{j' \in \mathcal{L}_i} \int_{\mathbb{R}^{d-1}} \left[\int_{y_1(\boldsymbol{x}_2^d)}^{u_i^*(\boldsymbol{x}_2^d)} \mathbb{1}\left\{\boldsymbol{x} \in \widetilde{\partial}_{ij}^\alpha\right\} \cdot \phi_\sigma(x_1) \, \mathrm{d}x_1\right] \cdot \mathbb{1}\left\{j_*(\boldsymbol{x}_2^d) = j'\right\} \cdot \phi_\sigma(\boldsymbol{x}_2^d) \, \mathrm{d}\boldsymbol{x}_2^d \quad (104)$$

where the inequality (a) holds because $\mathcal{V}_i \subseteq \widetilde{\mathcal{V}}_i^\alpha$ and (b) is due to (102) and (103).

Next, we observe that for each $j \in \mathcal{L}_i$, the following holds:

$$j_*(\boldsymbol{x}_2^d) = j \ \text{ and } \ x_1 - y_1\left(\boldsymbol{x}_2^d\right) \le \alpha\sigma \qquad \Longrightarrow \qquad \boldsymbol{x} \in \widetilde{\partial}_{ij}^\alpha. \quad (105)$$

Proof of Claim (105): Fix an $\boldsymbol{x} \in \mathcal{V}_i$ with $j_*(\boldsymbol{x}_2^d) = j$ and $x_1 - y_1\left(\boldsymbol{x}_2^d\right) \le \alpha\sigma$. Since $\boldsymbol{y}(\boldsymbol{x}_2^d) \in \partial_{ij}$, we have $\left\langle \boldsymbol{y}(\boldsymbol{x}_2^d) - \bar{\boldsymbol{\beta}}_{ij}, \boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_{ij}\right\rangle = 0$ by definition of ∂_{ij}; see (93). Therefore,

$$\left|\left\langle \boldsymbol{x} - \bar{\boldsymbol{\beta}}_{ij}, \boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_{ij}\right\rangle\right| = \left|\left\langle \boldsymbol{x} - \boldsymbol{y}(\boldsymbol{x}_2^d), \boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_{ij}\right\rangle\right|$$

$$\le \left\|\boldsymbol{x} - \boldsymbol{y}(\boldsymbol{x}_2^d)\right\| \cdot \left\|\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_{ij}\right\|$$

$$= \left|x_1 - y_1(\boldsymbol{x}_2^d)\right| \cdot \left\|\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_{ij}\right\|$$

$$\le \alpha\sigma \cdot \left\|\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}_{ij}\right\|.$$

The above inequality implies that $\boldsymbol{x} \in \widetilde{\partial}_{ij}^\alpha$ by definition of $\widetilde{\partial}_{ij}^\alpha$, cf. (94).

By the logical implication (105), we have the following inequality for all $\alpha \ge 0$:

$$\mathbb{1}\left\{\boldsymbol{x} \in \widetilde{\partial}_{ij}^\alpha\right\} \ge \mathbb{1}\left\{x_1 - y_1\left(\boldsymbol{x}_2^d\right) \le \alpha\sigma\right\} \cdot \mathbb{1}\left\{j_*(\boldsymbol{x}_2^d) = j\right\}. \quad (106)$$

Combining (104) and (106), we obtain

$$\mathbb{P}_s\left(\widetilde{\mathcal{V}}_i^\alpha \cap \widetilde{\partial}_{ij}^\alpha\right) \ge \int_{\mathbb{R}^{d-1}} \left[\int_{y_1(\boldsymbol{x}_2^d)}^{u_i^*(\boldsymbol{x}_2^d)} \mathbb{1}\left\{x_1 - y_1\left(\boldsymbol{x}_2^d\right) \le \alpha\sigma\right\} \cdot \phi_\sigma(x_1) \, \mathrm{d}x_1\right] \cdot \mathbb{1}\left\{j_*(\boldsymbol{x}_2^d) = j\right\} \cdot \phi_\sigma(\boldsymbol{x}_2^d) \, \mathrm{d}\boldsymbol{x}_2^d.$$

$$(107)$$

<u>*Step 4.*</u> By Lemma 10, we observe that for all $\alpha > 0$,

$$\int_{y_1(\boldsymbol{x}_2^d)}^{u_i^*(\boldsymbol{x}_2^d)} \phi_\sigma(x_1) \, dx_1 \le \left(\sqrt{2\pi} + 1\right) \cdot \max\left\{1, \, \frac{1}{\alpha}\right\} \cdot \int_{y_1(\boldsymbol{x}_2^d)}^{u_i^*(\boldsymbol{x}_2^d)} \mathbb{1}\left\{x_1 - y_1\left(\boldsymbol{x}_2^d\right) \le \alpha\sigma\right\} \cdot \phi_\sigma(x_1) \, dx_1.$$

Combining this inequality with (103) and the inequality (107), we have the following inequality hold for any choice of $\alpha_j \in \mathbb{R}_+$ for $j \in [k] \setminus \{i\}$:

$$\frac{1}{\sqrt{2\pi} + 1} \cdot \mathbb{P}_s(\mathcal{V}_i) \le \sum_{j \in \mathcal{L}_i} \max\left\{1, \, \frac{1}{\alpha_j}\right\} \cdot \mathbb{P}_s\left(\widetilde{\mathcal{V}}_i^{\alpha_j} \cap \widetilde{\partial}_{ij}^{\alpha_j}\right)$$

$$\le \sum_{j \in [k] \setminus \{i\}} \max\left\{1, \, \frac{1}{\alpha_j}\right\} \cdot \mathbb{P}_s\left(\widetilde{\mathcal{V}}_i^{\alpha_j} \cap \widetilde{\partial}_{ij}^{\alpha_j}\right).$$

$\square$

A.3   Completing the Proof of Proposition 1

Proof of Proposition 1. We begin by proving the first claim, and then use it to prove the second claim.

<u>*(1) Proof of Claim 1.*</u> We choose an arbitrary sequence $\left(\alpha_j \in \mathbb{R}_+ : j \in [k] \setminus \{i\}\right)$, and let $\delta_{ij}^\alpha := \frac{1}{k^2} \exp\left(-3\alpha_j \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right)$ for all $j \in [k] \setminus \{i\}$. Then we obtain the following upper bound

$$\frac{1}{\sqrt{2\pi} + 1} \cdot \mathbb{P}_s\left(\mathcal{V}_i\right) \overset{(a)}{\le} \sum_{j \in [k] \setminus \{i\}} \max\left\{1, \, \frac{1}{\alpha_j}\right\} \cdot \mathbb{P}_s\left(\widetilde{\mathcal{V}}_i^{\alpha_j} \cap \widetilde{\partial}_{ij}^{\alpha_j}\right)$$

$$= \sum_{j \in [k] \setminus \{i\}} \max\left\{1, \, \frac{1}{\alpha_j}\right\} \cdot \frac{1}{\delta_{ij}^\alpha} \cdot \mathbb{E}_s\left[\delta_{ij}^\alpha \cdot \mathbb{1}\left\{\widetilde{\mathcal{V}}_i^\alpha \cap \widetilde{\partial}_{ij}^\alpha\right\}\right]$$

$$\le \sum_{j \in [k] \setminus \{i\}} \max\left\{1, \, \frac{1}{\alpha_j}\right\} \cdot \frac{1}{\delta_{ij}^\alpha} \cdot \mathbb{E}_s\left[\Psi_i \Psi_j\right]$$

$$= k^2 \sum_{j \in [k] \setminus \{i\}} \max\left\{1, \, \frac{1}{\alpha_j}\right\} \cdot \mathbb{E}_s\left[\Psi_i \Psi_j\right] \cdot \exp\left(3\alpha_j \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right).$$

where (a) is obtained by Lemma 11, and (b) is due to Lemma 9, Claim 3.

<u>*(2) Proof of Claim 2.*</u> Summing $\mathbb{E}_s\left[\Psi_i \Psi_j\right]$ over $j \in [k] \setminus \{i\}$, we obtain that

$$\sum_{j \in [k] \setminus \{i\}} \mathbb{E}_s\left[\Psi_i \Psi_j\right] = \mathbb{E}_s\left[\Psi_i(1 - \Psi_i)\right]$$

$$= \mathbb{E}_s\left[\Psi_i \cdot \mathbb{1}\left\{\mathcal{V}_i\right\}\right] - \mathbb{E}_s\left[\Psi_i^2 \cdot \mathbb{1}\left\{\mathcal{V}_i\right\}\right] + \mathbb{E}_s\left[\Psi_i(1 - \Psi_i) \cdot \mathbb{1}\left\{\mathcal{V}_i^{\complement}\right\}\right]$$

$$\overset{(a)}{\ge} \frac{1}{2}\mathbb{E}_s\left[\Psi_i \cdot \mathbb{1}\left\{\mathcal{V}_i\right\}\right] - \mathbb{P}_s\left(\mathcal{V}_i\right) + \frac{1}{2}\mathbb{E}_s\left[\Psi_i \cdot \mathbb{1}\left\{\mathcal{V}_i^{\complement}\right\}\right]$$

$$= \frac{1}{2}\mathbb{E}_s\left[\Psi_i\right] - \mathbb{P}_s\left(\mathcal{V}_i\right)$$

where the inequality (a) holds because $\boldsymbol{x} \in \mathcal{V}_i^{\complement}$ implies $\psi_i(\boldsymbol{x}) < \frac{1}{2}$. Then it follows that

$$\mathbb{E}_s\left[\Psi_i\right] \le 2\left\{\sum_{j \in [k] \setminus \{i\}} \mathbb{E}_s\left[\Psi_i \Psi_j\right] + \mathbb{P}_s\left(\mathcal{V}_i\right)\right\}$$

$$\le 3\left(\sqrt{2\pi} + 1\right) \cdot k^2 \sum_{j \in [k] \setminus \{i\}} \max\left\{1, \, \frac{1}{\alpha_j}\right\} \cdot \mathbb{E}_s\left[\Psi_i \Psi_j\right] \cdot \exp\left(3\alpha_j \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma}\right).$$

The last inequality follows from Claim 1 of this proposition.

$\square$

B Deferred Proof of Proposition 2

Proof of Proposition 2. In this proof, we prove the three claims one by one.

(1) Proof of Claim 1. Fix an arbitrary $i \in [k]$ that satisfies $\mathcal{A}_i^\delta \neq \emptyset$. Let $\mathcal{S}_i := \{j \in [k] : \boldsymbol{\beta}_j = \boldsymbol{\beta}_i\}$. Then we may write

$$\mathbb{E}_s\left[\Psi_i\right] = \frac{1}{|\mathcal{S}_i|} \sum_{j \in \mathcal{S}_i} \mathbb{E}_s\left[\Psi_j\right] = \frac{1}{|\mathcal{S}_i|}\left(1 - \sum_{j \in [k] \setminus \mathcal{S}_i} \mathbb{E}_s\left[\Psi_j\right]\right). \tag{108}$$

By definition of the set $\mathcal{A}_i^\delta$, we observe that for all pair $(s, j) \in \mathcal{A}_i^\delta \times ([k] \setminus \mathcal{S}_i)$,

$$\boldsymbol{\theta}_s^* \notin \operatorname{int} \mathcal{V}_j \quad \text{and} \quad \max_{j' \in [k] \setminus \{j\}} \frac{\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'}\|}{\sigma} \cdot \mathbb{E}_s\left[\Psi_j \Psi_{j'}\right] < \delta. \tag{109}$$

It follows from Corollary 6 (claim 2) and (109) that for all $(s, j) \in \mathcal{A}_i^\delta \times ([k] \setminus \mathcal{S}_i)$,

$$\begin{aligned}
\mathbb{E}_s\left[\Psi_j\right] &\leq 9\left(\sqrt{2\pi} + 1\right) \cdot k^2 \cdot \sum_{j' \in [k] \setminus \{j\}} \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \cdot \mathbb{E}_s\left[\Psi_i \Psi_j\right] \\
&\leq 9\left(\sqrt{2\pi} + 1\right) \cdot k^2(k-1) \cdot \delta \\
&\leq 9\left(\sqrt{2\pi} + 1\right) \cdot k^3 \cdot \delta.
\end{aligned} \tag{110}$$

Plugging (110) into (108), we obtain that for all $s \in \mathcal{A}_i^\delta$,

$$\mathbb{E}_s\left[\Psi_i\right] \geq \frac{1}{|\mathcal{S}_i|}\left\{1 - 9\left(\sqrt{2\pi} + 1\right) \cdot k^4 \cdot \delta\right\}. \tag{111}$$

Now we assume there exists $i' \in \mathcal{S}_i$ such that $i' \neq i$. Then we have

$$\begin{aligned}
\mathbb{E}_s[\Psi_i \Psi_{i'}] = \mathbb{E}_s[\Psi_i^2] &\geq \mathbb{E}_s[\Psi_i]^2 \\
&\geq \frac{1}{|\mathcal{S}_i|^2}\left\{1 - 9\left(\sqrt{2\pi} + 1\right) \cdot k^4 \cdot \delta\right\}^2 \\
&\geq \frac{1}{k^2}\left\{1 - 9\left(\sqrt{2\pi} + 1\right) \cdot k^4 \cdot \delta\right\}^2.
\end{aligned} \tag{112}$$

Letting $\varphi_* := 9\left(\sqrt{2\pi} + 1\right)$ for a shorthand, we observe that if $\delta \leq \frac{1}{2\varphi_* k^4}$, then it follows from (112) that

$$\mathbb{E}_s[\Psi_i \Psi_{i'}] \geq \frac{1}{k^2}\left\{1 - \varphi_* \cdot k^4 \cdot \delta\right\}^2 \geq \frac{3}{4}\frac{1}{k^2} \geq \frac{1}{2\varphi_* k^4} \geq \delta,$$

which contradicts the assumption that $s \in \mathcal{A}_i^\delta$. Consequently, $\mathcal{S}_i = \{i\}$, and the claim is proved.

(2) Proof of Claim 2. Now that $\mathcal{S}_i = \{i\}$, it immediately follows from (111) that for all $s \in \mathcal{A}_i^\delta$,

$$1 - \mathbb{E}_s\left[\Psi_i\right] \leq 9\left(\sqrt{2\pi} + 1\right) \cdot k^4 \cdot \delta.$$

(3) Proof of Claim 3. It follows from Corollary 6 (claim 1) and (109) that for all $s \in \mathcal{A}_i^\delta$ and all $j \in [k]\setminus\{i\}$ (note that $\mathcal{S}_i = \{i\}$ by claim 1 proved above),

$$\mathbb{P}_s(\mathcal{V}_j) \leq 3\left(\sqrt{2\pi} + 1\right) \cdot k^4 \cdot \delta$$

by the same argument as in (110). Therefore, for all $s \in \mathcal{A}_i^\delta$,

$$\mathbb{P}_s\left(\mathcal{V}_i^c\right) = \sum_{j \in [k]\setminus\{i\}} \mathbb{P}_s\left(\mathcal{V}_j\right) \leq 3\left(\sqrt{2\pi} + 1\right) \cdot k^4 \cdot \delta.$$

$\square$

C Deferred Proof of Proposition 3

C.1 A Helper lemma for the Proof of Proposition 3

Recall from Section 3.1.2 that any local minimum $\boldsymbol{B}$ of L must satisfy the first-order stationary condition (12) and the second-order optimality condition $\nabla^2 L(\boldsymbol{\beta}) \succeq 0$, where the Hessian $\nabla^2 L$ can be computed using the expression (13) as stated in Lemma 2. Based on this fact, we state a lemma that is useful in our proof of Proposition 3.

Lemma 12. *Let $\boldsymbol{B} \in \mathbb{R}^{d \times k}$ be a local minimum of L. For any $i, j \in [k]$ such that $i \neq j$, and any $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^d$,*

$$\mathbb{E}_* \left[\Psi_i \Psi_j \left(\langle \boldsymbol{\beta}_i - \mathsf{x}, \boldsymbol{v}_1 \rangle - \langle \boldsymbol{\beta}_j - \mathsf{x}, \boldsymbol{v}_2 \rangle \right)^2 \right] \leq \sigma^2 \cdot \|\boldsymbol{v}_1\|^2 \cdot \mathbb{E}_* [\Psi_i] + \sigma^2 \cdot \|\boldsymbol{v}_2\|^2 \cdot \mathbb{E}_* [\Psi_j]. \tag{113}$$

Proof of Lemma 12. First of all, $\nabla^2 L(\boldsymbol{B}) \succeq 0$ because $\boldsymbol{B}$ is a local minimum of L. Thus, $\boldsymbol{v}^\top \nabla^2 L(\boldsymbol{B})\boldsymbol{v} \geq 0$ for all $\boldsymbol{v} \in \mathbb{R}^{dk}$. We choose $\boldsymbol{v}$ to be the flattened vector of $\boldsymbol{V} \in \mathbb{R}^{d \times k}$ such that $\boldsymbol{V}_i = \boldsymbol{v}_1$ and $\boldsymbol{V}_j = \boldsymbol{v}_2$, all other columns being 0. Then we obtain by applying Lemma 2 that

$$\sigma^4 \cdot \boldsymbol{v}^\top \nabla^2 L(\boldsymbol{B})\boldsymbol{v} = \sigma^4 \cdot \boldsymbol{v}_i^\top \left[\frac{\partial^2}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_i} L(\boldsymbol{B}) \right] \boldsymbol{v}_i + \sigma^4 \cdot \boldsymbol{v}_j^\top \left[\frac{\partial^2}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j} L(\boldsymbol{B}) \right] \boldsymbol{v}_j + 2\sigma^4 \cdot \boldsymbol{v}_i^\top \left[\frac{\partial^2}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_j} L(\boldsymbol{B}) \right] \boldsymbol{v}_j$$

$$= \mathbb{E}_* \left[(\Psi_i - 1)\Psi_i \langle \boldsymbol{\beta}_i - \mathsf{x}, \boldsymbol{v}_1 \rangle^2 + \sigma^2 \cdot \Psi_i \|\boldsymbol{v}_1\|^2 \right] + \mathbb{E}_* \left[(\Psi_j - 1)\Psi_j \langle \boldsymbol{\beta}_j - \mathsf{x}, \boldsymbol{v}_2 \rangle^2 + \sigma^2 \cdot \Psi_j \|\boldsymbol{v}_2\|^2 \right]$$

$$+ 2\mathbb{E}_* \left[\Psi_i \Psi_j \langle \boldsymbol{\beta}_i - \mathsf{x}, \boldsymbol{v}_1 \rangle \langle \boldsymbol{\beta}_j - \mathsf{x}, \boldsymbol{v}_2 \rangle \right]. \tag{114}$$

Noticing that $\sum_{i' \in [k]} \Psi_{i'} = 1$ and $\Psi_{i'} \geq 0$ for all $i' \in [k]$, we observe that $\Psi_i + \Psi_j \leq 1$. Then it follows from (114) that

$$\sigma^4 \cdot \boldsymbol{v}^\top \nabla^2 L(\boldsymbol{B})\boldsymbol{v} \leq -\mathbb{E}_* \left[\Psi_i \Psi_j \cdot \left\{ \langle \boldsymbol{\beta}_i - \mathsf{x}, \boldsymbol{v}_1 \rangle^2 + \langle \boldsymbol{\beta}_j - \mathsf{x}, \boldsymbol{v}_2 \rangle^2 \right\} \right] + 2\mathbb{E}_* \left[\Psi_i \Psi_j \cdot \langle \boldsymbol{\beta}_i - \mathsf{x}, \boldsymbol{v}_1 \rangle \langle \boldsymbol{\beta}_j - \mathsf{x}, \boldsymbol{v}_2 \rangle \right\} \right]$$

$$+ \sigma^2 \cdot \mathbb{E}_* \left[\Psi_i \|\boldsymbol{v}_1\|^2 + \Psi_j \|\boldsymbol{v}_2\|^2 \right]$$

$$= -\mathbb{E}_* \left[\Psi_i \Psi_j \cdot \left(\langle \boldsymbol{\beta}_i - \mathsf{x}, \boldsymbol{v}_1 \rangle - \langle \boldsymbol{\beta}_j - \mathsf{x}, \boldsymbol{v}_2 \rangle \right)^2 \right] + \sigma^2 \cdot \|\boldsymbol{v}_1\|^2 \cdot \mathbb{E}_* [\Psi_i] + \sigma^2 \cdot \|\boldsymbol{v}_2\|^2 \cdot \mathbb{E}_* [\Psi_j].$$

To complete the proof, it suffices to recall $\boldsymbol{v}^\top \nabla^2 L(\boldsymbol{B})\boldsymbol{v} \geq 0$. $\qquad \square$

C.2 Completing the Proof of Proposition 3

Proof of Proposition 3. In this proof, we prove the two claims separately.

(1) Proof of Claim 1. Fix arbitrary $i, j \in [k]$ such that $i \neq j$ and let $\boldsymbol{v}_{ij} := \frac{\boldsymbol{\beta}_i - \boldsymbol{\beta}_j}{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}$. Applying Lemma 12 with $\boldsymbol{v}_1 = \boldsymbol{v}_2 = \boldsymbol{v}_{ij}$, we obtain

$$\sigma^2 \cdot \mathbb{E}_* [\Psi_i + \Psi_j] \geq \mathbb{E}_* \left[\Psi_i \Psi_j \langle \boldsymbol{\beta}_i - \boldsymbol{\beta}_j, \boldsymbol{v}_{ij} \rangle^2 \right]$$

$$= \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|^2 \cdot \mathbb{E}_* [\Psi_i \Psi_j].$$

Rearranging the terms, we obtain the desired inequality.

(2) Proof of Claim 2. Choose an arbitrary pair $(s, i) \in [k_*] \times [k]$ and let $\boldsymbol{u}_{s \to i} := \frac{\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*}{\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\|}$. First, we observe that for any $\boldsymbol{x} \in \mathbb{R}^d$,

$$\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\|^2 = \langle \boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*, \boldsymbol{u}_{s \to i} \rangle^2$$

$$= \left(\langle \boldsymbol{\beta}_i - \boldsymbol{x}, \boldsymbol{u}_{s \to i} \rangle + \langle \boldsymbol{x} - \boldsymbol{\theta}_s^*, \boldsymbol{u}_{s \to i} \rangle \right)^2$$

$$\leq 2 \langle \boldsymbol{\beta}_i - \boldsymbol{x}, \boldsymbol{u}_{s \to i} \rangle^2 + 2 \langle \boldsymbol{x} - \boldsymbol{\theta}_s^*, \boldsymbol{u}_{s \to i} \rangle^2. \tag{115}$$

Second, applying Lemma 12 with $\boldsymbol{v}_1 = \boldsymbol{u}_{s\to i}$ and $\boldsymbol{v}_2 = 0$, we get for all $j \in [k] \setminus \{i\}$,

$$\sigma^2 \cdot \mathbb{E}_* [\Psi_i] \geq \mathbb{E}_* \left[\Psi_i \Psi_j \langle \boldsymbol{\beta}_i - \mathsf{x}, \boldsymbol{u}_{s\to i} \rangle^2 \right]. \tag{116}$$

Then it follows that for all $j \in [k] \setminus \{i\}$,

$$\sigma^2 \cdot \mathbb{E}_* [\Psi_i] \overset{(a)}{\geq} \mathbb{E}_* \left[\Psi_i \Psi_j \langle \boldsymbol{\beta}_i - \mathsf{x}, \boldsymbol{u}_{s\to i} \rangle^2 \right]$$

$$\overset{(b)}{=} \frac{1}{k_*} \sum_{s\in[k_*]} \mathbb{E}_s \left[\Psi_i \Psi_j \langle \boldsymbol{\beta}_i - \mathsf{x}, \boldsymbol{u}_{s\to i} \rangle^2 \right]$$

$$\overset{(c)}{\geq} \frac{1}{k_*} \sum_{s\in[k_*]} \left\{ \frac{1}{2} \mathbb{E}_s \left[\Psi_i \Psi_j \cdot \|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\|^2 \right] - \mathbb{E}_s \left[\Psi_i \Psi_j \cdot \langle \mathsf{x} - \boldsymbol{\theta}_s^*, \boldsymbol{u}_{s\to i} \rangle^2 \right] \right\} \tag{117}$$

where (a) follows from (116); (b) is by definition; see (2); and (c) is due to (115).

By our model assumption (Gaussian mixture model), the random variable $\mathsf{z}_i := \langle \mathsf{x} - \boldsymbol{\theta}_s^*, \boldsymbol{u}_{s\to i} \rangle$ is a centered Gaussian random variable with variance σ^2. Therefore, for all $(s, i, j) \in [k_*] \times [k] \times [k]$,

$$\mathbb{E}_s \left[\Psi_i \Psi_j \cdot \mathsf{z}_i^2 \right] \leq \mathbb{E}_s \left[\mathsf{z}_i^2 \right] = \sigma^2. \tag{118}$$

Plugging (118) into (117), we get for all $j \in [k] \setminus \{i\}$,

$$\sigma^2 \cdot \mathbb{E}_* [\Psi_i] \geq \frac{1}{k_*} \sum_{s\in[k_*]} \left(\frac{1}{2} \mathbb{E}_s [\Psi_i \Psi_j] \cdot \|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\|^2 - \sigma^2 \right) \tag{119}$$

$$\geq \frac{1}{2} \mathbb{E}_* [\Psi_i \Psi_j] \cdot \min_{s\in[k_*]} \|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\|^2 - \sigma^2.$$

Rearranging the terms, we complete the proof of Claim 2.

(3) Proof of Claim 3. Continuing on the inequality (119), we observe that

$$\sigma^2 \cdot \mathbb{E}_* [\Psi_i] \geq \frac{1}{k_*} \sum_{s\in[k_*]} \left(\frac{1}{2} \mathbb{E}_s [\Psi_i \Psi_j] \cdot \|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\|^2 - \sigma^2 \right)$$

$$\geq \frac{1}{2k_*} \mathbb{E}_s [\Psi_i \Psi_j] \cdot \|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\|^2 - \sigma^2$$

for all $s \in [k_*]$ and all $j \in [k] \setminus \{i\}$ because $\mathbb{E}_s [\Psi_i \Psi_j] \cdot \|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\|^2 \geq 0$ for all $(s, i, j) \in [k_*] \times [k] \times [k]$. $\qquad \square$

D   Deferred Proof of Proposition 4

D.1   Helper Lemmas

Basic properties   Recall the definition of the Gaussian Q-function $Q(t) = \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}s^2} ds$, and let $Q^{-1} : (0, 1) \to \mathbb{R}$ denote its inverse. That is, we write $z = Q^{-1}(t)$ if $t = Q(z)$.

Lemma 13. *Let $\mathsf{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ be the d-variate standard Gaussian random variable, and $\psi : \mathbb{R}^d \to [0, 1]$. Then*

$$\left\| \mathbb{E}_{\mathsf{x}} [\psi(\mathsf{x}) \cdot \mathsf{x}] \right\| \leq \phi(z_\psi)$$

where $z_\psi = Q^{-1} \big(\mathbb{E}[\psi(\mathsf{x})] \big)$.

Lemma 13 implies an upper bound on the mean displacement of a Gaussian distribution $\mathcal{N}(0, \boldsymbol{I}_d)$ when reweighted using $\frac{\psi(\mathsf{x})}{\mathbb{E}_x[\psi(\mathsf{x})]}$:

$$\left\| \mathbb{E}_{\mathsf{x}} \left[\frac{\psi(\mathsf{x})}{\mathbb{E}_{\mathsf{x}} [\psi(\mathsf{x})]} \cdot \mathsf{x} \right] \right\| \leq \frac{\phi(z_\psi)}{\mathbb{E}_{\mathsf{x}} [\psi(\mathsf{x})]}.$$

Proof of Lemma 13. It suffices to observe that

$$
\left\| \mathbb{E}_{\mathsf{x}}\big[\psi(\mathsf{x}) \cdot \mathsf{x} \big] \right\| = \sup_{u \in \mathbb{R}^d : \|u\| = 1} \mathbb{E}_{\mathsf{x}}\left[\psi(\mathsf{x}) \cdot \langle u, \mathsf{x} \rangle \right]
$$

$$
\leq \int_{z_\psi}^{\infty} t \cdot \phi(t) \ dt = \phi(z_\psi).
$$

$\square$

Lemma 14. *Let* $\boldsymbol{w}, \boldsymbol{w}' \in \{ \tilde{\boldsymbol{w}} \in \mathbb{R}^{k_*} : \tilde{w}_s \geq 0, \forall s \in [k_*], \sum_{s=1}^{k_*} \tilde{w}_s = 1 \}$, *and* $\{ \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k_*} \}, \{ \boldsymbol{v}'_1, \ldots, \boldsymbol{v}'_{k_*} \} \subset \mathbb{R}^d$. *Then*

$$
\left\| \sum_{s=1}^{k_*} w_s \boldsymbol{v}_s - \sum_{s=1}^{k_*} w'_s \boldsymbol{v}'_s \right\| \leq \|\boldsymbol{w} - \boldsymbol{w}'\|_1 \cdot \min_{s \in [k_*]} \max_{s' \in [k_*]} \|\boldsymbol{v}_s - \boldsymbol{v}_{s'}\| + \sum_{s=1}^{k_*} \|w'_s (\boldsymbol{v}_s - \boldsymbol{v}'_s)\|.
$$

Proof of Lemma 14. Applying triangle inequality, we obtain

$$
\left\| \sum_{s=1}^{k_*} w_s \boldsymbol{v}_s - \sum_{s=1}^{k_*} w'_s \boldsymbol{v}'_s \right\| \leq \left\| \sum_{s=1}^{k_*} (w_s - w'_s) \boldsymbol{v}_s \right\| + \left\| \sum_{s=1}^{k_*} w'_s (\boldsymbol{v}_s - \boldsymbol{v}'_s) \right\|
$$

$$
= \left\| \sum_{s=2}^{k_*} (w_s - w'_s)(\boldsymbol{v}_s - \boldsymbol{v}_1) \right\| + \left\| \sum_{s=1}^{k_*} w'_s (\boldsymbol{v}_s - \boldsymbol{v}'_s) \right\|
$$

$$
\leq \left(\sum_{s=2}^{k_*} |w_s - w'_s| \right) \cdot \max_{s \in [k_*]} \|\boldsymbol{v}_s - \boldsymbol{v}_1\| + \sum_{s=1}^{k_*} \|w'_s (\boldsymbol{v}_s - \boldsymbol{v}'_s)\|
$$

$$
\leq \|\boldsymbol{w} - \boldsymbol{w}'\|_1 \cdot \max_{s \in [k_*]} \|\boldsymbol{v}_s - \boldsymbol{v}_1\| + \sum_{s=1}^{k_*} \|w'_s (\boldsymbol{v}_s - \boldsymbol{v}'_s)\|.
$$

Since the choice of $\boldsymbol{v}_1$ is arbitrary, we can replace it with any $\boldsymbol{v}_{s'}$, $s' \in [k_*]$. $\square$

Properties of the coefficients of association Next, we present two useful properties of the coefficients of association, which we use later in the proof of Proposition 4.

Recall the definition of set $\mathcal{A}_i^\delta$ from (35): for any $i \in [k]$ and for any $\delta \geq 0$,

$$
\mathcal{A}_i^\delta := \left\{ s \in [k_*] \ \middle| \ \iota_{\boldsymbol{B}}(\boldsymbol{\theta}_s^*) = i \quad \text{and} \quad \max_{j, j' \in [k]} \frac{\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'}\|}{\sigma} \cdot \mathbb{E}_s \left[\Psi_j \Psi_{j'} \right] < \delta \right\}.
$$

We additionally define $\mathcal{B}_i^\delta$ as follows: for any $i \in [k]$ and for any $\delta \geq 0$,

$$
\mathcal{B}_i^\delta := \left\{ s \in [k_*] \ \middle| \ \iota_{\boldsymbol{B}}(\boldsymbol{\theta}_s^*) \neq i \quad \text{and} \quad \max_{j \in [k] \setminus \{i\}} \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \cdot \mathbb{E}_s \left[\Psi_i \Psi_j \right] < \delta \right\}. \tag{120}
$$

Lemma 15. *Let* $i \in [k]$ *and* $\delta > 0$. *If* $\mathcal{A}_i^\delta \neq \emptyset$ *and* $\delta > \frac{4 k_* \cdot \sigma}{\Delta_{\min}}$, *then* $\left| \mathcal{A}_i^\delta \cup \mathcal{B}_i^\delta \right| \geq k_* - 1$.

Proof of Lemma 15. Observe that if $s \in [k_*] \setminus \left(\mathcal{A}_i^\delta \cup \mathcal{B}_i^\delta \right)$, then either of the following must be true by definition of the sets $\mathcal{A}_i^\delta$ and $\mathcal{B}_i^\delta$:

(a) $\iota_{\boldsymbol{B}}(\boldsymbol{\theta}_s^*) = i$ and $\max_{j \in [k]} \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \cdot \mathbb{E}_s \left[\Psi_i \Psi_j \right] \geq \delta$; or

(b) $\iota_{\boldsymbol{B}}(\boldsymbol{\theta}_s^*) \neq i$ and $\max_{j \in [k]} \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \cdot \mathbb{E}_s \left[\Psi_i \Psi_j \right] \geq \delta$.

In either cases, there must exists $j \in [k] \setminus \{i\}$ such that $\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \cdot \mathbb{E}_s \left[\Psi_i \Psi_j \right] \geq \delta$. That is, $i, j \in \mathcal{E}_s^\delta$ by definition of the set $\mathcal{E}_s^\delta$ presented in (36). By Corollary 7,

$$
\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_s^*\|}{\sigma} \leq \frac{2 k_*}{\delta}. \tag{121}
$$

Now we assume that $\left|\mathcal{A}_i^\delta \cup \mathcal{B}_i^\delta\right| < k_* - 1$, i.e., $\left|[k_*] \setminus \left(\mathcal{A}_i^\delta \cup \mathcal{B}_i^\delta\right)\right| \geq 2$. We choose $s_1, s_2 \in [k_*] \setminus \left(\mathcal{A}_i^\delta \cup \mathcal{B}_i^\delta\right)$ such that $s_1 \neq s_2$. Observe that $i \in \mathcal{E}_{s_1}^\delta \cap \mathcal{E}_{s_2}^\delta$, and therefore, it follows from (121) that

$$\|\boldsymbol{\theta}_{s_1}^* - \boldsymbol{\theta}_{s_2}^*\| \leq \|\boldsymbol{\beta}_{l_1} - \boldsymbol{\theta}_{s_1}^*\| + \|\boldsymbol{\beta}_{l_1} - \boldsymbol{\theta}_{s_2}^*\|$$
$$\leq \frac{4k_* \cdot \sigma}{\delta}.$$

If $\delta > \frac{4k_* \cdot \sigma}{\Delta_{\min}}$, then $\frac{4k_* \cdot \sigma}{\delta} < \Delta_{\min}$, contradicting the definition $\Delta_{\min} := \min_{\substack{s,s' \in [k_*] \\ s \neq s}} \|\boldsymbol{\theta}_s^* - \boldsymbol{\theta}_{s'}^*\|$. Therefore we conclude that $\left|\mathcal{A}_i^\delta \cup \mathcal{B}_i^\delta\right| \geq k_* - 1$. $\qquad\square$

Lemma 16. *Let $i \in [k]$ and $\delta > 0$. If $\mathcal{A}_i^\delta \neq \emptyset$, then*

$$\max_{s \in \mathcal{B}_i^\delta} \left\|\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_i\right\| \cdot \mathbb{E}_s\left[\Psi_i\right] \leq 45\left(\sqrt{2\pi} + 1\right) \cdot k^3 \cdot \delta \cdot D_{\mathrm{cell}}(\boldsymbol{B}) + 4\sigma$$

where $D_{\mathrm{cell}}(\boldsymbol{B}) := \max_{j \in [k]} \max_{s \in [k_*] : \boldsymbol{\theta}_s^* \in \mathcal{V}_j} \|\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_j\|$.

Proof of Lemma 16. Choose an arbitrary $s \in \mathcal{B}_i^\delta$ and let $j = \iota_{\boldsymbol{B}}(\boldsymbol{\theta}_s^*)$. Notice that $j \neq i$, and we recall from (92) that $\bar{\boldsymbol{\beta}}_{ij} := \frac{\boldsymbol{\beta}_i + \boldsymbol{\beta}_j}{2}$. Then we define three quantities:

$$d_{ij} := \left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right\|,$$
$$t := \left\langle \boldsymbol{\theta}_s^* - \bar{\boldsymbol{\beta}}_{ij}, \frac{\boldsymbol{\beta}_j - \boldsymbol{\beta}_i}{\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_i\|} \right\rangle,$$
$$n := \sqrt{\|\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_j\|^2 - \left\langle \boldsymbol{\theta}_s^* - \boldsymbol{\beta}_j, \frac{\boldsymbol{\beta}_j - \boldsymbol{\beta}_i}{\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_i\|} \right\rangle}.$$

In what follows, we prove the lemma by considering two cases: (1) $d_{ij} \leq D_{\mathrm{cell}}(\boldsymbol{B})$ or $\left|t - \frac{d_{ij}}{2}\right| \geq \frac{d_{ij}}{4}$ or $n \geq \frac{d_{ij}}{4}$; and (2) $d_{ij} > D_{\mathrm{cell}}(\boldsymbol{B})$ and $\left|t - \frac{d_{ij}}{2}\right| < \frac{d_{ij}}{4}$ and $n < \frac{d_{ij}}{4}$.

Case 1. We observe that

$$\frac{1}{9\left(\sqrt{2\pi} + 1\right)} \mathbb{E}_s\left[\Psi_i\right] \overset{(a)}{\leq} k^2 \sum_{j \in [k] \setminus \{i\}} \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \cdot \mathbb{E}_s\left[\Psi_i \Psi_j\right]$$
$$\overset{(b)}{\leq} k^2 \sum_{j \in [k] \setminus \{i\}} \delta$$
$$\leq k^3 \cdot \delta. \tag{122}$$

where (a) is due to Corollary 6, Claim 2, and (b) is by definition of $\mathcal{B}_i^\delta$, cf. (120). Thereafter, we consider the three subcases individually.

Case 1-A. First of all, suppose that $d_{ij} \leq D_{\mathrm{cell}}(\boldsymbol{B})$. Observe that

$$\|\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_i\| \leq \|\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_j\| + \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_i\|$$
$$\leq D_{\mathrm{cell}}(\boldsymbol{B}) + d_{ij}$$
$$\leq 2D_{\mathrm{cell}}(\boldsymbol{B}).$$

This, combined with (122), yields

$$\left\|\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_i\right\| \cdot \mathbb{E}_s\left[\Psi_i\right] \leq 18\left(\sqrt{2\pi} + 1\right) \cdot k^3 \cdot \delta \cdot D_{\mathrm{cell}}(\boldsymbol{B}). \tag{123}$$

Case 1-B. Next, suppose that $\left|t - \frac{d_{ij}}{2}\right| \geq \frac{d_{ij}}{4}$. It is easy to observe that

$$\|\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_i\| \leq 5\|\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_j\| \leq 5D_{\mathrm{cell}}(\boldsymbol{B}).$$

Likewise, it follows from (122) that

$$\left\|\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_i\right\| \cdot \mathbb{E}_s\left[\Psi_i\right] \le 45\left(\sqrt{2\pi} + 1\right) \cdot k^3 \cdot \delta \cdot D_{\text{cell}}(\boldsymbol{B}). \tag{124}$$

Case 1-C. Lastly, suppose that $n \ge \frac{d_{ij}}{4}$. Similarly, we observe that

$$\|\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_i\| \le \left(\frac{(\sqrt{5}+2)^2 + 1}{(\sqrt{5}-2)^2 + 1}\right)^{1/2} \|\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_j\|$$

$$\le 5 D_{\text{cell}}(\boldsymbol{B}),$$

and therefore,

$$\left\|\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_i\right\| \cdot \mathbb{E}_s\left[\Psi_i\right] \le 45\left(\sqrt{2\pi} + 1\right) \cdot k^3 \cdot \delta \cdot D_{\text{cell}}(\boldsymbol{B}). \tag{125}$$

Case 2. Now we suppose that (i) $d_{ij} > D_{\text{cell}}(\boldsymbol{B})$, (ii) $\left|t - \frac{d_{ij}}{2}\right| < \frac{d_{ij}}{4}$, and (iii) $n < \frac{d_{ij}}{4}$. Observe that

$$\|\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_i\|^2 = \left(t + \frac{d_{ij}}{2}\right)^2 + n^2 < \frac{13}{8} d_{ij}^2. \tag{126}$$

Next, we fix a coordinate system: without loss of generality, we may let $\boldsymbol{\theta}_s^* = 0$ and let $\boldsymbol{e}_1 = \frac{\boldsymbol{\beta}_i - \boldsymbol{\beta}_j}{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}$; see Figure 6. Writing $\boldsymbol{x} = (x_1, x_2^d)$, we observe that for any $\alpha \in \left(0, \frac{1}{2}\right)$,

$$\mathbb{E}_s\left[\Psi_i\right] = \mathbb{E}_s\left[\Psi_i \cdot \mathbb{1}\left\{\mathsf{x}_1 \le \alpha \cdot d_{ij}\right\}\right] + \mathbb{E}_s\left[\Psi_i \cdot \mathbb{1}\left\{\mathsf{x}_1 > \alpha \cdot d_{ij}\right\}\right].$$

It suffices to establish upper bounds for the two terms on the right hand side.

- If $x_1 \le \alpha \cdot d_{ij}$, then

$$\psi_i(\boldsymbol{x}) = \frac{e^{-\frac{\|\boldsymbol{x}-\boldsymbol{\beta}_i\|^2}{2\sigma^2}}}{\sum_{j'=1}^k e^{-\frac{\|\boldsymbol{x}-\boldsymbol{\beta}_{j'}\|^2}{2\sigma^2}}}$$

$$\le \frac{e^{-\frac{\|\boldsymbol{x}-\boldsymbol{\beta}_i\|^2}{2\sigma^2}}}{e^{-\frac{\|\boldsymbol{x}-\boldsymbol{\beta}_j\|^2}{2\sigma^2}}}$$

$$= e^{\frac{1}{\sigma^2}\langle\boldsymbol{\beta}_i-\boldsymbol{\beta}_j, \boldsymbol{x}-\bar{\boldsymbol{\beta}}_{ij}\rangle}$$

$$\le e^{-\left(\frac{1}{2}-\alpha\right)\frac{d_{ij}^2}{\sigma^2}}.$$

Thus, $\mathbb{E}_s\left[\Psi_i \cdot \mathbb{1}\left\{\mathsf{x}_1 \le \alpha \cdot d_{ij}\right\}\right] \le e^{-\left(\frac{1}{2}-\alpha\right)\frac{d_{ij}^2}{\sigma^2}}$.

- Since $\psi_i(\boldsymbol{x}) \le 1$, we have

$$\mathbb{E}_s\left[\Psi_i \cdot \mathbb{1}\left\{\mathsf{x}_1 > \alpha \cdot d_{ij}\right\}\right] \le \mathbb{E}_s\left[\mathbb{1}\left\{\mathsf{x}_1 > \alpha \cdot d_{ij}\right\}\right]$$

$$= \mathbb{P}_s\left(\mathsf{x}_1 > \alpha \cdot d_{ij}\right)$$

$$= Q\left(\frac{\alpha \cdot d_{ij}}{\sigma}\right)$$

$$\overset{(a)}{\le} \sqrt{2\pi}\phi\left(\frac{\alpha \cdot d_{ij}}{\sigma}\right)$$

$$= e^{-\frac{\alpha^2}{2}\frac{d_{ij}^2}{\sigma^2}},$$
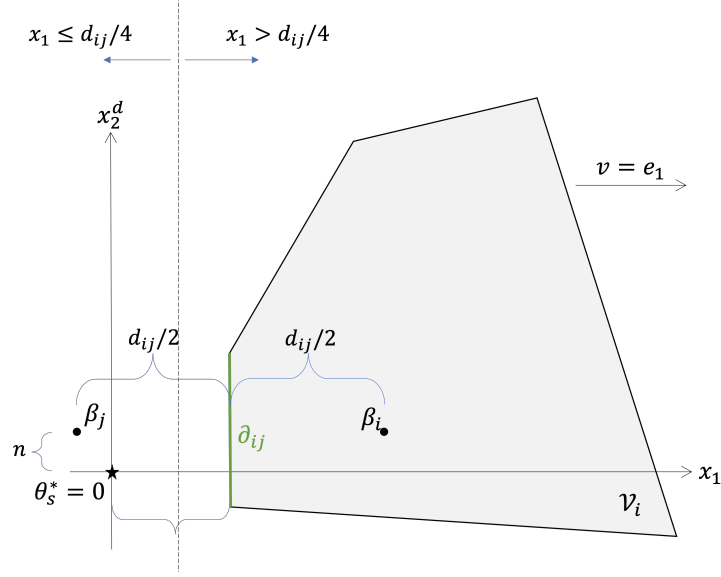
where (a) follows from Lemma 4.

Figure 6: Illustration for the proof of Lemma 16, specifically Case 2.

We optimize the value of $\alpha \in (0, 1/2)$ to balance these two upper bounds by solving $\frac{1}{2} - \alpha = \frac{\alpha^2}{2}$, and obtain $\alpha^\star = \sqrt{2} - 1$. Therefore,

$$
\begin{aligned}
\mathbb{E}_s\big[\Psi_i\big] &= \mathbb{E}_s\left[\Psi_i \cdot \mathbb{1}\Big\{\mathsf{x}_1 \le \alpha^\star \cdot d_{ij}\Big\}\right] + \mathbb{E}_s\left[\Psi_i \cdot \mathbb{1}\Big\{\mathsf{x}_1 > \alpha^\star \cdot d_{ij}\Big\}\right] \\
&\le 2e^{-\left(\frac{3}{2} - \sqrt{2}\right)\frac{d_{ij}^2}{\sigma^2}}.
\end{aligned}
\tag{127}
$$

Combining (126) and (127), we obtain

$$
\big\|\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_i\big\| \cdot \mathbb{E}_s\big[\Psi_i\big] \le \sigma \cdot \sqrt{\frac{13}{2}} \frac{d_{ij}}{\sigma} \cdot e^{-\left(\frac{3}{2} - \sqrt{2}\right)\frac{d_{ij}^2}{\sigma^2}} \le 4\sigma,
$$

where the last inequality follows from observing $\max_{z \ge 0}\left\{\sqrt{13/2} \cdot z \cdot e^{-\left(\frac{3}{2} - \sqrt{2}\right)z^2}\right\} < 4$.

$\square$

D.2   Proof of Proposition 4

Proof of Proposition 4. We present a proof of Proposition 4 in three steps.

Step 1. Preliminary observations   Fix $\delta > 0$ and choose an $i \in [k]$ such that $\mathcal{A}_i^\delta \ne \emptyset$. Observe that

$$
\begin{aligned}
\boldsymbol{\beta}_i &= \frac{\mathbb{E}_*[\Psi_i \mathsf{x}]}{\mathbb{E}_*[\Psi_i]} \\
&= \sum_{s \in [k_*]} \underbrace{\frac{\mathbb{E}_s[\Psi_i]}{\sum_{s \in [k_*]} \mathbb{E}_s[\Psi_i]}}_{=: w_s} \cdot \underbrace{\mathbb{E}_s\left[\frac{\Psi_i}{\mathbb{E}_s[\Psi_i]} \cdot \mathsf{x}\right]}_{=: \boldsymbol{v}_s}.
\end{aligned}
$$

by Lemma 1, and more specifically, by (12). We let for each $s \in [k_*]$,

$$
\widehat{w}_s := \frac{\mathbb{1}\{s \in \mathcal{A}_i^\delta\}}{\sum_{s \in [k_*]} \mathbb{1}\{s \in \mathcal{A}_i^\delta\}} = \begin{cases} \frac{1}{|\mathcal{A}_i^\delta|} & \text{if } s \in \mathcal{A}_i^\delta, \\ 0 & \text{otherwise.} \end{cases}
$$

62

Then we define

$$\widehat{\boldsymbol{\beta}}_i := \sum_{s \in [k_*]} \widehat{w}_s \cdot \boldsymbol{\theta}_s^* = \frac{1}{|\mathcal{A}_i^\delta|} \sum_{s \in \mathcal{A}_i^\delta} \boldsymbol{\theta}_s^*.$$

Further, we define two additional points in $\mathbb{R}^d$ that serve as proxies of $\boldsymbol{\beta}_i$ to facilitate our analysis:

$$\boldsymbol{\beta}_i' := \sum_{s \in \mathcal{B}_i^\delta} w_s \cdot \boldsymbol{\beta}_i + \sum_{s \in [k_*] \setminus \mathcal{B}_i^\delta} w_s \cdot \boldsymbol{v}_s, \tag{128}$$

$$\boldsymbol{\beta}_i'' := \sum_{s \in \mathcal{A}_i^\delta} \frac{w_s}{\sum_{s \in \mathcal{A}_i^\delta} w_s} \cdot \boldsymbol{v}_s. \tag{129}$$

Observe that

$$
\begin{aligned}
\boldsymbol{\beta}_i' - \boldsymbol{\beta}_i &\overset{(a)}{=} \left(\sum_{s \in \mathcal{B}_i^\delta} w_s \cdot \boldsymbol{\beta}_i + \sum_{s \in [k_*] \setminus \mathcal{B}_i^\delta} w_s \cdot \boldsymbol{v}_s \right) - \boldsymbol{\beta}_i \\
&= \sum_{s \in \mathcal{A}_i^\delta} w_s \cdot (\boldsymbol{v}_s - \boldsymbol{\beta}_i) + \sum_{s \in \mathcal{C}_i^\delta} w_s \cdot (\boldsymbol{v}_s - \boldsymbol{\beta}_i) \\
&\overset{(b)}{=} \left(\sum_{s \in \mathcal{A}_i^\delta} w_s \right) \cdot (\boldsymbol{\beta}_i'' - \boldsymbol{\beta}_i) + \sum_{s \in \mathcal{C}_i^\delta} w_s \cdot (\boldsymbol{v}_s - \boldsymbol{\beta}_i),
\end{aligned}
$$

where (a) is due to (128) and (b) is due to (129). Therefore,

$$\left\| \boldsymbol{\beta}_i'' - \boldsymbol{\beta}_i \right\| \leq \frac{1}{\sum_{s \in \mathcal{A}_i^\delta} w_s} \left(\left\| \boldsymbol{\beta}_i' - \boldsymbol{\beta}_i \right\| + \sum_{s \in \mathcal{C}_i^\delta} w_s \cdot \left\| \boldsymbol{v}_s - \boldsymbol{\beta}_i \right\| \right). \tag{130}$$

By triangle inequality and (130), we obtain

$$
\begin{aligned}
\left\| \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i \right\| &\leq \left\| \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i'' \right\| + \left\| \boldsymbol{\beta}_i'' - \boldsymbol{\beta}_i \right\| \\
&\leq \left\| \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i'' \right\| + \frac{1}{\sum_{s \in \mathcal{A}_i^\delta} w_s} \left\| \boldsymbol{\beta}_i' - \boldsymbol{\beta}_i \right\| + \frac{1}{\sum_{s \in \mathcal{A}_i^\delta} w_s} \sum_{s \in \mathcal{C}_i^\delta} w_s \cdot \left\| \boldsymbol{v}_s - \boldsymbol{\beta}_i \right\|.
\end{aligned} \tag{131}
$$

Step 2. Establishing separate upper bounds Next, we establish upper bounds for the three terms on the right hand side of (131).

Step 2-A. An upper bound on the first term in (131). Recall that $\widehat{\boldsymbol{\beta}}_i = \sum_{s \in \mathcal{A}_i^\delta} \widehat{w}_s \cdot \widehat{\boldsymbol{v}}_s$ and $\boldsymbol{\beta}_i'' = \sum_{s \in \mathcal{A}_i^\delta} w_s'' \cdot \boldsymbol{v}_s$ where

$$w_s'' = \frac{w_s}{\sum_{s \in \mathcal{A}_i^\delta} w_s} = \frac{\mathbb{E}_s[\Psi_i]}{\sum_{s \in \mathcal{A}_i^\delta} \mathbb{E}_s[\Psi_i]}.$$

Therefore, by Lemma 14, we obtain

$$\left\| \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i'' \right\| \leq \left(\sum_{s \in \mathcal{A}_i^\delta} |\widehat{w}_s - w_s''| \right) \cdot \min_{s \in \mathcal{A}_i^\delta} \max_{s' \in \mathcal{A}_i^\delta} \left\| \boldsymbol{\theta}_s^* - \boldsymbol{\theta}_{s'}^* \right\| + \sum_{s \in \mathcal{A}_i^\delta} \left\| w_s'' \cdot (\boldsymbol{\theta}_s^* - \boldsymbol{v}_s) \right\|. \tag{132}$$

Let $\varepsilon_\delta := 9 \left(\sqrt{2\pi} + 1 \right) \cdot k^3 \cdot \delta$, to avoid cluttered notation. Recall from Corollary 6 and Proposition 2 that

$$1 - \mathbb{E}_s[\Psi_i] \leq 9 \left(\sqrt{2\pi} + 1 \right) \cdot k^4 \cdot \delta = k \cdot \varepsilon_\delta, \qquad \forall s \in \mathcal{A}_i^\delta, \tag{133}$$

$$\mathbb{E}_s[\Psi_i] \leq 9 \left(\sqrt{2\pi} + 1 \right) \cdot k^3 \cdot \delta = \varepsilon_\delta, \qquad \forall s \in \mathcal{B}_i^\delta. \tag{134}$$

63

It follows from (133) that for all $s \in \mathcal{A}_i^\delta$,

$$\left| \widehat{w}_s - w_s'' \right| \leq \max \left\{ \frac{1}{|\mathcal{A}_i^\delta|} - \frac{1 - k \cdot \varepsilon_\delta}{|\mathcal{A}_i^\delta|}, \frac{1}{|\mathcal{A}_i^\delta|(1 - k \cdot \varepsilon_\delta)} - \frac{1}{|\mathcal{A}_i^\delta|} \right\}$$

$$= \frac{1}{|\mathcal{A}_i^\delta|} \cdot \frac{k \cdot \varepsilon_\delta}{1 - k \cdot \varepsilon_\delta} \tag{135}$$

Let $z_{s,i} \coloneqq Q^{-1}\left(\mathbb{E}_s[\Psi_i]\right)$ for all $s \in \mathcal{A}_i^\delta$, where Q is the Gaussian Q-funciton. Observe that

$$\sum_{s \in \mathcal{A}_i^\delta} \left\| w_s'' \cdot (\boldsymbol{\theta}_s^* - \boldsymbol{v}_s) \right\| = \frac{1}{\sum_{s \in \mathcal{A}_i^\delta} \mathbb{E}_s[\Psi_i]} \sum_{s \in \mathcal{A}_i^\delta} \left\| \mathbb{E}_s[\Psi_i] \cdot (\boldsymbol{\theta}_s^* - \boldsymbol{v}_s) \right\|$$

$$\overset{(a)}{=} \frac{1}{\sum_{s \in \mathcal{A}_i^\delta} \mathbb{E}_s[\Psi_i]} \sum_{s \in \mathcal{A}_i^\delta} \left\| \mathbb{E}_s\left[\Psi_i \cdot (\mathsf{x} - \boldsymbol{\theta}_s^*)\right] \right\|$$

$$\overset{(b)}{\leq} \frac{1}{\sum_{s \in \mathcal{A}_i^\delta} \mathbb{E}_s[\Psi_i]} \sum_{s \in \mathcal{A}_i^\delta} \phi(z_{s,i}) \cdot \sigma$$

$$\overset{(c)}{\leq} \frac{3k \cdot \varepsilon_\delta}{1 - k \cdot \varepsilon_\delta} \cdot \sigma \tag{136}$$

where (a) holds because $\boldsymbol{v}_s = \mathbb{E}_s\left[\frac{\Psi_i}{\mathbb{E}_s[\Psi_i]} \cdot \mathsf{x}\right]$; (b) is by Lemma 13; and (c) follows from the observation: for all $s \in \mathcal{A}_i^\delta$,

$$1 - \mathbb{E}_s[\Psi_i] \leq k \cdot \varepsilon_\delta \leq \frac{1}{4}$$

$$\implies \quad z_{s,i} \leq -0.5$$

$$\implies \quad \phi(z_\delta) = \phi(-z_\delta) \leq \frac{z_\delta^2 + 1}{-z_\delta} Q(-z_{\varepsilon_\delta}) \leq 3k \cdot \varepsilon_\delta.$$

Combining the inequalities (132), (135), (136), we obtain

$$\left\| \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i'' \right\| \leq \frac{k \cdot \varepsilon_\delta}{1 - k \cdot \varepsilon_\delta} \cdot \left(\min_{s \in \mathcal{A}_i^\delta} \max_{s' \in \mathcal{A}_i^\delta} \left\| \boldsymbol{\theta}_s^* - \boldsymbol{\theta}_{s'}^* \right\| + 3 \cdot \sigma \right). \tag{137}$$

Step 2-B. An upper bound on the second term in (131). Recall the definition of $\mathcal{B}_i^\delta$ from (120) and that $\boldsymbol{\beta}_i' = \sum_{s \in \mathcal{B}_i^\delta} w_s \cdot \boldsymbol{\beta}_i + \sum_{s \in [k_*] \setminus \mathcal{B}_i^\delta} w_s \cdot \boldsymbol{v}_s$, cf. (128). Thus, we observe that

$$\frac{1}{\sum_{s \in \mathcal{A}_i^\delta} w_s} \cdot \left\| \boldsymbol{\beta}_i' - \boldsymbol{\beta}_i \right\| = \frac{1}{\sum_{s \in \mathcal{A}_i^\delta} w_s} \cdot \left\| \sum_{s \in \mathcal{B}_i^\delta} w_s \cdot (\boldsymbol{v}_s - \boldsymbol{\beta}_i) \right\|$$

$$\leq \frac{1}{\sum_{s \in \mathcal{A}_i^\delta} \mathbb{E}_s[\Psi_i]} \cdot \sum_{s \in \mathcal{B}_i^\delta} \left\| \mathbb{E}_s[\Psi_i] \cdot (\boldsymbol{v}_s - \boldsymbol{\beta}_i) \right\|. \tag{138}$$

Recall that $\boldsymbol{v}_s = \mathbb{E}_s\left[\frac{\Psi_i}{\mathbb{E}_s[\Psi_i]} \cdot \mathsf{x}\right]$, and thus,

$$\mathbb{E}_s[\Psi_i] \cdot (\boldsymbol{v}_s - \boldsymbol{\beta}_i) = \mathbb{E}_s[\Psi_i \cdot \mathsf{x}] - \mathbb{E}_s[\Psi_i] \cdot \boldsymbol{\beta}_i$$

$$= \mathbb{E}_s\left[\Psi_i \cdot (\mathsf{x} - \boldsymbol{\theta}_s^*)\right] + \mathbb{E}_s[\Psi_i] \cdot (\boldsymbol{\theta}_s^* - \boldsymbol{\beta}_i).$$

Observe that

$$\left\| \mathbb{E}_s\left[\Psi_i \cdot (\mathsf{x} - \boldsymbol{\theta}_s^*)\right] \right\| \leq \mathbb{E}_s\left[\Psi_i \cdot \|\mathsf{x} - \boldsymbol{\theta}_s^*\|\right]$$

$$\leq E_s \|\mathsf{x} - \boldsymbol{\theta}_s^*\|$$

$$\leq \sigma \sqrt{d}. \tag{139}$$

64

By Lemma 16 and (139), we have

$$\left\| \mathbb{E}_s[\Psi_i] \cdot \left(\boldsymbol{v}_s - \boldsymbol{\beta}_i\right) \right\| \leq 45\left(\sqrt{2\pi} + 1\right) \cdot k^3 \cdot \delta \cdot D_{\mathrm{cell}}(\boldsymbol{B}) + \sigma\left(\sqrt{d} + 4\right)$$
$$= 5\varepsilon_\delta \cdot D_{\mathrm{cell}}(\boldsymbol{B}) + \sigma\left(\sqrt{d} + 4\right). \tag{140}$$

Combining (138) and (140), we obtain

$$\frac{1}{\sum_{s \in \mathcal{A}_i^\delta} w_s} \cdot \left\| \boldsymbol{\beta}_i' - \boldsymbol{\beta}_i \right\| \leq \frac{|\mathcal{B}_i^\delta|}{|\mathcal{A}_i^\delta|} \cdot \frac{5\varepsilon_\delta \cdot D_{\mathrm{cell}}(\boldsymbol{B}) + \sigma\left(\sqrt{d} + 4\right)}{1 - k \cdot \varepsilon_\delta}. \tag{141}$$

Step 2-C. An upper bound on the third term in (131). Recall that $\mathcal{C}_i^\delta$ is either an empty set or a singleton. If $\mathcal{C}_i^\delta = \emptyset$, there is nothing to prove. Thus, we assume $\mathcal{C}_i^\delta = \{s_0\}$. Let $z_{s_0} := Q^{-1}\left(\mathbb{E}_{s_0}[\Psi_i]\right)$, and observe that

$$
\begin{aligned}
\frac{w_{s_0}}{\sum_{s \in \mathcal{A}_i^\delta} w_s} \cdot \left\| \boldsymbol{v}_{s_0} - \boldsymbol{\beta}_i \right\| &\leq \frac{\mathbb{E}_{s_0}[\Psi_i]}{\sum_{s \in \mathcal{A}_i^\delta} \mathbb{E}_s[\Psi_i]} \cdot \left(\left\| \boldsymbol{v}_{s_0} - \boldsymbol{\theta}_{s_0}^* \right\| + \left\| \boldsymbol{\theta}_{s_0}^* - \boldsymbol{\beta}_i \right\| \right) \\
&\leq \frac{1}{|\mathcal{A}_i^\delta| \cdot (1 - k \cdot \varepsilon_\delta)} \cdot \left(\mathbb{E}_{s_0}\left[\Psi_i \cdot \left(\mathsf{x} - \boldsymbol{\theta}_{s_0}^*\right)\right] + \left\| \boldsymbol{\theta}_{s_0}^* - \boldsymbol{\beta}_i \right\| \right) \\
&\overset{(a)}{\leq} \frac{1}{|\mathcal{A}_i^\delta| \cdot (1 - k \cdot \varepsilon_\delta)} \cdot \left(\phi(z_{s_0}) \cdot \sigma + \frac{2k_*}{\delta} \cdot \sigma \right) \\
&\leq \frac{1}{1 - k \cdot \varepsilon_\delta} \cdot \frac{\sigma}{|\mathcal{A}_i^\delta|} \left(\frac{1}{\sqrt{2\pi}} + \frac{2k_*}{\delta} \right),
\end{aligned}
\tag{142}
$$

where (a) follows from Lemma 13 and Corollary 7.

Step 3. Concluding the proof To conclude the proof, we insert the upper bounds (137), (141), and (142) into (131) and observe that $\varepsilon_\delta := 9\left(\sqrt{2\pi} + 1\right) \cdot k^3 \cdot \delta \leq \frac{1}{2k}$ due to the assumption $\delta \leq \frac{1}{18(\sqrt{2\pi}+1)\cdot k^4}$:

$$
\begin{aligned}
\left\| \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i \right\| &\leq \frac{k \cdot \varepsilon_\delta}{1 - k \cdot \varepsilon_\delta} \cdot \left(\min_{s \in \mathcal{A}_i^\delta} \max_{s' \in \mathcal{A}_i^\delta} \left\| \boldsymbol{\theta}_s^* - \boldsymbol{\theta}_{s'}^* \right\| + 3 \cdot \sigma \right) + \frac{|\mathcal{B}_i^\delta|}{|\mathcal{A}_i^\delta|} \cdot \frac{5\varepsilon_\delta \cdot D_{\mathrm{cell}}(\boldsymbol{B}) + \sigma\left(\sqrt{d} + 4\right)}{1 - k \cdot \varepsilon_\delta} \\
&\quad + \frac{1}{1 - k \cdot \varepsilon_\delta} \cdot \frac{\sigma}{|\mathcal{A}_i^\delta|} \left(\frac{1}{\sqrt{2\pi}} + \frac{2k_*}{\delta} \right) \\
&\leq \left(2k \cdot \varepsilon_\delta \cdot \widetilde{\Delta}_{\mathrm{cell}}^{i,\delta}(\boldsymbol{B}) + 3 \cdot \sigma \right) + \frac{2|\mathcal{B}_i^\delta|}{|\mathcal{A}_i^\delta|} \cdot \left(5\varepsilon_\delta \cdot D_{\mathrm{cell}}(\boldsymbol{B}) + \sigma\left(\sqrt{d} + 4\right) \right) + \frac{2\sigma}{|\mathcal{A}_i^\delta|} \left(\frac{1}{\sqrt{2\pi}} + \frac{2k_*}{\delta} \right) \\
&\leq 18\left(\sqrt{2\pi} + 1 \right) \cdot k^3 \cdot \left(\frac{5k_*}{|\mathcal{A}_i^\delta|} \cdot D_{\mathrm{cell}}(\boldsymbol{B}) + k \cdot \widetilde{\Delta}_{\mathrm{cell}}^{i,\delta}(\boldsymbol{B}) \right) \cdot \delta + \frac{4k_* \cdot \sigma}{|\mathcal{A}_i^\delta|} \cdot \frac{1}{\delta} \\
&\quad + \sigma \left\{ \frac{1}{|\mathcal{A}_i^\delta|} \cdot \left(2k_* \cdot (\sqrt{d} + 4) + \sqrt{\frac{2}{\pi}} \right) + 3 \right\}.
\end{aligned}
$$

$\square$

E   Deferred Proof of Proposition 5

Proof of Proposition 5. In this proof, we prove the four claims one by one.

(1) Proof of Claim 1 (non-emptiness). We begin by showing the non-emptiness of $\mathcal{T}_a^\delta$. First, it is clear from line 5 of Algorithm 1 that $\mathcal{T}_a^\delta \neq \emptyset$ for all $a \in [q_0]$. Second, to prove $\mathcal{T}_a^\delta \neq \emptyset$ for all $a \in [q] \setminus [q_0]$, we assume $\exists a \in [q] \setminus [q_0]$ such that $\mathcal{E}_{s_a}^\delta = \emptyset$ where $s_a \in [k_*]$ denotes the unique element in $\mathcal{S}_a^\delta$. By this assumption, $\max_{i \in [k]} \max_{j \in [k] \setminus \{i\}} \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|}{\sigma} \cdot \mathbb{E}_{s_a}[\Psi_i \Psi_j] < \delta$. In addition, we observe that $s_a \notin \mathcal{A}_i^\delta$ for all $i \in [k]$ because

$s_a \in \mathcal{R}^\delta$. These two observations together imply that $\iota_{\boldsymbol{B}}(\boldsymbol{\theta}^*_{s_a}) \neq i$ for all $i \in [k]$, which is a contradiction. As a result, we conclude that $\mathcal{T}_a^\delta \neq \emptyset$ for all $a \in [q] \setminus [q_0]$. Next, it is easy to observe that $\mathcal{S}_a^\delta \neq \emptyset$ for all $a \in [q]$ by construction; see line 6 and line 15 of Algorithm 1.

Moreover, it is clear from the construction (see lines 5 and 15 in Algorithm 1) that $|\mathcal{T}_a^\delta| = 1$ for all $a \in [q_0]$, and $|\mathcal{S}_a^\delta| = 1$ for all $a \in [q] \setminus [q_0]$. Lastly, we can verify that $|\mathcal{T}_a^\delta| \geq 2$ for all $a \in [q] \setminus [q_0]$ because $|\mathcal{E}_s^\delta| \geq 2$ unless $\mathcal{E}_s^\delta = \emptyset$ by definition, cf. (36).

<u>*(2) Proof of Claim 2 (covering).*</u> It is clear from line 19 of Algorithm 1 that $\bigcup_{a=0}^{k} \mathcal{T}_a^\delta = [k]$.

Next, we show that for all $s \in [k_*]$, there exists $a \in [q]$ such that $s \in \mathcal{S}_a^\delta$. To this end, we choose an arbitrary $s \in [k_*]$ and consider two possibilities: (i) $\mathcal{E}_s^\delta = \emptyset$; and (ii) $\mathcal{E}_s^\delta \neq \emptyset$. If (i) is the case, then there exists $a \in [q_0]$ such that $s \in \mathcal{S}_a^\delta = \mathcal{A}_i^\delta$ where i_a is the unique element of $\mathcal{T}_a^\delta$; see lines 5 and 6 of Algorithm 1. If (ii) is the case, then there exists $a \in [q] \setminus [q_0]$ such that $\mathcal{S}_a^\delta = \{s\}$ by construction.

<u>*(3) Proof of Claim 3 (inclusion of Voronoi centers).*</u> For $a \in [q_0]$, if $s \in \mathcal{S}_a^\delta$, then $s \in \mathcal{A}_{i_a}^\delta$ where $i_a = \iota_{\boldsymbol{B}}(\boldsymbol{\theta}^*_s)$ is the unique element in $\mathcal{T}_a^\delta$. Thus, it is clearly true by construction

Next, we let $a \in [q] \setminus [q_0]$ and let s_a denote the unique element in $\mathcal{S}_a^\delta$. Then it suffices to show that $i = \iota_{\boldsymbol{B}}(\boldsymbol{\theta}^*_{s_a}) \in \mathcal{E}_{s_a}^\delta$. If we assume otherwise, then there must exist $b \in [q] \setminus [q_0]$ such that $b \neq a$ and $i \in \mathcal{T}_b^\delta$. Then it follows from Corollary 7 that

$$\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\theta}^*_{s_b}\|}{\sigma} \leq \frac{2k_*}{\delta}$$

where s_b is the unique element in $\mathcal{S}_b^\delta$. Also, we get

$$\frac{\|\boldsymbol{\beta}_i - \boldsymbol{\theta}^*_{s_a}\|}{\sigma} \leq \min_{j \in \mathcal{T}_a^\delta} \frac{\|\boldsymbol{\beta}_j - \boldsymbol{\theta}^*_{s_a}\|}{\sigma} \leq \frac{2k_*}{\delta}$$

because $\boldsymbol{\theta}^*_{s_a} \in \mathcal{V}_i$. Then we get

$$\begin{aligned}\left\|\boldsymbol{\theta}^*_{s_a} - \boldsymbol{\theta}^*_{s_b}\right\| &\leq \left\|\boldsymbol{\beta}_i - \boldsymbol{\theta}^*_{s_a}\right\| + \left\|\boldsymbol{\beta}_i - \boldsymbol{\theta}^*_{s_b}\right\| \\ &\leq \frac{4k_* \cdot \sigma}{\delta} \\ &< \Delta_{\min}\end{aligned}$$

where the last inequality follows from the premise of the theorem that $\delta > \frac{4k_* \cdot \sigma}{\Delta_{\min}}$. This is a contradiction due to the definition of $\Delta_{\min}$ in (4).

<u>*(4) Proof of Claim 4 (disjointness of $\mathbb{S}^\delta$).*</u>

(a) Let $a \in [q_0]$.

- Suppose that $b \in [q_0] \setminus \{a\}$. Assume that $\mathcal{S}_a^\delta \cap \mathcal{S}_b^\delta \neq \emptyset$, i.e., $\exists s \in \mathcal{S}_a^\delta \cap \mathcal{S}_b^\delta \neq \emptyset$. Letting i_a and i_b denote the unique elements of $\mathcal{T}_a^\delta$ and $\mathcal{T}_b^\delta$, respectively, we observe that $s \in \mathcal{A}_{i_a}^\delta \cap \mathcal{A}_{i_b}^\delta$ by construction; see lines 5 and 6 of Algorithm 1. Then it follows that

$$\begin{aligned}\mathbb{E}_s\left[\Psi_{i_a}\right] &\overset{(a)}{\geq} 1 - 9\big(\sqrt{2\pi} + 1\big) \cdot k^4 \cdot \delta \\ &\overset{(b)}{>} \frac{1}{2},\end{aligned}$$

  where (a) is by Proposition 2 (Claim 2); and (b) is by the premise that $\delta < \frac{1}{18(\sqrt{2\pi}+1) \cdot k^4}$. Likewise, we obtain $\mathbb{E}_s\left[\Psi_{i_b}\right] > \frac{1}{2}$, thereby, $\mathbb{E}_s\left[\Psi_{i_a} + \Psi_{i_b}\right] > 1$, which is a contradiction. Therefore, $\mathcal{S}_a^\delta \cap \mathcal{S}_b^\delta = \emptyset$ for all $b \in [q_0] \setminus \{a\}$.

- Suppose that $b \in [q] \setminus [q_0]$. We observe that $\mathcal{T}_b^\delta \subseteq \mathcal{R}^\delta = [k_*] \setminus \left(\bigcup_{a'=1}^{q_0} \mathcal{S}_{a'}^\delta\right)$ for any $b \in [q] \setminus [q_0]$, by construction (see line 10 and line 15 in Algorithm 1). Thus, $\mathcal{S}_a^\delta \cap \mathcal{S}_b^\delta = \emptyset$.

(b) Let $a \in [q] \setminus [q_0]$ and $b \in [q] \setminus ([q_0] \cup \{i\})$. It is obvious from construction that $\mathcal{S}_a^\delta \cap \mathcal{S}_b^\delta = \emptyset$.

66

(5) Proof of Claim 5 (partial disjointness of $\mathbb{T}^\delta$).

(a) Let $a = 0$. It is obvious that $\mathcal{T}_a^\delta \cap \mathcal{T}_b^\delta = \emptyset$ for all $b \in [q]$ by definition; see line 19 of Algorithm 1.

(b) Let $a \in [q_0]$.

- Suppose that $b \in [q_0] \setminus \{a\}$. It is clear that $\mathcal{T}_a^\delta \cap \mathcal{T}_b^\delta = \emptyset$ for all $b \in [q_0] \setminus \{a\}$ from line 5 of Algorithm 1.
- For the case $b \in [q] \setminus [q_0]$, we do not claim disjointness for $\mathbb{T}^\delta$.

(c) Let $a \in [q] \setminus [q_0]$ and $b \in [q] \setminus ([q_0] \cup \{i\})$. We assume there exists $i \in [k]$ such that $i \in \mathcal{T}_a^\delta \cap \mathcal{T}_b^\delta$. Letting s_a and s_b denote the unique elements in $\mathcal{S}_a^\delta$ and $\mathcal{S}_b^\delta$, respectively, we have

$$\max_{j \in [k] \setminus \{i\}} \frac{\|\beta_i - \beta_j\|}{\sigma} \cdot \mathbb{E}_{s_a}\left[\Psi_i \Psi_j\right] \geq \delta \quad \text{and}$$

$$\max_{j \in [k] \setminus \{i\}} \frac{\|\beta_i - \beta_j\|}{\sigma} \cdot \mathbb{E}_{s_b}\left[\Psi_i \Psi_j\right] \geq \delta.$$

Then it follows from Corollary 7 that

$$\begin{aligned}
\left\|\boldsymbol{\theta}_{s_a}^* - \boldsymbol{\theta}_{s_b}^*\right\| &\leq \left\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_{s_a}^*\right\| + \left\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_{s_b}^*\right\| \\
&\leq \frac{4k_* \cdot \sigma}{\delta} \\
&< \Delta_{\min}
\end{aligned}$$

where the last inequality follows from the premise of the theorem that $\delta > \frac{4k_* \cdot \sigma}{\Delta_{\min}}$. This is a contradiction due to the definition of $\Delta_{\min}$ in (4): $\left\|\boldsymbol{\theta}_{s_a}^* - \boldsymbol{\theta}_{s_b}^*\right\| \geq \Delta_{\min}$. Therefore, $\mathcal{T}_a^\delta \cap \mathcal{T}_b^\delta = \emptyset$.

$\square$

F   Deferred Proof of Proposition 6

Proof of Proposition 6. Our proof is based on the induction argument. We let $q^{(t)}, q_0^{(t)}, \mathbb{T}^{(t)}, \mathbb{S}^{(t)}$ denote the values and the collections of sets after the t-th iteration of the outermost loop in Algorithm 2. Assuming the conclusions hold at the end of the t-th iteration, we consider the two possible scenarios at the $(t+1)$-th iteration.

<u>Scenario A:</u> $\mathcal{T}_a \setminus \bigcup_{b \in \mathcal{B}_a^{\mathrm{conf}}} \mathcal{T}_b \neq \emptyset$, i.e., Line 3 of Algortihm 2 is executed.

1. Claim 1:

    (a) Comparing $\mathbb{T}^{(t+1)}$ to $\mathbb{T}^{(t)}$, we only have made one $\mathcal{T}_a$ smaller (but $\mathcal{T}_a$ is still non-empty). To show that the new $\mathbb{T}^{(t+1)} \setminus \{\mathcal{T}_0\}$ is still a $\left(\{\mathcal{T}_a\}_{a=1}^{q_0}, \{\mathcal{T}_a\}_{a=q_0+1}^q\right)$-quasi-partition of $[k] \setminus \mathcal{T}_0$, it suffices to verify that $\bigcup_{\mathcal{T} \in \mathbb{T} \setminus \{\mathcal{T}_0\}} \mathcal{T} = [k] \setminus \mathcal{T}_0$. This is true because $\bigcup_{b \in \mathcal{B}_a^{\mathrm{conf}}} \mathcal{T}_b$ still contains the elements that are removed from $\mathcal{T}_b$.

    (b) $\mathbb{S}^{(t+1)}$ is a partition because $\mathbb{S}^{(t)}$ is a partition and $\mathbb{S}^{(t+1)} = \mathbb{S}^{(t)}$.

    (c) Claims 1-(c) and 1-(d) of Theorem 4 hold as the sets $\mathcal{T}_{a'}$, $a' \in [q] \setminus \{a\}$ and $\mathcal{S}_{b'}$, $b' \in [q]$ are unaffected.

2. Claim 2 continues to hold as the sets $\mathcal{T}_{a'}$, $a' \in [q] \setminus \{a\}$ and $\mathcal{S}_{b'}$, $b' \in [q]$ are unaffected.

3. Claim 3:

    (a) Equation (32) still holds because the sets $\mathcal{T}_a, \mathcal{S}_a$ remain unchanged for all $a \in [q_0]$.

    (b) Equation (33) still holds because

$$\max_{i \in \mathcal{T}_a \setminus \bigcup_{b \in \mathcal{B}_a^{\mathrm{conf}}} \mathcal{T}_b} \left\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_{s_a}^*\right\| \leq \max_{i \in \mathcal{T}_a} \left\|\boldsymbol{\beta}_i - \boldsymbol{\theta}_{s_a}^*\right\|.$$

67

<u>Scenario B:</u> $\mathcal{T}_a \setminus \bigcup_{b \in \mathcal{B}_a^{\mathrm{conf}}} \mathcal{T}_b = \emptyset$, i.e., Lines 5–12 of Algorithm 2 are executed.

1. Claim 1:

   (a) The only change to $\mathbb{T}$ was removing $\mathcal{T}_b$, so we only need to check the sets in the new $\mathbb{T} \setminus \{\mathcal{T}_0\}$ still cover $[k] \setminus \mathcal{T}_0$. This is true since the condition $\mathcal{T}_b \setminus \bigcup_{a \in A_b} \mathcal{T}_a = \emptyset$ implies that all the elements in $\mathcal{T}_b$ are still contained in other $\mathcal{T}_a$'s.

   (b) $\mathbb{S}^{(t)}$ is a partition of $[k_*]$ and we only merged two sets in $\mathbb{S}^{(t)}$. Thus, $\mathbb{S}^{(t+1)}$ is still a partition.

   (c) It is trivial to see that Claims 1-(c) and 1-(d) of Theorem 4 continue to hold.

2. It is straightforward that Claim 2 continues to hold.

3. Claim 3:

   (a) Recall from Lines 5–6 of Algorithm 1 that $\mathcal{T}_{b_0} = \{i\}$ where $i = \arg\min \mathcal{T}_a$. By the premise for $\mathbb{T}^{(t)}$, we have

   - $\frac{1}{\sigma}\left\| \boldsymbol{\beta}_i - \frac{1}{|\mathcal{S}_{b_0}|} \sum_{s \in \mathcal{S}_{b_0}} \boldsymbol{\theta}_s^* \right\| \le \varepsilon_1$ where $\varepsilon_1 = \left(2^6 \cdot 3^3 \cdot (\sqrt{2\pi} + 1) \cdot k_* \cdot k^3 \cdot (k + k_*) \cdot \frac{\Delta_{\max}}{\sigma}\right)^{1/2} + 2 d_{\mathrm{eff}}^{1/2}$, and

   - $\frac{1}{\sigma}\left\| \boldsymbol{\beta}_i - \boldsymbol{\theta}_{s_a}^* \right\|$, where $\varepsilon_2 = \left(2^2 \cdot 3^3 \cdot (\sqrt{2\pi} + 1) \cdot k_* \cdot k^3 \cdot (k + k_*) \cdot \frac{\Delta_{\max}}{\sigma}\right)^{1/2}$ and s_a denotes the unique element in $\mathcal{S}_a$.

   Therefore, it follows that

   $$
   \frac{1}{\sigma}\left\| \boldsymbol{\beta}_i - \frac{1}{|\mathcal{S}_a \cup \mathcal{S}_{b_0}|} \sum_{s \in \mathcal{S}_a \cup \mathcal{S}_{b_0}} \boldsymbol{\theta}_s^* \right\| \le \frac{|\mathcal{S}_a|}{|\mathcal{S}_a \cup \mathcal{S}_{b_0}|} \cdot \frac{1}{\sigma}\left\| \boldsymbol{\beta}_i - \frac{1}{|\mathcal{S}_a|} \sum_{s \in \mathcal{S}_a} \boldsymbol{\theta}_s^* \right\|
   $$

   $$
   + \frac{|\mathcal{S}_{b_0}|}{|\mathcal{S}_a \cup \mathcal{S}_{b_0}|} \cdot \frac{1}{\sigma}\left\| \boldsymbol{\beta}_i - \frac{1}{|\mathcal{S}_{b_0}|} \sum_{s \in \mathcal{S}_{b_0}} \boldsymbol{\theta}_s^* \right\|
   $$

   $$
   = \frac{1}{|\mathcal{S}_{b_0}| + 1} \cdot \frac{1}{\sigma}\left\| \boldsymbol{\beta}_i - \boldsymbol{\theta}_{s_a}^* \right\| + \frac{|\mathcal{S}_{b_0}|}{|\mathcal{S}_{b_0}| + 1} \cdot \frac{1}{\sigma}\left\| \boldsymbol{\beta}_i - \frac{1}{|\mathcal{S}_{b_0}|} \sum_{s \in \mathcal{S}_{b_0}} \boldsymbol{\theta}_s^* \right\|
   $$

   $$
   \le \frac{1}{|\mathcal{S}_{b_0}| + 1} \cdot (|\mathcal{S}_{b_0}| \cdot \varepsilon_1 + \varepsilon_2)
   $$

   $$
   \le \max\{\varepsilon_1, \varepsilon_2\} = \varepsilon_1.
   $$

   (b) Equation (33) holds because it is only $\mathcal{T}_a, \mathcal{S}_a$ that are removed during the $(t+1)$-th iteration, and all the other sets $\mathcal{T}_{a'}, \mathcal{S}_{a'}$, $a' \in [q] \setminus [q_0] \setminus \{a\}$ remain the same.

   $\square$

G   Deferred Proofs of Technical Lemmas in Section 8.1.2

G.1   Proof of Lemma 5

Proof of Lemma 5. In this proof we let $\mathbb{E}_0$ denote $\mathbb{E}_{\mathsf{x} \sim \mathcal{N}(0,1)}$. By Stein's identity (Lemma 3) and the premise that $\beta \ge |\alpha|$, we have

$$
\mathbb{E}_0\left[\psi_{\alpha,\beta}(\mathsf{x}) \cdot (\beta - \mathsf{x})\right] = \beta \cdot \mathbb{E}_0\left[\psi_{\alpha,\beta}(\mathsf{x})^2\right] + \alpha \cdot \mathbb{E}_0\left[(1 - \psi_{\alpha,\beta}(\mathsf{x})) \cdot \psi_{\alpha,\beta}(\mathsf{x})\right]
$$
$$
\ge \beta \cdot \mathbb{E}_0\left[\psi_{\alpha,\beta}(\mathsf{x}) \cdot (2\psi_{\alpha,\beta}(\mathsf{x}) - 1)\right]. \tag{143}
$$

Letting $c := \frac{\alpha + \beta}{2}$ and $\delta := \frac{\beta - \alpha}{2}$, we may write

$$
\psi_{\alpha,\beta}(x) \cdot (2\psi_{\alpha,\beta}(x) - 1) = e^{-\frac{(x-\beta)^2}{2}} \cdot g(x)
$$

68

where

$$g(x) = \frac{e^{(x-c)\delta} - e^{-(x-c)\delta}}{\left(e^{(x-c)\delta} + e^{-(x-c)\delta}\right)} \cdot \frac{1}{\left(e^{-(x-c-\delta)^2/2} + e^{-(x-c+\delta)^2/2}\right)}. \tag{144}$$

Observe that $g(c+z) = -g(c-z)$. By a change of variable $z = x - c$, we obtain

$$\begin{aligned}
\mathbb{E}_0\left[\psi_{\alpha,\beta}(\mathsf{x}) \cdot \left(2\psi_{\alpha,\beta}(\mathsf{x}) - 1\right)\right] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-\beta)^2/2} \cdot g(x) \cdot e^{-x^2/2} \, \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(z+c) \cdot e^{-(z+c-\beta)^2/2} \cdot e^{-(z+c)^2/2} \, \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} g(z+c) \cdot \underbrace{e^{-(z+c-\beta)^2/2} \cdot e^{-(z+c)^2/2}}_{=:h_1(z)} \mathrm{d}z \\
&\quad - \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} g(z+c) \cdot \underbrace{e^{-(-z+c-\beta)^2/2} \cdot e^{-(-z+c)^2/2}}_{=:h_2(z)} \mathrm{d}z. \tag{145}
\end{aligned}$$

Note that $\frac{h_1(z)}{h_2(z)} = e^{-2\alpha z} \geq 1$, $\forall z \geq 0$ because $\alpha < 0$ by assumption. Because $g(z+c) \geq 0$ for all $z \geq 0$, we have

$$\mathbb{E}_0\left[\psi_{\alpha,\beta}(\mathsf{x}) \cdot \left(2\psi_{\alpha,\beta}(\mathsf{x}) - 1\right)\right] \geq \frac{1}{\sqrt{2\pi}} \int_{-1/\alpha}^{\infty} g(z+c) \cdot \left[h_1(z) - h_2(z)\right] \mathrm{d}z.$$

Then we observe that by definition, cf. (144), for all $z \geq -1/\alpha > 1/\delta$,

$$\begin{aligned}
g(z+c) &= \frac{e^{z\delta} - e^{-z\delta}}{\left(e^{z\delta} + e^{-z\delta}\right) \cdot \left(e^{-(z-\delta)^2/2} + e^{-(z+\delta)^2/2}\right)} \\
&\geq \frac{e - e^{-1}}{e + e^{-1}} \cdot \frac{1}{2e^{-(z-\delta)^2/2}} \\
&\geq \frac{1}{4} e^{(z-\delta)^2/2}.
\end{aligned}$$

Moreover, if $z \geq -1/\alpha$, then $\frac{h_1(z)}{h_2(z)} = e^{-2\alpha z} = e^2 \geq 2$, and therefore,

$$\begin{aligned}
\mathbb{E}_0\left[\psi_{\alpha,\beta}(\mathsf{x}) \cdot \left(2\psi_{\alpha,\beta}(\mathsf{x}) - 1\right)\right] &\geq \frac{1}{\sqrt{2\pi}} \int_{-1/\alpha}^{\infty} \frac{1}{4} e^{(z-\delta)^2/2} \cdot \frac{1}{2} h_1(z) \, \mathrm{d}z \\
&= \frac{1}{8\sqrt{2\pi}} \int_{-1/\alpha}^{\infty} e^{(z-\delta)^2/2} \cdot e^{-(z+c-\beta)^2/2} \cdot e^{-(z+c)^2/2} \, \mathrm{d}z \\
&= \frac{1}{8\sqrt{2\pi}} \int_{-1/\alpha}^{\infty} e^{-(z-c)^2/2} \, \mathrm{d}z \\
&= \frac{1}{8} \cdot Q\left(c - \frac{1}{\alpha}\right). \tag{146}
\end{aligned}$$

Combining (143) and (146) completes the proof. $\qquad\square$

G.2   Proof of Lemma 6

Proof of Lemma 6. Recall the definition of $\psi_{\alpha,\beta}$ from (47) and observe that

$$\begin{aligned}
\psi_{\alpha,\beta}(x) &= \frac{e^{-(x-\beta)^2/2}}{e^{-(x-\alpha)^2/2} + e^{-(x-\beta)^2/2}} \\
&= \frac{1}{1 + e^{-(\beta-\alpha)\cdot\left(x - \frac{\alpha+\beta}{2}\right)}}. \tag{147}
\end{aligned}$$

In this proof we let $\mathbb{E}_0$ denote $\mathbb{E}_{\mathsf{x} \sim \mathcal{N}(0,1)}$ and Var_0 denote $\mathrm{Var}_{\mathsf{x} \sim \mathcal{N}(0,1)}$. Letting $c := \frac{\alpha+\beta}{2}$, $\delta := \beta - \alpha$, and $\mu = \mathbb{E}_0 [\psi_{\alpha,\beta}(\mathsf{x})]$, we may assume $c \geq 0$. Also, we may assume $\delta > 0$; if $\delta < 0$, we consider $\psi'_{\alpha,\beta}(x) = 1 - \psi_{\alpha,\beta}(\mathsf{x})$ instead of $\psi_{\alpha,\beta}(x)$ as $\mathrm{Var}_0 [\psi_{\alpha,\beta}(\mathsf{x})] = \mathrm{Var}_0 \left[\psi'_{\alpha,\beta}(\mathsf{x}) \right]$. We have

$$
\begin{aligned}
\sqrt{2\pi} \cdot \mathrm{Var}_0 (\psi_{\alpha,\beta}(\mathsf{x})) &= \sqrt{2\pi} \cdot \mathbb{E}_0 \left[(\psi_{\alpha,\beta}(\mathsf{x}) - \mu)^2 \right] \\
&\overset{(a)}{=} \int_{-\infty}^{\infty} \left(\frac{1}{1 + e^{-\delta \cdot (x-c)}} - \mu \right)^2 \cdot e^{-x^2/2} \, dx \\
&= \int_{-\infty}^{c} \left(\frac{1}{1 + e^{-\delta \cdot (x-c)}} - \mu \right)^2 \cdot e^{-x^2/2} \, dx + \int_{c}^{\infty} \left(\frac{1}{1 + e^{-\delta \cdot (x-c)}} - \mu \right)^2 \cdot e^{-x^2/2} \, dx \\
&= \int_{-c}^{\infty} \left(\frac{1}{1 + e^{\delta \cdot (x+c)}} - \mu \right)^2 \cdot e^{-x^2/2} \, dx + \int_{c}^{\infty} \left(\frac{1}{1 + e^{-\delta \cdot (x-c)}} - \mu \right)^2 \cdot e^{-x^2/2} \, dx \\
&\overset{(b)}{\geq} \int_{c}^{\infty} \left[\left(\frac{1}{1 + e^{\delta \cdot (x+c)}} - \mu \right)^2 + \left(\frac{1}{1 + e^{-\delta \cdot (x-c)}} - \mu \right)^2 \right] \cdot e^{-x^2/2} \, dx,
\end{aligned}
$$

where (a) follows from (147), and (b) follows from that $c \geq 0$ and $\left(\frac{1}{1 + e^{\delta \cdot (x+c)}} - \mu \right)^2 \geq 0$ for all $x \in [-c, c]$. Because $u^2 + v^2 \geq \frac{1}{2}(u - v)^2$ for all $u, v \in \mathbb{R}$, it follows that

$$
\begin{aligned}
\sqrt{2\pi} \cdot \mathrm{Var}_0 (\psi_{\alpha,\beta}(\mathsf{x})) &\geq \frac{1}{2} \int_{c}^{\infty} \left(\frac{1}{1 + e^{-\delta \cdot (x-c)}} - \frac{1}{1 + e^{\delta \cdot (x+c)}} \right)^2 \cdot e^{-x^2/2} \, dx \\
&= \frac{1}{2} \int_{c}^{\infty} \frac{\left(e^{\delta \cdot (x+c)} - e^{-\delta \cdot (x-c)} \right)^2}{\left(1 + e^{-\delta \cdot (x-c)} \right)^2 \left(1 + e^{\delta \cdot (x+c)} \right)^2} \cdot e^{-x^2/2} \, dx \\
&\overset{(a)}{\geq} \frac{1}{8} \int_{c}^{\infty} \frac{\left(e^{\delta \cdot (x+c)} - 1 \right)^2}{\left(1 + e^{\delta \cdot (x+c)} \right)^2} \cdot e^{-x^2/2} \, dx \\
&\geq \frac{1}{8} \int_{c+\epsilon_1}^{c+\epsilon_2} \left(\frac{e^{\delta \cdot (x+c)} - 1}{e^{\delta \cdot (x+c)} + 1} \right)^2 \cdot e^{-x^2/2} \, dx, \quad (148)
\end{aligned}
$$

for any $\epsilon_1, \epsilon_2 \in \mathbb{R}$ such that $0 \leq \epsilon_1 \leq \epsilon_2$. Note that we used $e^{-\delta \cdot (x-c)} \leq 1 \leq e^{\delta \cdot (x+c)}$, $\forall x \geq c$ to get the inequality (a). Since the function $x \mapsto \frac{e^{\delta \cdot (x+c)} - 1}{e^{\delta \cdot (x+c)} + 1}$ is non-decreasing, we have

$$
\begin{aligned}
\int_{c+\epsilon_1}^{c+\epsilon_2} \left(\frac{e^{\delta \cdot (x+c)} - 1}{e^{\delta \cdot (x+c)} + 1} \right)^2 \cdot e^{-x^2/2} \, dx &\geq (\epsilon_2 - \epsilon_1) \cdot \left(\frac{e^{\delta \cdot (2c+\epsilon_1)} - 1}{e^{\delta \cdot (2c+\epsilon_1)} + 1} \right)^2 \cdot e^{-(c+\epsilon_2)^2/2} \\
&\geq (\epsilon_2 - \epsilon_1) \cdot \left(\frac{\delta \cdot (2c + \epsilon_1)}{2 e^{\delta \cdot (2c+\epsilon_1)}} \right)^2 \cdot e^{-(c+\epsilon_2)^2/2} \\
&\overset{(a)}{=} (c + \delta) \cdot \left(\frac{\delta \cdot (2c + \delta)}{2 e^{\delta \cdot (2c+\delta)}} \right)^2 \cdot e^{-2(c+\delta)^2} \\
&\geq \frac{1}{4} \delta^2 (c + \delta)^3 \cdot e^{-4(c+\delta)^2}, \quad (149)
\end{aligned}
$$

where (a) is attained by choosing $\epsilon_1 = \delta$, $\epsilon_2 = c + 2\delta$. Inserting the lower bound (149) to (148), we obtain

$$
\mathrm{Var}_0 (\psi_{\alpha,\beta}(\mathsf{x})) \geq \frac{1}{32\sqrt{2\pi}} \delta^2 (c + \delta)^3 \cdot e^{-4(c+\delta)^2}.
$$

$\square$

G.3 Proof of Lemma 7

Proof of Lemma 7. Let $\alpha \geq 0$ be a parameter whose value will be determined later in this proof. Then we write

$$\mathbb{E}_s\left[\Psi_j\right] = \mathbb{E}_s\left[\Psi_j \cdot \mathbb{1}\left\{|\mathsf{x} - \boldsymbol{\beta}_j| - |\mathsf{x} - \boldsymbol{\beta}_{i_s}| \geq \alpha\right\}\right] + \mathbb{E}_s\left[\Psi_j \cdot \mathbb{1}\left\{|\mathsf{x} - \boldsymbol{\beta}_j| - |\mathsf{x} - \boldsymbol{\beta}_{i_s}| < \alpha\right\}\right]. \tag{150}$$

We establish upper bounds for the two terms on the right-hand side of (150) separately.

First, for x such that $|x - \boldsymbol{\beta}_j| - |x - \boldsymbol{\beta}_{i_s}| \geq \alpha$, we have $|x - \boldsymbol{\beta}_j|^2 \geq |x - \boldsymbol{\beta}_{i_s}|^2 + \alpha^2$, and thus,

$$\begin{aligned}
\psi_j(x) &= \frac{e^{-|x - \boldsymbol{\beta}_j|^2/2}}{\sum_{i' \in [k]} e^{-|x - \boldsymbol{\beta}_{i'}|^2/2}} \\
&\leq \frac{e^{-|x - \boldsymbol{\beta}_j|^2/2}}{e^{-|x - \boldsymbol{\beta}_j|^2/2} + e^{-|x - \boldsymbol{\beta}_{i_s}|^2/2}} \\
&\leq \frac{1}{1 + e^{\alpha^2/2}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}_s\left[\Psi_j \cdot \mathbb{1}\left\{|\mathsf{x} - \boldsymbol{\beta}_j| - |\mathsf{x} - \boldsymbol{\beta}_{i_s}| \geq \alpha\right\}\right] &\leq \mathbb{E}_s\left[\frac{1}{1 + e^{\alpha^2/2}} \cdot \mathbb{1}\left\{|\mathsf{x} - \boldsymbol{\beta}_j| - |\mathsf{x} - \boldsymbol{\beta}_{i_s}| \geq \alpha\right\}\right] \\
&\leq \frac{1}{1 + e^{\alpha^2/2}}.
\end{aligned} \tag{151}$$

Second, we observe that

$$\left\{x \in \mathbb{R} : |x - \boldsymbol{\beta}_j| - |x - \boldsymbol{\beta}_{i_s}| < \alpha\right\} \subseteq \left\{x \in \mathbb{R} : \left\langle x, \frac{\boldsymbol{\beta}_j - \boldsymbol{\theta}_s^*}{|\boldsymbol{\beta}_j - \boldsymbol{\theta}_s^*|}\right\rangle \geq \frac{\delta_j}{2} - \alpha\right\} =: S_\alpha.$$

Therefore,

$$\begin{aligned}
\mathbb{E}_s\left[\Psi_j \cdot \mathbb{1}\left\{|\mathsf{x} - \boldsymbol{\beta}_j| - |\mathsf{x} - \boldsymbol{\beta}_{i_s}| < \alpha\right\}\right] &\leq \mathbb{E}_s\left[\Psi_j \cdot \mathbb{1}\left\{\mathsf{x} \in S_\alpha\right\}\right] \\
&\leq \mathbb{E}_s\left[\mathbb{1}\left\{\mathsf{x} \in S_\alpha\right\}\right] \\
&= Q\left(\frac{\delta_j}{2} - \alpha\right)
\end{aligned} \tag{152}$$

where Q is the Gaussian Q-function, cf. (42).

Plugging in (151) and (152) to (150) and using the Q-function upper bound in (43), we obtain

$$\mathbb{E}_s\left[\Psi_j\right] \leq \frac{1}{1 + e^{\alpha^2/2}} + \frac{1}{\frac{\delta_j}{2} - \alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\delta_j}{2} - \alpha\right)^2}.$$

Choosing $\alpha = \frac{\delta_j}{4}$ and further simplifying this upper bound, we complete the proof. $\square$

G.4 Proof of Lemma 8

Proof of Lemma 8. Recall that $\boldsymbol{\beta}_i = \frac{\mathbb{E}_*[\Psi_i \mathsf{x}]}{\mathbb{E}_*[\Psi_i]}$, cf. (12), for all $i \in [k]$ and $\boldsymbol{\theta}_s^* = \mathbb{E}_s[\mathsf{x}]$, for all $s \in [k_*]$. Moreover, we may assume $\frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \boldsymbol{\theta}_s^* = \mathbf{0}$ without loss of generality (see Remark 1). Thus,

$$\boldsymbol{\beta}_i - \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \boldsymbol{\theta}_s^* = \frac{\sum_{s \in [k_*]} \mathbb{E}_s\left[\Psi_i \mathsf{x}\right]}{\sum_{s \in [k_*]} \mathbb{E}_s\left[\Psi_i\right]}. \tag{153}$$

Then we observe that

$$\sum_{s \in [k_*]} \mathbb{E}_s \left[\Psi_i \mathsf{x} \right] = \sum_{s \in \mathcal{S}} \mathbb{E}_s \left[\Psi_i \mathsf{x} \right] + \sum_{s \in [k_*] \setminus \mathcal{S}} \mathbb{E}_s \left[\Psi_i \mathsf{x} \right]$$

$$= \sum_{s \in \mathcal{S}} \mathbb{E}_s[\mathsf{x}] + \underbrace{\left(\sum_{s \in \mathcal{S}} \sum_{j \in [k] \setminus \{i\}} \mathbb{E}_s \left[\Psi_j \mathsf{x} \right] + \sum_{s \in [k_*] \setminus \mathcal{S}} \mathbb{E}_s \left[\Psi_i \mathsf{x} \right] \right)}_{=: \epsilon_1} \tag{154}$$

because $\sum_{j \in [k]} \Psi_j = 1$ with probability 1 (w.r.t. all s). Likewise, we can see that

$$\sum_{s \in [k_*]} \mathbb{E}_s \left[\Psi_i \right] = \sum_{s \in \mathcal{S}} \mathbb{E}_s 1 + \underbrace{\left(\sum_{s \in \mathcal{S}} \sum_{j \in [k] \setminus \{i\}} \mathbb{E}_s \left[\Psi_j \right] + \sum_{s \in [k_*] \setminus \mathcal{S}} \mathbb{E}_s \left[\Psi_i \right] \right)}_{=: \epsilon_2}. \tag{155}$$

Combining (154) and (155) with (153), we get

$$\boldsymbol{\beta}_i - \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \boldsymbol{\theta}_s^* = \frac{\sum_{s \in \mathcal{S}} \mathbb{E}_s[\mathsf{x}] + \epsilon_1}{\sum_{s \in \mathcal{S}} \mathbb{E}_s 1 + \epsilon_2} = \frac{\epsilon_1}{|\mathcal{S}| + \epsilon_2} \tag{156}$$

because $\sum_{s \in \mathcal{S}} \mathbb{E}_s[\mathsf{x}] = \sum_{s \in \mathcal{S}} \boldsymbol{\theta}_s^* = \mathbf{0}$.

Next, we argue that the "perturbation terms" — ϵ_1 and ϵ_2 in (154) and (155) — have small norms. To this end, we begin by observing that

$$\| \mathbb{E}_s[\Psi_j \mathsf{x}] \| \le \left(\mathbb{E}_s[\Psi_j^2] \cdot \mathbb{E}_s \left[\|\mathsf{x}\|^2 \right] \right)^{1/2} \qquad \because \text{Cauchy-Schwarz}$$

$$\le \left(\mathbb{E}_s[\Psi_j] \cdot \mathbb{E}_s \left[\|\mathsf{x}\|^2 \right] \right)^{1/2} \qquad \because 0 \le \Psi_j \le 1.$$

It follows from the premise (50) and Lemma 7 that for any $(s, j) \in (\mathcal{S} \times [k] \setminus \{i\}) \cup ([k_*] \setminus \mathcal{S} \times \{i\})$,

$$\mathbb{E}_s \left[\Psi_j \right] \le \left(1 + \frac{4}{\sqrt{2\pi} \cdot \delta} \right) e^{-\frac{\delta^2}{32}}.$$

Also, we can easily observe that

$$\mathbb{E}_s \left[\|\mathsf{x}\|^2 \right] \le \mathbb{E}_s \left[\|\mathsf{x} - \boldsymbol{\theta}_s^*\|^2 \right] + \|\boldsymbol{\theta}_s^*\|^2$$

$$\le \sigma^2 \cdot d + \Delta_{\max}^2$$

$$= 1 + \Delta_{\max}^2.$$

All in all,

$$|\epsilon_2| \le (|\mathcal{S}| \cdot (k-1) + (k_* - |\mathcal{S}|)) \cdot \left(1 + \frac{4}{\sqrt{2\pi} \cdot \delta} \right)^{1/2} e^{-\frac{\delta^2}{64}}$$

$$\le k k_* \cdot \left(1 + \frac{4}{\sqrt{2\pi} \cdot \delta} \right)^{1/2} e^{-\frac{\delta^2}{64}}, \tag{157}$$

$$\|\epsilon_1\| \le (|\mathcal{S}| \cdot (k-1) + (k_* - |\mathcal{S}|)) \cdot (1 + \Delta_{\max}^2)^{1/2} \cdot \left(1 + \frac{4}{\sqrt{2\pi} \cdot \delta} \right)^{1/2} e^{-\frac{\delta^2}{64}}$$

$$\le \sqrt{2} k k_* \Delta_{\max} \cdot \left(1 + \frac{4}{\sqrt{2\pi} \cdot \delta} \right)^{1/2} e^{-\frac{\delta^2}{64}}. \tag{158}$$

Observe that $|\epsilon_2| \leq 1/2$ because $\delta \geq \max\left\{\frac{4}{\sqrt{2\pi}}, 8\sqrt{\log(2\sqrt{2} \cdot kk_*)}\right\}$. Combining the upper bounds (157) and (158) with (156), we obtain

$$\left\|\boldsymbol{\beta}_i - \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\boldsymbol{\theta}_s^*\right\| = \frac{\|\epsilon_1\|}{\big||\mathcal{S}| + \epsilon_2\big|}$$

$$\leq \frac{\|\epsilon_1\|}{|\mathcal{S}| - |\epsilon_2|} \leq 2\|\epsilon_1\|$$

$$\leq 4kk_*\Delta_{\max} \cdot e^{-\frac{\delta^2}{64}}.$$

$\square$