Demystifying Softmax Gating in Gaussian Mixture of Experts

Huy Nguyen[†] TrungTin Nguyen[◇] Nhat Ho[†]

The University of Texas at Austin[†],
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France[◇]
May 5, 2023

Abstract

Understanding parameter estimation of softmax gating Gaussian mixture of experts has remained a long-standing open problem in the literature. It is mainly due to three fundamental theoretical challenges associated with the softmax gating: (i) the identifiability only up to the translation of the parameters; (ii) the intrinsic interaction via partial differential equation between the softmax gating and the expert functions in Gaussian distribution; (iii) the complex dependence between the numerator and denominator of the conditional density of softmax gating Gaussian mixture of experts. We resolve these challenges by proposing novel Vononoi loss functions among parameters and establishing the convergence rates of the maximum likelihood estimator (MLE) for solving parameter estimation in these models. When the number of experts is unknown and over-specified, our findings show a connection between the rate of MLE and a solvability problem of a system of polynomial equations.

1 Introduction

Softmax gating Gaussian mixture of experts [20, 22], a class of statistical machine learning models that combine multiple simpler models, known as expert functions of the covariates, via softmax gating network to form more complex and accurate models, have found widespread use in various applications, including speech recognition [27, 35, 36], natural language processing [11, 8, 29, 12], computer vision [28], and other applications [15, 25]. In softmax gating Gaussian mixture of experts, the parameters of each expert function play an important role in capturing the heterogeneity of data. However, a comprehensive theoretical understanding of parameter estimation in these models has still remained a long-standing open problem in the literature.

Parameter estimation had been studied quite extensively in standard mixture models. In his seminal work, Chen [4] established the convergence rate $\mathcal{O}(n^{-1/4})$ of parameter estimation in over-fitted univariate mixture models, namely, the settings when the number of true components is unknown and over-specified, when the family of distribution is strongly identifiable in the second order, e.g., location Gaussian distribution. That slow and non-standard convergence rate is due to the collapse of some parameters into single pararameter or the vanishing of weights to 0, which leads to the singularity of Fisher information matrix around the true parameters. Then, Nguyen [26] and Ho et al. [18] utilized Wasserstein metric to achieve this rate under the multivariate settings of second-order strongly identifiable mixture models. Recently, Ho et al. [17] demonstrated that rates of the MLE can strictly depend on the amount of overspecified components when the mixture models are not strongly identifiable, such as location-scale Gaussian mixtures. The minimax optimal behaviors of parameter estimation were studied in [16, 24]. From the computational side, the statistical guarantee

of the expectation-maximization (EM) and moment methods had also been studied under both exact-fitted [2, 1, 14] and over-fitted settings [10, 9, 32, 7, 33] of mixture models.

Compared to mixture models, there has been less research on parameter estimation of mixture of experts. When the gating networks are independent of the covariates, Ho et al. [19] employed generalized Wasserstein to study the rates of parameter estimation in Gaussian mixture of experts. They proved that these rates are determined by the algebraic independence of the expert functions and the partial differential equations associated with the parameters. Later, Do et al. [6] extended these results to general mixture of experts with covariate-free gating network. Statistical guarantees of optimization methods for solving parameter estimation in Gaussian mixture of experts with covariate-free gating were studied in [5, 37, 23, 34]. When the gating networks are softmax functions, parameter estimation becomes more challenging to understand due to the complex structures of the softmax gating in the Gaussian mixture of experts. Before describing these phenomena in further detail, we begin by formally introducing softmax gating Gaussian mixture of experts and related notions.

Problem setting: Assume that $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ are i.i.d. samples from the softmax gating Gaussian mixture of experts of order k_* whose conditional density function $g_{G_*}(Y|X)$ is given by:

$$\sum_{i=1}^{k_*} \frac{\exp((\beta_{1i}^*)^\top X + \beta_{0i}^*)}{\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)} f(Y|(a_i^*)^\top X + b_i^*, \sigma_i^*), \tag{1}$$

where $f(.|\mu, \sigma)$ is a Gaussian distribution with mean μ and variance σ. Here, we define $G_* := \sum_{i=1}^{k_*} \exp(\beta_{0i}^*)\delta_{(\beta_{1i}^*, a_i^*, b_i^*, \sigma_i^*)}$ is a true but unknown mixing measure associated with the true parameters and δ is denoted as Dirac delta measure. Notably, G_* is not necessarily a probability measure as the summation of its weights can be different from 1. For the purpose of the theory, we assume that $(\beta_{0i}^*, \beta_{1i}^*, a_i^*, b_i^*, \sigma_i^*) \in \Theta \subset \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+$ where Θ is a compact set and $X \in \mathcal{X} \subset \mathbb{R}^d$ where $\mathcal{X}$ is a bounded set. Furthermore, $(a_1^*, b_1^*, \sigma_1^*), \ldots, (a_{k_*}^*, b_{k_*}^*, \sigma_{k_*}^*)$ are pairwise different and at least one of $\beta_{11}^*, \ldots, \beta_{1k_*}^*$ is different from 0. Since the value of true order k_* is unknown in practice, to estimate the unknown parameters in softmax gating Gaussian mixture of experts (1), we consider using maximum likelihood estimation (MLE) within a class of at most k Gaussian mixture of experts, which is defined as follows:

$$\widehat{G}_n \in \arg\max_{G \in \mathcal{O}_k(\Theta)} \frac{1}{n} \sum_{i=1}^{n} \log(g_G(Y_i|X_i)), \tag{2}$$

where $\mathcal{O}_k(\Theta) := \{G = \sum_{i=1}^{k'} \exp(\beta_{0i})\delta_{(\beta_{1i}, a_i, b_i, \sigma_i)} : k' \leq k \text{ and } (\beta_{0i}, \beta_{1i}, a_i, b_i, \sigma_i) \in \Theta\}$. To guarantee that the MLE $\widehat{G}_n$ is a consistent estimator of G_*, we need $k \geq k_*$. When $k = k_*$, we refer this setting to exact-specified softmax gating Gaussian mixture of experts. When $k > k_*$, we call this setting as over-specified softmax gating Gaussian mixture of experts.

In this paper, we study the convergence rate of the MLE $\widehat{G}_n$ to the true mixing measure G_* under both the exact-specified and over-specified settings of the softmax gating Gaussian mixture of experts.

Fundamental challenges from the softmax gating: There are three fundamental challenges arising from the softmax gating that create various obstacles when trying to comprehend the

behaviors of the MLE: (i) the parameters $\beta_{1i}^*, \beta_{0i}^*$ are only identifiable up to translation, namely, the softmax weights remain identical when we translate β_{1i}^* to $\beta_{1i}^* + t_1$ and β_{0i}^* to $\beta_{0i}^* + t_2$ for some t_1, t_2; (ii) the numerators in softmax weights have an intrinsic interaction with the expert functions in Gaussian distribution via the following partial differential equation (PDE):

$$\frac{\partial u(X,Y)}{\partial \beta_1} \cdot \frac{\partial u(X,Y)}{\partial b} = \frac{\partial u(X,Y)}{\partial \beta_0} \cdot \frac{\partial u(X,Y)}{\partial a}, \tag{3}$$

where $u(X,Y) = \exp(\beta_1^\top X + \beta_0) \cdot f(Y|a^\top X + b, \sigma)$; (iii) the numerator and denominator of the conditional density of Gaussian mixture of experts (1) are dependent.

These fundamental challenges from the softmax gating suggest that the previous loss functions, such as Wasserstein distance [26, 17, 19], being employed to study parameter estimation in standard mixture models or mixture of experts with covariate-free gating functions are no longer sufficient as these loss functions heavily rely on the assumptions that the weights of these models are independent of the covariates.

Main contributions: To tackle these challenges of the softmax gating, we propose novel Voronoi losses among parameters and establish the lower bounds of the Hellinger distance of the mixing densities of softmax gating Gaussian mixture of experts in terms of these Voronoi losses to capture the behaviors of the MLE. Our results can be summarized as follows:

1. Exact-fitted settings: When $k = k_*$, we demonstrate that $h(g_G, g_{G_*}) \geq C \cdot \mathcal{D}_1(G, G_*)$ for any $G \in \mathcal{O}_k(\Theta)$ where C is some universal constant and the Voronoi metric $\mathcal{D}_1(G, G_*)$ is defined as:

$$\mathcal{D}_1(G, G_*) := \inf_{t_1,t_2} \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}) \|(\Delta_{t_2}\beta_{1ij}, \Delta a_{ij}, \Delta b_{ij}, \Delta \sigma_{ij})\| + |\sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}) - \exp(\beta_{0j}^* + t_1)|, \tag{4}$$

where $\Delta_{t_2}\beta_{1ij} := \beta_{1i} - \beta_{1j}^* - t_2$, $\Delta a_{ij} := a_i - a_j^*$, $\Delta b_{ij} := b_i - b_j^*$, $\Delta \sigma_{ij} := \sigma_i - \sigma_j^*$. The infimum over t_1, t_2 is to account for the identifiability up to the translation of $(\beta_{0j}^*, \beta_{1j}^*)_{j=1}^{k_*}$. Furthermore, $\mathcal{A}_j := \{i \in \{1, 2, \ldots, k\} : \|\theta_i - \theta_j^*\| \leq \|\theta_i - \theta_\ell^*\| \ \forall \ell \neq j\}$ is Voronoi cell of $\theta_j^* = (a_j^*, b_j^*, \sigma_j^*)$ for all $1 \leq j \leq k_*$ where $\theta_i = (a_i, b_i, \sigma_i)$.

As $h(g_{\widehat{G}_n}, g_{G_*}) = \mathcal{O}(n^{-1/2})$ (up to logarithmic term), that lower bound of Hellinger distance indicates that $\mathcal{D}_1(\widehat{G}_n, G_*) = \mathcal{O}(n^{-1/2})$. Therefore, the rates of MLE to estimate $\exp(\beta_{0j}^*), \beta_{1j}^*$ (up to translations) and $a_j^*, \beta_j^*, \sigma_j^*$ are $\mathcal{O}(n^{-1/2})$, which are optimal.

2. Over-fitted settings: When $k > k_*$, the lower bound of Hellinger distance in terms of the Voronoi metric D_1 in the exact-fitted settings is no longer enough due to the collapse of softmax of vector in possibly k dimensions to softmax of vector in k_* dimensions. Our approach is to define

more fine-grained Vononoi metric $\mathcal{D}_2(G, G_*)$ to capture such collapse, which is given by:

$$\mathcal{D}_2(G, G_*) := \inf_{t_1,t_2} \sum_{j:|\mathcal{A}_j|>1} \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i})(\|(\Delta_{t_2}\beta_{1ij}, \Delta b_{ij})\|^{\bar{r}(|\mathcal{A}_j|)} + \|(\Delta a_{ij}, \Delta\sigma_{ij})\|^{\bar{r}(|\mathcal{A}_j|)/2})$$

$$+ \sum_{j:|\mathcal{A}_j|=1} \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i})\|(\Delta_{t_2}\beta_{1ij}, \Delta a_{ij}, \Delta b_{ij}, \Delta\sigma_{ij})\| + \sum_{j=1}^{k_*} |\sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}) - \exp(\beta^*_{0j} + t_1)|, \tag{5}$$

for any $G := \sum_{i=1}^{k'} \exp(\beta_{0i})\delta_{(\beta_{1i}, a_i, b_i, \sigma_i)} \in \mathcal{O}_k(\Theta)$. Here, $\bar{r}(|\mathcal{A}_j|)$ is the smallest number r such that the following system of polynomial equations does not have any non-trivial solutions:

$$\sum_{l=1}^{|\mathcal{A}_j|} \sum_{(\alpha_1,\alpha_2,\alpha_3,\alpha_4)\in\mathcal{I}_{\ell_1,\ell_2}} \frac{1}{\alpha!} p_{5l}^2 p_{1l}^{\alpha_1} p_{2l}^{\alpha_2} p_{3l}^{\alpha_3} p_{4l}^{\alpha_4} = 0, \tag{6}$$

where $\mathcal{I}_{\ell_1,\ell_2} = \{\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \mathbb{N}^d \times \mathbb{N}^d \times \mathbb{N} \times \mathbb{N} : \ \alpha_1 + \alpha_2 = \ell_1, \ |\alpha_2| + \alpha_3 + 2\alpha_4 = \ell_2\}$ for any $(\ell_1, \ell_2) \in \mathbb{N}^d \times \mathbb{N}$ such that $0 \leq |\ell_1| \leq r$, $0 \leq \ell_2 \leq r - |\ell_1|$ and $|\ell_1| + \ell_2 \geq 1$. In this system, $\{p_{1l}, p_{2l}, p_{3l}, p_{4l}, p_{5l}\}_{l=1}^{|\mathcal{A}_j|}$ are unknown variables. A solution is considered to be non-trivial if at least one among p_{3l} is different from 0 and all of p_{5l} are non-zero. Some specific values of $\bar{r}(|\mathcal{A}_j)$ can be found in Lemma 1.

In high level, the system of polynomial equations (6) arises from the PDE (3) when we establish the lower bound $h(g_G, g_{G_*}) \geq C'\mathcal{D}_2(G, G_*)$ for any $G \in \mathcal{O}_k(\Theta)$ for some universal constant C'. Since $h(g_{\widehat{G}_n}, g_{G_*}) = \mathcal{O}(n^{-1/2})$, we also have $\mathcal{D}_2(\widehat{G}_n, G_*) = \mathcal{O}(n^{-1/2})$ under the over-fitted settings of the softmax gating Gaussian mixture of experts. As a consequence, the rates for estimating true parameters whose Voronoi cells have only one component of the MLE are $\mathcal{O}(n^{-1/2})$. On the other hand, for true parameters $\exp(\beta^*_{0i}), \beta^*_{1i}, a^*_i, b^*_i, \sigma^*_i$ whose Voronoi cells have more than one component of the MLE, the estimation rates are respectively $\mathcal{O}(n^{-1/2\bar{r}(|\mathcal{A}_i)|})$ for β^*_{1i}, b^*_i, $\mathcal{O}(n^{-1/\bar{r}(|\mathcal{A}_i)})$ for a^*_i, σ^*_i, and $\mathcal{O}(n^{-1/2})$ for $\exp(\beta^*_{0i})$. That rich spectrum of the convergence rates of the parameters is due to the complex interaction between the softmax gating and the expert functions.

Practical implication: In practice, as the true number of experts k_* is generally unknown and the rates of MLE depend on the number of extra components under the over-fitted settings of softmax gating Gaussian mixture of experts, the value of the number of experts k should not be chosen very large compared to k_*. Furthermore, the slow convergence rates of the MLE may provide important thresholds in the merge-truncate-merge procedure, a procedure that was used to estimate the true number of components in standard mixture models [13], to consistently estimate the true number of experts k_*. A high-level idea of that procedure is that we can merge the MLE parameters that are close and within the range of their rates of convergence or truncate the parameters that lead to small weights of the experts. As the sample size becomes sufficiently large, the reduced number of experts may converge to the true number of experts. We leave a theoretical investigation of that procedure in future work.

Organization: The paper is organized as follows. In Section 2, we first provide background on the identifiability and rate of conditional density estimation in softmax gating Gaussian mixture of experts. Then, we proceed to establish the convergence rate of the MLE under both the exact-fitted and over-fitted settings of these models in Section 3. The conclusion of the paper is in Section 4. Finally, proofs of the results in the paper are in the Appendices.

Notation: For any positive integer n, we denote $[n] = \{1, 2, \ldots, n\}$. For any $\alpha \in \mathbb{N}^d$, we denote $|\alpha|$ as the summation of elements of α. For any positive sequences $\{a_n\}_{n\geq 1}, \{b_n\}_{n\geq 1}$, we write $a_n = \mathcal{O}(b_n)$ or $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for all $n \geq 1$ where $C > 0$ is some universal constant. Furthermore, we write $a_n \asymp b_n$ when $a_n \lesssim b_n \lesssim a_n$. Given two probability density functions p, q dominated by measure μ, we denote $h^2(p, q) = \frac{1}{2}\int(\sqrt{p} - \sqrt{q})^2 d\mu$ as the their squared Hellinger distance and $V(p, q) = \frac{1}{2}\int |p - q|d\mu$ as their Total Variation distance.

2 Background

In this section, we first start with the following result about the identifiability of the softmax gating Gaussian mixture of experts, which was studied previously in [21].

Proposition 1 (Identifiability of softmax gating Gaussian mixture of experts)**.** *For any mixing measures $G = \sum_{i=1}^{k} \exp(\beta_{0i})\delta_{(\beta_{1i}, a_i, b_i, \sigma_i)}$ and $G' = \sum_{i=1}^{k'} \exp(\beta'_{0i})\delta_{(\beta'_{1i}, a'_i, b'_i, \sigma'_i)}$, if we have $g_G(Y|X) = g_{G'}(Y|X)$ for almost surely (X, Y), then it follows that $k = k'$ and $G \equiv G'_{t_1,t_2}$ where $G'_{t_1,t_2} := \sum_{i=1}^{k'} \exp(\beta'_{0i} + t_1)\delta_{(\beta'_{1i}+t_2, a'_i, b'_i, \sigma'_i)}$ for some $t_1 \in \mathbb{R}$ and $t_2 \in \mathbb{R}^d$.*

Proof of Proposition 1 is in Appendix B.1. The identifiability of the softmax gating Gaussian mixture of experts guarantees that the MLE $\widehat{G}_n$ (2) converges to the true mixing measure G_* (up to the translation of the parameters in the softmax gating).

Given the consistency of the MLE, it is natural to ask about its convergence rate to the true parameters. Our next result establishes the convergence rate of conditional density estimation $g_{\widehat{G}_n}(Y|X)$ to the true conditional density estimation $g_{G_*}(Y|X)$, which lays an important foundation for the study of MLE's convergence rate.

Proposition 2 (Convergence rate of conditional density estimator)**.** *Given the MLE in equation (2), the conditional density function $g_{\widehat{G}_n}(Y|X)$ has the following convergence rate:*

$$\mathbb{P}(h(g_{\widehat{G}_n}, g_{G_*}) > C(\log(n)/n)^{1/2}) \lesssim \exp(-c\log n),$$

where c and C are universal constants.

Proof of Proposition 2 is in Appendix B.2. The result of Proposition 2 indicates that under either the exact-fitted or over-fitted settings of the softmax gating Gaussian mixture of experts, the rate of the conditional density function $g_{\widehat{G}_n}(Y|X)$ to the true one $g_{G_*}(Y|X)$ under Hellinger distance is parametric $\mathcal{O}(n^{-1/2})$ (up to some logarithmic factors).

From density estimation to parameter estimation: The parametric rate of the conditional density function from the MLE in Proposition 2 suggests that as long as we can establish the following lower bound $h(g_G(Y|X), g_{G_*}(Y|X)) \gtrsim \mathcal{D}(G, G_*)$ for any $G \in \mathcal{O}_k(\Theta)$ for some metric $\mathcal{D}$ among the parameters, then we obtain directly the parametric convergence rate of the MLE under the metric $\mathcal{D}$. Therefore, the main focus of the next section is to determine such metric $\mathcal{D}$ and establish that lower bound under either exact-fitted or over-fitted settings of the softmax gating Gaussian mixture of experts.

3 Convergence Rate of the Maximum Likelihood Estimation

In this section, we first study the convergence rate of the MLE under the exact-fitted settings of the softmax gating Gaussian mixture of experts in Section 3.1. Then, we move to the over-fitted settings in Section 3.2. Finally, we provide proof sketch of the theories in Section 3.3.

3.1 Exact-fitted Settings

For the exact-fitted settings, namely, when the chosen number of experts k is equal to the true number of experts k_*, as we mentioned in the introduction the proper metric between the MLE and the true mixing measure is the metric $\mathcal{D}_1$ defined in equation (4), which is given by:

$$\mathcal{D}_1(G, G_*) := \inf_{t_1, t_2} \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}) \|(\Delta_{t_2}\beta_{1ij}, \Delta a_{ij}, \Delta b_{ij}, \Delta \sigma_{ij})\| + |\sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}) - \exp(\beta^*_{0j} + t_1)|,$$

where $\Delta_{t_2}\beta_{1ij} := \beta_{1i} - \beta^*_{1j} - t_2$, $\Delta a_{ij} := a_i - a^*_j$, $\Delta b_{ij} := b_i - b^*_j$, $\Delta \sigma_{ij} := \sigma_i - \sigma^*_j$. Here, $\mathcal{A}_j$ is Voronoi cell of $(a^*_j, b^*_j, \sigma^*_j)$ for all $1 \leq j \leq k_*$. Furthermore, the infimum is taken with respect to $(t_1, t_2) \in \mathbb{R} \times \mathbb{R}^d$ such that $\beta^*_{0j} + t_1$ and $\beta^*_{1j} + t_2$ still lie inside the domain of the parameter space Θ.

It is clear that $\mathcal{D}_1(G, G_*) = 0$ if and only if $G \equiv G_*$ (up to translation). When $\mathcal{D}_1(G, G_*)$ is sufficiently small, there exist t_1, t_2 such that all of $\Delta_{t_2}\beta_{1ij}$, Δa_{ij}, Δb_{ij}, $\Delta \sigma_{ij}$, and $\sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}) - \exp(\beta^*_{0j} + t_1)$ are sufficiently small as well. Therefore, the loss function $\mathcal{D}_1$ provides a useful metric to measure the difference between the MLE and the true mixing measure. For any fixed t_1, t_2, the computation of the summations in $\mathcal{D}_1$ only has the complexity of the order $\mathcal{O}(k_*^2)$. To solve the optimization with respect to t_1, t_2 in the metric $\mathcal{D}_1$, we can utilize the projected subgradient method with fixed step size [3], which has the complexity of the order $\mathcal{O}(\varepsilon^{-2})$ as the functions of t_1 and t_2 are convex where ε is a desired tolerance. Therefore, the total computational complexity of approximating the value of the Voronoi loss function $\mathcal{D}_1$ is at the order of $\mathcal{O}(k_*^2/\varepsilon^2)$.

The following result establishes the lower bound of the Hellinger distance between the conditional densities in terms of the loss function $\mathcal{D}_1$ between corresponding mixing measures, which in turn leads to the convergence rate of the MLE.

Theorem 1. *Given the exact-fitted settings of the softmax gating Gaussian mixture of experts (1), i.e., $k = k_*$, we find that*

$$h(g_G, g_{G_*}) \geq C \cdot \mathcal{D}_1(G, G_*), \tag{7}$$

for any $G \in \mathcal{E}_{k_*}(\Theta) := \mathcal{O}_{k_*}(\Theta) \setminus \mathcal{O}_{k_*-1}(\Theta)$ where C is some universal constant depending only on G_* and Θ. As a consequence, there exist universal constants C' and c such that the convergence rate of the MLE $\widehat{G}_n$ under the exact-fitted settings satisfies:

$$\mathbb{P}(\mathcal{D}_1(\widehat{G}_n, G_*) > C'(\log(n)/n)^{1/2}) \lesssim \exp(-c \log n). \tag{8}$$

Proof of Theorem 1 is in Appendix A.1. The parametric convergence rate of the MLE to G_* under the metric $\mathcal{D}_1$ suggests that the rates of estimating the true parameters $\exp(\beta^*_{0j}), \beta^*_{1j}$ (up to translation), a^*_j, b^*_j, σ^*_j for $j \in [k_*]$ are $\mathcal{O}((\log(n)/n)^{1/2})$, which are optimal up to logarithmic factors.

3.2 Over-fitted Settings

We now consider the over-fitted settings of the softmax gating Gaussian mixture of experts. Different from the exact-fitted settings, the softmax weights associated with the MLE collapse to softmax weights of the mixture of true experts as long as the MLE approaches the true mixing measure G_*. More concretely, we can relabel the supports of the MLE $\widehat{G}_n$ with k_n components ($k_n \leq k$) such that we can rewrite it as $\widehat{G}_n = \sum_{i=1}^{k_*}\sum_{j=1}^{s_i^n}\exp(\widehat{\beta}_{0ij}^n)\delta_{(\widehat{\beta}_{1ij}^n,\widehat{a}_{ij}^n,\widehat{b}_{ij}^n,\widehat{\sigma}_{ij}^n)}$ where $\sum_{i=1}^{k_*} s_i^n = k_n$, $(\widehat{a}_{ij}^n,\widehat{b}_{ij}^n,\widehat{\sigma}_{ij}^n) \to (a_i^*,b_i^*,\sigma_i^*)$ for all $1 \leq j \leq s_i^n$, and

$$\sum_{j=1}^{s_i^n}\frac{\exp((\widehat{\beta}_{1ij}^n)^\top X + \widehat{\beta}_{0ij}^n)}{\sum_{i=1}^{k_*}\sum_{j=1}^{s_i^n}\exp((\widehat{\beta}_{1ij}^n)^\top X + \widehat{\beta}_{0ij}^n)} \to \frac{\exp((\beta_{1i}^*)^\top X + \beta_{0i}^*)}{\sum_{j=1}^{k_*}\exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)}$$

as n approaches infinity for all $i \in [k_*]$.

The collapse of softmax weights along with the PDE (3) between the softmax gating and the expert functions in the Gaussian distribution create a complex interaction among the estimated parameters. To disentangle such interaction, we rely on the solvability of a novel system of polynomial equations defined in equation (6). In particular, for any $m \geq 2$, we define $\bar{r}(m)$ as the smallest number r such that the following system of polynomial equations

$$\sum_{j=1}^{m}\sum_{(\alpha_1,\alpha_2,\alpha_3,\alpha_4)\in\mathcal{I}_{\ell_1,\ell_2}}\frac{1}{\alpha!}p_{5j}^2 p_{1j}^{\alpha_1}p_{2j}^{\alpha_2}p_{3j}^{\alpha_3}p_{4j}^{\alpha_4} = 0,$$

for any $(\ell_1,\ell_2) \in \mathbb{N}^d \times \mathbb{N}$ such that $0 \leq |\ell_1| \leq r$, $0 \leq \ell_2 \leq r - |\ell_1|$ and $|\ell_1| + \ell_2 \geq 1$, does not have any non-trivial solution for the unknown variables $\{p_{1j}, p_{2j}, p_{3j}, p_{4j}, p_{5j}\}_{j=1}^m$, namely, all of p_{5j} are non-zeros and at least one among p_{3j} is different from 0. The ranges of $\alpha_1,\alpha_2,\alpha_3,\alpha_4$ in the sum satisfy $\mathcal{I}_{\ell_1,\ell_2} = \{\alpha = (\alpha_1,\alpha_2,\alpha_3,\alpha_4) \in \mathbb{N}^d \times \mathbb{N}^d \times \mathbb{N} \times \mathbb{N} : \ \alpha_1 + \alpha_2 = \ell_1, \ |\alpha_2| + \alpha_3 + 2\alpha_4 = \ell_2\}$. When $d = 1$ and $r = 2$, that system of equations becomes

$$\sum_{j=1}^{m} p_{5j}^2 p_{1j} = 0, \ \sum_{j=1}^{m} p_{5j}^2 p_{1j}^2 = 0, \ \sum_{j=1}^{m} p_{5j}^2(p_{1j}p_{3j} + p_{2j}) = 0,$$
$$\sum_{j=1}^{m} p_{5j}^2 p_{3j} = 0, \ \sum_{j=1}^{m} p_{5j}^2\Big(\frac{1}{2}p_{3j}^2 + p_{4j}\Big) = 0.$$

It is clear from that we have non-trivial solutions $p_{5j} = 1$, $p_{1j} = 0$ for all $j \in [m]$, $|p_{21}| = p_{31} = 1$, $|p_{22}| = p_{32} = -1$, $p_{41} = p_{42} = -1/2$, $p_{2j} = p_{3j} = p_{4j} = 0$ for $j \geq 3$. Therefore, $\bar{r}(m) \geq 3$ when $d = 1$ and $m \geq 2$.

In general, when $d = 1$, the system of equations has $(r^2 + 3r)/2$ equations. Intuitively, when m is sufficiently larger than $(r^2 + 3r)/2$, the system may not have non-trivial solution. In general, for general dimension d and parameter $m \geq 2$, finding the exact value of $\bar{r}(m)$ is non-trivial and a central problem in algebraic geometry [30]. When m is small, the following lemma provides specific values for $\bar{r}(m)$.

Lemma 1. *For any $d \geq 1$, when $m = 2$, $\bar{r}(m) = 4$. When $m = 3$, $\bar{r}(m) = 6$.*

Proof of Lemma 1 is in Appendix B.3. When m increases, the value of $\bar{r}(m)$ also increases. Given the results of Lemma 1, we conjecture that $\bar{r}(m) = 2m$ and leave the proof for that conjecture for the future work.

The following result demonstrates that the convergence rates of the MLE under the over-fitted settings of the softmax gating Gaussian mixture of experts are determined by $\bar{r}(.)$.

Theorem 2. *Under the over-fitted settings of the softmax gating Gaussian mixture of experts (1), namely, when $k > k_*$, we obtain that*

$$h(g_G, g_{G_*}) \geq C \cdot \mathcal{D}_2(G, G_*), \tag{9}$$

for any $G \in \mathcal{O}_k(\Theta)$ where the Voronoi loss $\mathcal{D}_2$ is defined in equation (5) and C is some universal constant depending only on G_* and Θ. Therefore, that lower bound leads to the following convergence rate of the MLE:

$$\mathbb{P}(\mathcal{D}_2(\widehat{G}_n, G_*) > C'(\log(n)/n)^{1/2}) \lesssim \exp(-c \log n). \tag{10}$$

Proof of Theorem 2 is in Appendix A.2. A few comments with the result of Theorem 2 are in order.

First, the convergence rate $\mathcal{O}(n^{-1/2})$ (up to some logarithmic term) of the MLE under the loss function $\mathcal{D}_2$ implies that for the true parameters $\exp(\beta_{0i}^*), \beta_{1i}^*, a_i^*, b_i^*, \sigma_i^*$ whose Voronoi cells have only one component of the MLE, the rates for estimating them are $\mathcal{O}(n^{-1/2})$. On the other hand, for true parameters with greater than one components in their Voronoi cells, the rates for estimating β_{1i}^*, b_i^* are $\mathcal{O}(n^{-1/2\bar{r}(|\mathcal{A}_i|)})$ while those for a_i^*, σ_i^* are $\mathcal{O}(n^{-1/\bar{r}(|\mathcal{A}_i|)})$. As the maximum value of $|\mathcal{A}_i|$ is $k - k_* + 1$, it indicates that these rates can be as worse as $\mathcal{O}(n^{-1/\bar{r}(k-k_*+1)})$ for estimating a_i^*, σ_i^* and $\mathcal{O}(n^{-1/2\bar{r}(k-k_*+1)})$ for estimating the remaining parameters. Finally,.....

Although the slow rates of the MLE under the over-fitted settings of the softmax gating Gaussian mixture of experts may seem discouraging, a practical implication of these results is that we should not choose k to be very large compared to the true number of experts k_*. Furthermore, the slow rates can also be useful for post-processing procedure, such as merge-truncate-merge procedure [13], with the MLE to reduce the number of experts so as to consistently estimate k_* when the number of data is sufficiently large. We leave an investigation of model selection with Gaussian mixture of experts via the rates of MLE for the future work.

Second, similar to the Voronoi loss function $\mathcal{D}_1$ in the exact-fitted settings, the loss function $\mathcal{D}_2$ is also computationally efficient. In particular, for any fixed t_1, t_2, the computation of the summations in the formulation of $\mathcal{D}_2$ is at the order $\mathcal{O}(k \times k_*)$, which is linear on k when k_* is fixed. Furthermore, we can solve the convex optimization problem with respect to t_1, t_2 with computational complexity at the order of $\mathcal{O}(\varepsilon^{-2})$ via the projected gradient descent method with fixed step size where ε is the error. Therefore, the total computational complexity of approximating the Voronoi loss function $\mathcal{D}_2$ is at the order of $\mathcal{O}(k \times k_*/\varepsilon^2)$.

3.3 Proof Sketch

In this section, we provide a proof sketch for Theorems 1 and 2. To simplify the ensuing discussion, the loss function $\mathcal{D}$ in the proof sketch is implicitly understood as either the loss function $\mathcal{D}_1$ or $\mathcal{D}_2$

depending on the settings of the softmax gating Gaussian mixture of experts. To obtain the bound of Hellinger distance between g_G and g_{G_*} in terms of $\mathcal{D}(G, G_*)$, it is sufficient to consider the lower bound of the total variation distance $V(g_G, g_{G_*})$ in terms of $\mathcal{D}(G, G_*)$. To establish this bound, we respectively prove its local and global versions by contradiction as follows:

Local version: $\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{O}_k(\Theta), \mathcal{D}(G, G_*) \leq \varepsilon} V(g_G, g_{G_*})/\mathcal{D}(G, G_*) > 0$. Assume that this claim does not hold true, that is, there exists a sequence $G_n = \sum_{i=1}^{k_n} \exp(\beta_{0i}^n)\delta_{(\beta_{1i}^n, a_i^n, b_i^n, \sigma_i^n)} \in \mathcal{O}_k(\Theta)$ such that both $V(g_{G_n}, g_{G_*})/\mathcal{D}(G_n, G_*)$ and $\mathcal{D}(G_n, G_*)$ approach zero as n tends to infinity. This implies that for any $j \in [k_*]$, we have $\sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n) \to \exp(\beta_{0j}^*)$ and $(\beta_{1i}^n, a_i^n, b_i^n, \sigma_i^n) \to (\beta_{1j}^*, a_j^*, b_j^*, \sigma_j^*)$ and for all $i \in \mathcal{A}_j$. For the sake of presentation, we simplify the loss function $\mathcal{D}$ by assuming that it is minimized when $t_1 = 0$ and $t_2 = \mathbf{0}_d$. Now, we decompose the quantity $Q_n = [\sum_{j'=1}^{k_*} \exp((\beta_{1j'}^*)^\top X + \beta_{0j'}^*)] \cdot [g_{G_n}(Y|X) - g_{G_*}(Y|X)]$ as follows:

$$
\begin{aligned}
Q_n = \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n) \Big[& u(X, Y|\beta_{1i}^n, a_i^n, b_i^n, \sigma_i^n) - u(X, Y|\beta_{1j}^*, a_j^*, b_j^*, \sigma_j^*) - v(X, Y|\beta_{1i}^n) \\
& + v(X, Y|\beta_{1j}^*) \Big] + \sum_{j=1}^{k_*} \Big(\sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^*) \Big) \Big[u(X, Y|\beta_{0j}^*, a_j^*, b_j^*, \sigma_j^*) - v(X, Y|\beta_{1j}^*) \Big],
\end{aligned}
$$

where we define $u(X, Y|\beta_1, a, b, \sigma) := \exp(\beta_1^\top X) f(Y|a^\top X + b, \sigma)$ and $v(X, Y|\beta_1) := \exp(\beta_1^\top X) g_{G_n}(Y|X)$. Next, for each $j \in [k_*]$ and $i \in \mathcal{A}_j$, we apply the Taylor expansions to the functions $u(X, Y|\beta_{1i}^n, a_i^n, b_i^n, \sigma_i^n)$ and $v(X, Y|\beta_{1i}^n)$ up to orders r_{1j} and r_{2j} (which we will choose later), respectively, as follows:

$$
\begin{aligned}
& u(X, Y|\beta_{1i}^n, a_i^n, b_i^n, \sigma_i^n) - u(X, Y|\beta_{1j}^*, a_j^*, b_j^*, \sigma_j^*) \\
& = \sum_{|\ell_1| + \ell_2 = 1}^{2r_{1j}} \sum_{\alpha \in \mathcal{I}_{\ell_1, \ell_2}} t_{\ell_1, \ell_2}(j) X^{\ell_1} \exp((\beta_{1j}^*)^\top X) \frac{\partial^{\ell_2} f}{\partial h_1^{\ell_2}}(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) + R_{1ij}(X, Y), \\
& v(X, Y|\beta_{1i}^n) - v(X, Y|\beta_{1j}^*) = \sum_{|\gamma| = 1}^{r_{2j}} s_\gamma(j) X^\gamma \exp((\beta_{1j}^*)^\top X) g_{G_n}(Y|X) + R_{2ij}(X, Y),
\end{aligned}
$$

where $R_{1ij}(X, Y)$ and $R_{2ij}(X, Y)$ are Taylor remainders such that $R_{\rho ij}(X, Y)/\mathcal{D}(G_n, G_*)$ vanishes as $n \to \infty$ for $\rho \in \{1, 2\}$. As a result, the limit of $Q_n/\mathcal{D}(G_n, G_*)$ when n goes to infinity can be seen as a linear combination of elements of the following set:

$$
\begin{aligned}
\mathcal{W} := & \left\{ X^{\ell_1} \exp((\beta_{1j}^*)^\top X) \frac{\partial^{\ell_2} f}{\partial h_1^{\ell_2}}(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) : j \in [k_*],\ 0 \leq 2|\ell_1| + \ell_2 \leq 2r_{1j} \right\} \\
& \cup \left\{ X^\gamma \exp((\beta_{1j}^*)^\top X) g_{G_*}(Y|X) : j \in [k_*],\ 0 \leq |\gamma| \leq r_{2j} \right\},
\end{aligned}
$$

which is shown to be linearly independent. By the Fatou's lemma, we demonstrate that $Q_n/\mathcal{D}(G_n, G_*)$ goes to zero as $n \to \infty$, implying that all the coefficients in the representation of $Q_n/\mathcal{D}(G_n, G_*)$, denoted by $T_{\ell_1, \ell_2}(j)/\mathcal{D}(G_n, G_*)$ and $S_\gamma(j)/\mathcal{D}(G_n, G_*)$, vanish when $n \to \infty$. Given that result, we aim to select the Taylor orders r_{1j} and r_{2j} such that at least one among the limits of $T_{\ell_1, \ell_2}(j)/\mathcal{D}(G_n, G_*)$ and $S_\gamma(j)/\mathcal{D}(G_n, G_*)$ is different from zero, which leads to a contradiction. Hence, we obtain the local version of the desired inequality. Below are the details of choosing appropriate Taylor orders in each setting.

Exact-fitted settings: Under this setting, since k_* is known, each of the Voronoi cells $\mathcal{A}_j$ for $j \in [k_*]$ has only one element. Thus, for any $i \in \mathcal{A}_j$, we have $\exp(\beta_{0i}^n) \to \exp(\beta_{0j}^*)$ and $(\beta_{1i}^n, a_i^n, b_i^n, \sigma_i^n) \to (\beta_{1j}^*, a_j^*, b_j^*, \sigma_j^*)$. Given that result, we can select $r_{1j} = r_{2j} = 1$ for all $j \in [k_*]$ as it suffices to show that at least one among the limits of $T_{\ell_1,\ell_2}(j)/\mathcal{D}(G_n, G_*)$ and $S_\gamma(j)/\mathcal{D}(G_n, G_*)$ is different from zero. In particular, if all of them vanish, we would take the sum of all the limits of $T_{\ell_1,\ell_2}(j)/\mathcal{D}(G_n, G_*)$ for (ℓ_1, ℓ_2) such that $0 \leq 2|\ell_1| + \ell_2 \leq 2$, which leads to a contradiction that $1 = \mathcal{D}(G_n, G_*)/\mathcal{D}(G_n, G_*) \to 0$.

Over-fitted settings: As k_* becomes unknown in this scenario, we need higher Taylor orders to obtain the same result as in the exact-fitted setting. We will reuse the proof by contradiction method to figure out those orders. More specifically, assume that all the limits of $T_{\ell_1,\ell_2}(j)/\mathcal{D}(G_n, G_*)$ and $S_\gamma(j)/\mathcal{D}(G_n, G_*)$ equal zero. After some steps of considering typical limits as in the previous setting which requires $r_{2j} = 2$ for all $j \in [k_*]$, we encounter the following system of polynomial equations:

$$\sum_{l=1}^{|\mathcal{A}_j|} \sum_{(\alpha_1,\alpha_2,\alpha_3,\alpha_4) \in \mathcal{I}_{\ell_1,\ell_2}} \frac{1}{\alpha!} p_{5l}^2 p_{1l}^{\alpha_1} p_{2l}^{\alpha_2} p_{3l}^{\alpha_3} p_{4l}^{\alpha_4} = 0,$$

for all (ℓ_1, ℓ_2) such that $1 \leq |\ell_1| + \ell_2 \leq r_{1j}$ for some $j \in [k_*]$. Due to the construction of this system, it must have at least one non-trivial solution. Therefore, if we choose $r_{1j} = \bar{r}(|\mathcal{A}_j|)$ for all $j \in [k_*]$, then the above system does not admit any non-trivial solutions, which leads to a contradiction.

Global version: The local inequality suggests that there exists a positive constant $\varepsilon' > 0$ such that $\inf_{G \in \mathcal{O}_k(\Theta), \mathcal{D}(G,G_*) \leq \varepsilon'} V(g_G, g_{G_*})/\mathcal{D}(G, G_*) > 0$. Thus, it suffices to show the following version of this inequality $\inf_{G \in \mathcal{O}_k(\Theta), \mathcal{D}(G,G_*) > \varepsilon'} V(g_G, g_{G_*})/\mathcal{D}(G, G_*) > 0$. Assume that this claim is not true, then we can find a mixing measure $G' \in \mathcal{O}_k(\Theta)$ such that $g_{G'}(Y|X) = g_{G_*}(Y|X)$ for almost surely (X, Y). According to Proposition 1, we get that $\mathcal{D}(G', G_*) = 0$, which contradicts the hypothesis $\mathcal{D}(G', G_*) > \varepsilon'$. These arguments hold for both exact-fitted and over-fitted settings up to some changes of notations.

4 Conclusion

In the paper, we study the convergence rates of parameter estimation under both the exact-fitted and over-fitted settings of the softmax gating Gaussian mixture of experts. We introduce novel Voronoi loss functions among parameters to resolve fundamental theoretical challenges posed by softmax gating, including identifiability up to the translation of parameters, the interaction between softmax weight and expert functions, and dependence between the numerator and denominator of the conditional density function. When the number of experts is known, we demonstrate that the rates of estimating true parameters are parametric. On the other hand, when the number of experts is unknown and overspecified, these rates are determined by a solvability of a system of polynomial equations.

A Proofs of Main Results

In this appendix, we provide proofs for Theorems 1 and 2.

A.1 Proof of Theorem 1

General Picture: It is worth noting that given the bound in equation (7), we can directly deduce the result in equation (8) from Proposition 2. Moreover, since the Hellinger distance h is lower bounded by the total variation distance V, we only need to show that

$$V(g_G, g_{G_*}) \geq C \cdot \mathcal{D}_1(G, G_*). \tag{11}$$

to obtain the bound in equation (7).

Local version: Firstly, we prove the local version of the above inequality, i.e., we will verify that

$$\lim_{\varepsilon \to 0} \inf_{\substack{G \in \mathcal{E}_{k_*}(\Theta), \\ \mathcal{D}_1(G, G_*) \leq \varepsilon}} V(g_G, g_{G_*}) / \mathcal{D}_1(G, G_*) > 0. \tag{12}$$

Suppose that the inequality in equation (12) does not hold, then we can find a sequence $G_n := \sum_{i=1}^{k_*} \exp(\beta_{0i}^n) \delta_{(\beta_{1i}^n, a_i^n, b_i^n, \sigma_i^n)} \in \mathcal{E}_{k_*}(\Theta)$ such that $V(g_{G_n}, g_{G_*}) / \mathcal{D}_1(G_n, G_*) \to 0$ and $\mathcal{D}_1(G_n, G_*) \to 0$ as $n \to \infty$. Next, for each $j \in [k_*]$, let us define the Voronoi cells corresponding to the mixing measure G_n as follows:

$$\mathcal{A}_j^n = \mathcal{A}_j(G_n) = \{i \in [k_*] : \|\theta_i^n - \theta_j^*\| \leq \|\theta_i^n - \theta_\ell^*\|, \forall \ell \neq j\|\},$$

where $\theta_i^n := (a_i^n, b_i^n, \sigma_i^n)$ and $\theta_j^* := (a_j^*, b_j^*, \sigma_j^*)$. As the number of distinct sets $\mathcal{A}_1^n \times \ldots \times \mathcal{A}_{k_*}^n$ is finite, there exist a subsequence of G_n such that $\mathcal{A}_j = \mathcal{A}_j^n$, i.e., the Voronoi cells are independent of n, for all $j \in [k_*]$. Since the argument in this proof is asymptotic, without loss of generality we assume that these Voronoi cells are independent of n for all n. Additionally, since k_* is known under the exact-fitted setting and $\mathcal{D}_1(G_n, G_*) \to 0$, the Voronoi cells $\mathcal{A}_j$ has only one element for any $j \in [k_*]$. Without loss of generality, we assume that $\mathcal{A}_j = \{j\}$ for all $j \in [k_*]$, i.e., $(a_j^n, b_j^n, \sigma_j^n) \to (a_j^*, b_j^*, \sigma_j^*)$ as $n \to \infty$. Furthermore, there exist t_1 and t_2 independent of n and a subsequence of G_n, which we again assume without loss of generality to hold for all n, such that $\exp(\beta_{0j}^n) \to \exp(\beta_{0j}^* + t_1)$ and $\beta_{1j}^n \to \beta_{1j}^* + t_2$ as n approaches infinity for all $j \in [k_*]$. It indicates that we can upper bound the Voronoi loss function $\mathcal{D}_1(G_n, G_*)$ as follows:

$$\begin{aligned}
\mathcal{D}_1(G_n, G_*) &\leq \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n) \|(\Delta_{t_2} \beta_{1ij}^n, \Delta a_{ij}^n, \Delta b_{ij}^n, \Delta \sigma_{ij}^n)\| + |\sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^* + t_1)| \\
&:= \mathcal{D}_1'(G_n, G_*).
\end{aligned}$$

As $V(g_{G_n}, g_{G_*}) / \mathcal{D}_1(G_n, G_*) \to 0$, we obtain that $V(g_{G_n}, g_{G_*}) / \mathcal{D}_1'(G_n, G_*) \to 0$.

Step 1: Decomposition

Subsequently, we consider $Q_n := [\sum_{j=1}^{k_*} \exp((\beta_{1j}^* + t_2)^\top X + \beta_{0j}^* + t_1)] \cdot [g_{G_n}(Y|X) - g_{G_*}(Y|X)]$,

which is decomposed as

$$
\begin{aligned}
Q_n &= \sum_{j=1}^{k_*}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\Big[u(X,Y|\beta_{1i}^n,a_i^n,b_i^n,\sigma_i^n)-u(X,Y|\beta_{1j}^*+t_2,a_j^*,b_j^*,\sigma_j^*)\Big]\\
&\quad-\sum_{j=1}^{k_*}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\Big[v(X,Y|\beta_{1i}^n)-v(X,Y|\beta_{1j}^*+t_2)\Big]\\
&\quad+\sum_{j=1}^{k_*}\Big(\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)-\exp(\beta_{0j}^*+t_1)\Big)\Big[u(X,Y|\beta_{1j}^*+t_2,a_j^*,b_j^*,\sigma_j^*)-v(X,Y|\beta_{1j}^*+t_2)\Big],\\
&:= A_n+B_n+E_n, \qquad (13)
\end{aligned}
$$

where we denote $u(X,Y|\beta_1,a,b,\sigma):=\exp(\beta_1^\top X)f(Y|a^\top X+b,\sigma)$ and $v(X,Y|\beta_1):=\exp(\beta_1^\top X)g_{G_n}(Y|X)$. Next, by means of the first-order Taylor expansion, we rewrite A_n as

$$
\begin{aligned}
A_n &= \sum_{j=1}^{k_*}\sum_{i\in\mathcal{A}_j}\sum_{|\alpha|=1}\frac{\exp(\beta_{0i}^n)}{2^{\alpha_4}\alpha!}(\Delta_{t_2}\beta_{1ij}^n)^{\alpha_1}(\Delta a_{ij}^n)^{\alpha_2}(\Delta b_{ij}^n)^{\alpha_3}(\Delta\sigma_{ij}^n)^{\alpha_4}\\
&\qquad\times X^{\alpha_1+\alpha_2}\exp((\beta_{1j}^*+t_2)^\top X)\cdot\frac{\partial^{|\alpha_2|+\alpha_3+2\alpha_4}f}{\partial h_1^{|\alpha_2|+\alpha_3+2\alpha_4}}(Y|(a_j^*)^\top X+b_j^*,\sigma_j^*)+R_1(X,Y)\\
&= \sum_{j=1}^{k_*}\sum_{i\in\mathcal{A}_j}\sum_{2|\ell_1|+\ell_2=1}^{2}\sum_{\alpha\in\mathcal{I}_{\ell_1,\ell_2}}\frac{\exp(\beta_{0i}^n)}{2^{\alpha_4}\alpha!}(\Delta_{t_2}\beta_{1ij}^n)^{\alpha_1}(\Delta a_{ij}^n)^{\alpha_2}(\Delta b_{ij}^n)^{\alpha_3}(\Delta\sigma_{ij}^n)^{\alpha_4}\\
&\qquad\times X^{\ell_1}\exp((\beta_{1j}^*+t_2)^\top X)\cdot\frac{\partial^{\ell_2}f}{\partial h_1^{\ell_2}}(Y|(a_j^*)^\top X+b_j^*,\sigma_j^*)+R_1(X,Y), \qquad (14)
\end{aligned}
$$

where $R_1(X,Y)$ is a Taylor remainder such that $R_1(X,Y)/\mathcal{D}_1'(G_n,G_*)\to 0$ as $n\to\infty$. The last equality in the above equation is obtained by defining $\ell_1=\alpha_1+\alpha_2$, $\ell_2=|\alpha_2|+\alpha_3+\alpha_4$ and

$$
\mathcal{I}_{\ell_1,\ell_2}:=\Big\{\alpha=(\alpha_i)_{i=1}^4\in\mathbb{N}^d\times\mathbb{N}^d\times\mathbb{N}\times\mathbb{N}:\alpha_1+\alpha_2=\ell_1,\ \alpha_3+2\alpha_4=\ell_2-|\alpha_2|\Big\}, \qquad (15)
$$

for all $(\ell_1,\ell_2)\in\mathbb{R}^d\times\mathbb{R}$ such that $1\le 2|\ell_1|+\ell_2\le 2$. Analogously, B_n can be rewritten as

$$
B_n=-\sum_{j=1}^{k_*}\sum_{i\in\mathcal{A}_j}\sum_{|\gamma|=1}\frac{\exp(\beta_{0i}^n)}{\gamma!}(\Delta_{t_2}\beta_{1ij}^n)^{\gamma}X^{\gamma}\exp((\beta_{1j}^*+t_2)^\top X)g_{G_n}(Y|X)+R_2(X,Y), \qquad (16)
$$

where $R_2(X,Y)$ is a Taylor remainder such that $R_2(X,Y)/\mathcal{D}_1'(G_n,G_*)\to 0$ as $n\to\infty$. From the formulations of A_n, B_n and E_n, we can represent Q_n as the following linear combination

$$
\begin{aligned}
Q_n &= \sum_{j=1}^{k_*}\sum_{2|\ell_1|+\ell_2=0}^{2}T_{\ell_1,\ell_2}(j)\cdot X^{\ell_1}\exp((\beta_{1j}^*+t_2)^\top X)\frac{\partial^{\ell_2}f}{\partial h_1^{\ell_2}}(Y|(a_j^*)^\top X+b_j^*,\sigma_j^*)\\
&\quad+\sum_{j=1}^{k_*}\sum_{|\gamma|=0}^{1}S_\gamma(j)\cdot X^{\gamma}\exp((\beta_{1j}^*+t_2)^\top X)g_{G_n}(Y|X),
\end{aligned}
$$

with coefficients being denoted by $T_{\ell_1,\ell_2}(j)$ and $S_\gamma(j)$ for all $j \in [k_*]$, $0 \le 2|\ell_1| + \ell_2 \le 2$ and $0 \le |\gamma| \le 1$ where

$$T_{\ell_1,\ell_2}(j) = \begin{cases} \sum_{i\in\mathcal{A}_j} \sum_{\alpha\in\mathcal{I}_{\ell_1,\ell_2}} \frac{\exp(\beta_{0i}^n)}{2^{\alpha_4}\alpha!}(\Delta_{t_2}\beta_{1ij}^n)^{\alpha_1}(\Delta a_{ij}^n)^{\alpha_2}(\Delta b_{ij}^n)^{\alpha_3}(\Delta\sigma_{ij}^n)^{\alpha_4}, \\ \qquad\qquad\qquad\qquad\qquad\qquad (\ell_1,\ell_2) \neq (\mathbf{0}_d, 0) \\ \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^* + t_1), \qquad\qquad (\ell_1,\ell_2) = (\mathbf{0}_d, 0) \end{cases} \tag{17}$$

and

$$S_\gamma(j) = \begin{cases} -\sum_{i\in\mathcal{A}_j} \frac{\exp(\beta_{0i}^n)}{\gamma!}(\Delta_{t_2}\beta_{1ij}^n)^\gamma, \qquad |\gamma| \neq 0 \\ -\sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n) + \exp(\beta_{0j}^* + t_1), \quad |\gamma| = 0. \end{cases} \tag{18}$$

Step 2: Non-vanishing coefficients

Now, we will demonstrate by contradiction that at least one of the terms $T_{\ell_1,\ell_2}(j)/\mathcal{D}_1'(G_n, G_*)$, $S_\gamma(j)/\mathcal{D}_1'(G_n, G_*)$ does not approach zero. Indeed, assume that all of them vanish when $n \to \infty$, then we get

$$\sum_{j=1}^{k_*} \frac{|T_{0,0}(j)|}{\mathcal{D}_1'(G_n, G_*)} = \sum_{j=1}^{k_*} \frac{|\sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^* + t_1)|}{\mathcal{D}_1'(G_n, G_*)} \to 0,$$

which implies that

$$\frac{1}{\mathcal{D}_1'(G_n, G_*)} \cdot \sum_{j=1}^{k_*} \Big| \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^* + t_1) \Big| \to 0. \tag{19}$$

Similarly, by considering the limits of $T_{\ell_1,\ell_2}(j)/\mathcal{D}_1'(G_n, G_*)$ for all $j \in [k_*]$ and $1 \le 2|\ell_1| + \ell_2 \le 2$, we obtain that

$$\frac{1}{\mathcal{D}_1'(G_n, G_*)} \cdot \sum_{j=1}^{k_*} \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n) \|(\Delta_{t_2}\beta_{1ij}^n, \Delta a_{ij}^n, \Delta b_{ij}^n, \Delta\sigma_{ij}^n)\| \to 0. \tag{20}$$

Combine the results in equations (19) and (20), we have $1 = \mathcal{D}_1'(G_n, G_*)/\mathcal{D}_1'(G_n, G_*) \to 0$, which is a contradiction. As a result, not all the limits of $T_{\ell_1,\ell_2}(j)/\mathcal{D}_1'(G_n, G_*)$ and $S_\gamma(j)/\mathcal{D}_1'(G_n, G_*)$ equal to zero.

Step 3: Fatou's lemma involvement

Thus, let m_n be the maximum of the absolute values of those terms, we have that $1/m_n \not\to \infty$. Subsequently, the Fatou's lemma says that

$$\lim_{n\to\infty} \frac{1}{m_n} \cdot \frac{V(g_{G_n}, g_{G_*})}{\mathcal{D}_1'(G_n, G_*)} \ge \int \liminf_{n\to\infty} \frac{1}{m_n} \cdot \frac{|g_{G_n}(Y|X) - g_{G_*}(Y|X)|}{2\mathcal{D}_1'(G_n, G_*)} \, \mathrm{d}(X, Y). \tag{21}$$

By assumption, the left-hand side of the above equation equals to zero, therefore, the integrand in the right-hand side also equals to zero for almost surely (X, Y), which leads to the following limit: $Q_n/[m_n\mathcal{D}_1(G_n, G_*)] \to 0$ as $n \to \infty$ for almost surely (X, Y). More specifically, we have

$$\sum_{j=1}^{k_*} \sum_{2|\ell_1|+\ell_2=0}^{2} \eta_{\ell_1,\ell_2}(j) \cdot X^{\ell_1} \exp((\beta_{1j}^*)^\top X) \frac{\partial^{\ell_2} f}{\partial h_1^{\ell_2}}(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*)$$
$$+ \sum_{j=1}^{k_*} \sum_{|\gamma|=0}^{1} \omega_\gamma(j) \cdot X^\gamma \exp((\beta_{1j}^*)^\top X) g_{G_*}(Y|X) = 0,$$

for almost surely (X, Y), where $\eta_{\ell_1,\ell_2}(j)$ and $\omega_\gamma(j)$ are the limits of $T_{\ell_1,\ell_2}(j)/[m_n\mathcal{D}_1'(G_n, G_*)]$ and $S_\gamma(j)/[m_n\mathcal{D}_1'(G_n, G_*)]$, respectively, for all $j \in [k_*]$, $0 \leq 2|\ell_1| + \ell_2 \leq 2$ and $0 \leq |\gamma| \leq 1$. Here, at least one among $\eta_{\ell_1,\ell_2}(j)$ and $\omega_\gamma(j)$ is different from zero. On the other hand, since the set

$$\mathcal{W}_1 := \left\{ X^{\ell_1} \exp((\beta_{1j}^*)^\top X) \frac{\partial^{\ell_2} f}{\partial h_1^{\ell_2}}(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) : j \in [k_*],\ 0 \leq 2|\ell_1| + \ell_2 \leq 2 \right\}$$
$$\cup \left\{ X^\gamma \exp((\beta_{1j}^*)^\top X) g_{G_*}(Y|X) : j \in [k_*],\ 0 \leq |\gamma| \leq 1 \right\}, \tag{22}$$

is linearly independent (see Lemma 2 at the end of this proof), we obtain that $\eta_{\ell_1,\ell_2}(j) = \omega_\gamma(j) = 0$ for all $j \in [k_*]$, $0 \leq 2|\ell_1| + \ell_2 \leq 2$ and $0 \leq |\gamma| \leq 1$, which is a contradiction. Thus, we reach the local inequality in equation (12), that is, there exists $\varepsilon' > 0$ that satisfies

$$\inf_{\substack{G \in \mathcal{E}_{k_*}(\Theta), \\ \mathcal{D}_1(G_n, G_*)}} V(g_G, g_{G_*})/\mathcal{D}_1(G, G_*) > 0.$$

Then, in order to obtain the conclusion in equation (11), it suffices to prove its following global version:

Global version:

$$\inf_{\substack{G \in \mathcal{E}_{k_*}(\Theta), \\ \mathcal{D}_1(G, G_*) > \varepsilon'}} V(g_G, g_{G_*})/\mathcal{D}_1(G, G_*) > 0. \tag{23}$$

Assume by contrary that there exists a sequence $G_n' \in \mathcal{E}_{k_*}(\Theta)$ that satisfies

$$\begin{cases} \lim_{n\to\infty} V(g_{G_n'}, g_{G_*})/\mathcal{D}_1(G_n', G_*) = 0, \\ \mathcal{D}_1(G_n', G_*) > \varepsilon'. \end{cases}$$

Therefore, we obtain that $V(g_{G_n'}, g_{G_*}) \to 0$ as $n \to \infty$. Since the set Θ is compact, we are able to replace the sequence G_n' by its subsequence which converges to some mixing measure $G' \in \mathcal{E}_{k_*}(\Theta)$ such that $\mathcal{D}(G', G_*) > \varepsilon'$. Then, by the Fatou's lemma, we get

$$\lim_{n\to\infty} V(g_{G_n'}, g_{G_*}) \geq \frac{1}{2} \int \liminf_{n\to\infty} |g_{G_n'}(Y|X) - g_{G_*}(Y|X)| \, \mathrm{d}(X, Y),$$

which implies that

$$\int |g_{G'}(Y|X) - g_{G_*}(Y|X)| \, \mathrm{d}(X, Y) = 0$$

Thus, we obtain that $g_{G'}(Y|X) = g_{G_*}(Y|X)$ for almost surely (X, Y). Now that the softmax gating Gaussian mixture of experts is identifiable up to a translation (see Proposition 1), the mixing measure G' admits the form $G' = \sum_{i=1}^{k_*} \exp(\beta^*_{0\tau(i)} + t_1)\delta_{(\beta^*_{1\tau(i)}+t_2, a^*_{\tau(i)}, b^*_{\tau(i)}, \sigma^*_{\tau(i)})}$ for some t_1, t_2, where τ is some permutation of the set $\{1, 2, \ldots, k\}$. This leads to $\mathcal{D}_1(G', G_*) = 0$, which contradicts the hypothesis that $\mathcal{D}_1(G', G_*) > \varepsilon' > 0$. Hence, we obtain the inequality in equation (11).

To complete the proof, we will show the previous claim regarding the independence of elements in $\mathcal{W}_1$ in the following lemma:

Lemma 2. *The set $\mathcal{W}_1$ defined in equation (22) is linearly independent.*

Proof of Lemma 2. Assume that we have the following equality for almost surely (X, Y):

$$\sum_{j=1}^{k_*} \sum_{2|\ell_1|+\ell_2=0}^{2} \eta_{\ell_1,\ell_2}(j) \cdot X^{\ell_1} \exp((\beta^*_{1j})^\top X) \frac{\partial^{\ell_2} f}{\partial h_1^{\ell_2}}(Y|(a^*_j)^\top X + b^*_j, \sigma^*_j)$$
$$+ \sum_{j=1}^{k_*} \sum_{|\gamma|=0}^{1} \omega_\gamma(j) \cdot X^\gamma \exp((\beta^*_{1j})^\top X) g_{G_*}(Y|X) = 0,$$

where $\eta_{\ell_1,\ell_2}(j) \in \mathbb{R}$ and $\omega_\gamma(j) \in \mathbb{R}$, we need to show that $\eta_{\ell_1,\ell_2}(j) = \omega_\gamma(j) = 0$, for all $j \in [k_*]$, $0 \leq 2|\ell_1| + \ell_2 \leq 2$ and $0 \leq |\gamma| \leq 1$. The above equation is equivalent to

$$\sum_{j=1}^{k_*} \sum_{\zeta=0}^{1} \Big[\sum_{\ell_2=0}^{2-2\zeta} \eta_{\zeta,\ell_2}(j) \frac{\partial^{\ell_2} f}{\partial h_1^{\ell_2}}(Y|(a^*_j)^\top X + b^*_j, \sigma^*_j) + \omega_\zeta(j) g_{G_*}(Y|X) \Big] X^\zeta \exp((\beta^*_{1j})^\top X) = 0,$$

for almost surely (X, Y). Since $\beta^*_{11}, \ldots, \beta^*_{1k_*}$ are k_* distinct values, we get that the set $\Big\{\exp((\beta^*_{1j})^\top X) : j \in [k_*]\Big\}$ is linearly independent, which implies that

$$\sum_{\zeta=0}^{1} \Big[\sum_{\ell_2=0}^{2-2\zeta} \eta_{\zeta,\ell_2}(j) \frac{\partial^{\ell_2} f}{\partial h_1^{\ell_2}}(Y|(a^*_j)^\top X + b^*_j, \sigma^*_j) + \omega_\zeta(j) g_{G_*}(Y|X) \Big] X^\zeta = 0,$$

for all $j \in [k_*]$ for almost surely (X, Y). Obviously, the above equation is a polynomial of $X \in \mathcal{X}$, where $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$. Then, we achieve that

$$\sum_{\ell_2=0}^{2-2\zeta} \eta_{\zeta,\ell_2}(j) \frac{\partial^{\ell_2} f}{\partial h_1^{\ell_2}}(Y|(a^*_j)^\top X + b^*_j, \sigma^*_j) + \omega_\zeta(j) g_{G_*}(Y|X) = 0,$$

for all $j \in [k_*]$ and $\zeta \in \{0, 1\}$, for almost surely (X, Y). Again, as $(a^*_j, b^*_j, \sigma^*_j)$ for $j \in [k_*]$ are k_* distinct tuples, we have that $((a^*_j)^\top X + b^*_j, \sigma^*_j)$ for $j \in [k_*]$ are also k_* distinct tuples for almost surely X. Therefore, $\Big\{ \frac{\partial^{\ell_2} f}{\partial h_1^{\ell_2}}(Y|(a^*_j)^\top X + b^*_j, \sigma^*_j),\ g_{G_*}(Y|X) \Big\}$ is a linearly independent set. As a result, $\eta_{\ell_1,\ell_2}(j) = \omega_\gamma(j) = 0$ for all $j \in [k_*]$, $0 \leq 2|\ell_1| + \ell_2 \leq 2$ and $0 \leq |\gamma| \leq 1$. Hence, the proof is completed. □

A.2 Proof of Theorem 2

In this proof, we will adapt the framework in Appendix A.1 to the setting of Theorem 2. However, since the arguments utilized for the global version part remain the same (up to some changes of notations) for the over-fitted setting, they will not be presented here again and we focus only on proving the following local inequality by following the steps outlined in Appendix A.1:

$$\lim_{\varepsilon \to 0} \inf_{\substack{G \in \mathcal{O}_k(\Theta), \\ \mathcal{D}_2(G_n, G_*) \leq \varepsilon}} V(g_G, g_{G_*})/\mathcal{D}_2(G, G_*) > 0. \tag{24}$$

Assume that the above claim is not true, then there exists a sequence $G_n := \sum_{i=1}^{k_n} \exp(\beta_{0i}^n)\delta_{(\beta_{1i}^n, a_i^n, b_i^n, \sigma_i^n)} \in \mathcal{O}_k(\Theta)$ such that both terms $V(g_{G_n}, g_{G_*})/\mathcal{D}_2(G_n, G_*)$ and $\mathcal{D}_2(G_n, G_*)$ vanish to 0 when n tends to infinity. As $k_n \leq k$ for all $n \in \mathbb{N}$, we are able to substitute the sequence G_n with its subsequence which has the number of atoms $k_n = k' \leq k_*$ being independent of n. Since the proof argument is asymptotic, we also assume that $k_n = k'$ for all $n \geq 1$. Following the proof argument of Theorem 1 in Appendix A.1, we also assume that the Voronoi cells $\mathcal{A}_j = \mathcal{A}_j^n$ does not change with n for all $j \in [k_*]$. Furthermore, for any $(a_i^n, b_i^n, \sigma_i^n)$ that $i \in \mathcal{A}_j$, we have $(a_i^n, b_i^n, \sigma_i^n) \to (a_j^*, b_j^*, \sigma_j^*)$ as n approaches infinity. Furthermore, there exist t_1, t_2 such that $\exp(\beta_{0i}^n) \to \exp(\beta_{0j}^* + t_1)$ and $\beta_{1i}^n \to \beta_{1j}^* + t_2$ for any $i \in \mathcal{A}_j$ and $j \in [k_*]$. Given t_1, t_2, we can upper bound the Voronoi loss function $\mathcal{D}_2(G_n, G_*)$ as follows:

$$\begin{aligned}
\mathcal{D}_2(G_n, G_*) &\leq \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n)(\|(\Delta_{t_2}\beta_{1ij}^n, \Delta b_{ij}^n)\|^{\bar{r}(|\mathcal{A}_j|)} + \|(\Delta a_{ij}^n, \Delta\sigma_{ij}^n)\|^{\bar{r}(|\mathcal{A}_j|)/2}) \\
&+ \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n)\|(\Delta_{t_2}\beta_{1ij}, \Delta a_{ij}^n, \Delta b_{ij}^n, \Delta\sigma_{ij}^n)\| + \sum_{j=1}^{k_*} |\sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^* + t_1)| \\
&:= \mathcal{D}_2'(G_n, G_*).
\end{aligned}$$

Since $V(g_{G_n}, g_{G_*})/\mathcal{D}_2(G_n, G_*) \to 0$, the above inequality leads to $V(g_{G_n}, g_{G_*})/\mathcal{D}_2'(G_n, G_*) \to 0$ as $n \to \infty$.

Step 1: Decomposition

In this step, we reuse the decomposition $Q_n = A_n + B_n + E_n$ in equation (13). However, under the over-fitted setting, since there are some Voronoi cells $\mathcal{A}_j$ possibly having more than one element, we continue to decompose A_n and B_n as follows:

$$\begin{aligned}
A_n &= \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n)\Big[u(X, Y|\beta_{1i}^n, a_i^n, b_i^n, \sigma_i^n) - u(X, Y|\beta_{1j}^* + t_2, a_j^*, b_j^*, \sigma_j^*)\Big] \\
&+ \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n)\Big[u(X, Y|\beta_{1i}^n, a_i^n, b_i^n, \sigma_i^n) - u(X, Y|\beta_{1j}^* + t_2, a_j^*, b_j^*, \sigma_j^*)\Big] \\
&:= A_{n,1} + A_{n,2},
\end{aligned}$$

and

$$\begin{aligned}B_n = &- \sum_{j:|\mathcal{A}_j|=1} \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)\Big[v(X,Y|\beta_{1i}^n) - v(X,Y|\beta_{1j}^* + t_2)\Big] \\ &- \sum_{j:|\mathcal{A}_j|>1} \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)\Big[v(X,Y|\beta_{1i}^n) - v(X,Y|\beta_{1j}^* + t_2)\Big] \\ &:= B_{n,1} + B_{n,2}.\end{aligned}$$

Now, we apply the first-order Taylor expansions for two terms $A_{n,1}$ and $B_{n,1}$ as in equations (14) and (16), while for $A_{n,2}$ and $B_{n,2}$, we use the Taylor expansions of orders $\bar{r}(|\mathcal{A}_j|)$ and 2, respectively, for each $j : |\mathcal{A}_j| > 1$ as

$$\begin{aligned}A_{n,2} = &\sum_{j:|\mathcal{A}_j|>1} \sum_{i\in\mathcal{A}_j} \sum_{2|\ell_1|+\ell_2=1}^{2\bar{r}(|\mathcal{A}_j|)} \sum_{\alpha\in\mathcal{I}_{\ell_1,\ell_2}} \frac{\exp(\beta_{0i}^n)}{2^{\alpha_4}\alpha!}(\Delta_{t_2}\beta_{1ij}^n)^{\alpha_1}(\Delta a_{ij}^n)^{\alpha_2}(\Delta b_{ij}^n)^{\alpha_3}(\Delta\sigma_{ij}^n)^{\alpha_4} \\ &\times X^{\ell_1}\exp((\beta_{1j}^* + t_2)^\top X)\cdot\frac{\partial^{\ell_2} f}{\partial h_1^{\ell_2}}(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) + R_3(X,Y), \\ B_{n,2} = &-\sum_{j:|\mathcal{A}_j|>1} \sum_{i\in\mathcal{A}_j} \sum_{|\gamma|=1}^{2} \frac{\exp(\beta_{0i}^n)}{\gamma!}(\Delta_{t_2}\beta_{1ij}^n)^\gamma X^\gamma \exp((\beta_{1j}^* + t_2)^\top X) g_{G_n}(Y|X) + R_4(X,Y),\end{aligned}$$

where $\mathcal{I}_{\ell_1,\ell_2}$ is defined in equation (15) and $R_3(X,Y)$, $R_4(X,Y)$ are Taylor remainders such that $R_p(X,Y)/\mathcal{D}_2'(G_n,G_*) \to 0$ when $n\to\infty$ for $p\in\{3,4\}$. As a result, Q_n can represented as

$$\begin{aligned}Q_n = &\sum_{j=1}^{k_*} \sum_{2|\ell_1|+\ell_2=0}^{2\bar{r}(|\mathcal{A}_j|)} T_{\ell_1,\ell_2}(j)\cdot X^{\ell_1}\exp((\beta_{1j}^* + t_2)^\top X)\frac{\partial^{\ell_2} f}{\partial h_1^{\ell_2}}(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) \\ &+ \sum_{j=1}^{k_*} \sum_{|\gamma|=0}^{1+\mathbf{1}_{\{|\mathcal{A}_j|>1\}}} S_\gamma(j)\cdot X^\gamma \exp((\beta_{1j}^* + t_2)^\top X) g_{G_n}(Y|X),\end{aligned} \tag{25}$$

where $T_{\ell_1,\ell_2}(j)$ and $S_\gamma(j)$ are defined in equations (17) and (18).

Step 2: Non-vanishing coefficients

Next, we will show that not all the quantities $T_{\ell_1,\ell_2}(j)/\mathcal{D}_2'(G_n,G_*)$ and $S_\gamma(j)/\mathcal{D}_2'(G_n,G_*)$ go to 0 as $n\to\infty$. Assume that all of them vanish when n tends to infinity. Then, by arguing similarly as in equations (19) and (20), we obtain that

$$\begin{aligned}\frac{1}{\mathcal{D}_2'(G_n,G_*)}\cdot\Big[&\sum_{j=1}^{k_*}\Big|\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n) - \exp(\beta_{0j}^* + t_1)\Big| \\ &+ \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\|(\Delta_{t_2}\beta_{1ij}^n, \Delta a_{ij}^n, \Delta b_{ij}^n, \Delta\sigma_{ij}^n)\|\Big] \to 0.\end{aligned}$$

Putting the above limit and the formulation of $\mathcal{D}_2(G_n,G_*)$ together, we deduce that

$$\frac{1}{\mathcal{D}_2'(G_n,G_*)}\cdot\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i})(\|\Delta_{t_2}\beta_{1ij}^n, \Delta b_{ij}^n\|^{\bar{r}(|\mathcal{A}_j|)} + \|(\Delta a_{ij}^n, \Delta\sigma_{ij}^n\|^{\bar{r}(|\mathcal{A}_j|)/2}) \not\to 0,$$

which indicates that there exists some index $j^* \in [k_*]$ such that $|\mathcal{A}_{j^*}| > 1$ and

$$\frac{1}{\mathcal{D}_2'(G_n, G_*)} \cdot \sum_{i \in \mathcal{A}_{j^*}} \exp(\beta_{0i}) \|(\Delta_{t_2} \beta_{1ij^*}^n, \Delta b_{ij^*}^n)\|^{\bar{r}(|\mathcal{A}_j|)} + \|(\Delta a_{ij^*}^n, \Delta \sigma_{ij^*}^n)\|^{\bar{r}(|\mathcal{A}_j|)/2} \not\to 0,$$

for all $t_2 \in \mathbb{R}^d$. Without loss of generality, we may assume that $j^* = 1$. Recall that for (ℓ_1, ℓ_2) such that $1 \leq |\ell_1| + \ell_2 \leq \bar{r}(|\mathcal{A}_1|)$, we have $T_{\ell_1,\ell_2}(1)/\mathcal{D}_2'(G_n, G_*) \to 0$ as $n \to \infty$. Thus, by dividing this ratio and the left hand side of the above equation and let $t_2 = 0$, we obtain that

$$\frac{\sum_{i \in \mathcal{A}_1} \sum_{\alpha \in \mathcal{I}_{\ell_1,\ell_2}} \frac{\exp(\beta_{0i}^n)}{2^{\alpha_4} \alpha!} (\Delta \beta_{1i1}^n)^{\alpha_1} (\Delta a_{i1}^n)^{\alpha_2} (\Delta b_{i1}^n)^{\alpha_3} (\Delta \sigma_{i1}^n)^{\alpha_4}}{\sum_{i \in \mathcal{A}_1} \exp(\beta_{0i}^n) (\|(\Delta \beta_{1i1}^n, \Delta b_{i1}^n)\|^{\bar{r}(|\mathcal{A}_1|)} + \|(\Delta a_{i1}^n, \Delta \sigma_{i1}^n)\|^{\bar{r}(|\mathcal{A}_1|)/2})} \to 0, \tag{26}$$

for all (ℓ_1, ℓ_2) such that $1 \leq |\ell_1| + \ell_2 \leq \bar{r}(|\mathcal{A}_1|)$.

Let us define $\overline{M}_n := \max\{\|\Delta \beta_{1i1}^n\|, \|\Delta a_{i1}^n\|^{1/2}, |\Delta b_{i1}^n|, |\Delta \sigma_{i1}^n|^{1/2} : i \in \mathcal{A}_1\}$ and $\overline{\beta}_n := \max_{i \in \mathcal{A}_1} \exp(\beta_{0i}^n)$. Note that the sequence $\exp(\beta_{0i}^n)/\overline{\beta}_n$ is bounded, we will replace it by its subsequence that has a positive limit $p_{5i}^2 := \lim_{n \to \infty} \exp(\beta_{0i}^n)/\overline{\beta}_n$. Thus, at least one among p_{5i}^2 for $i \in \mathcal{A}_1$ equals 1.

In addition, we also define

$$\begin{aligned} (\Delta \beta_{1i1}^n)/\overline{M}_n &\to p_{1i}, \quad (\Delta a_{i1}^n)/\overline{M}_n \to p_{2i}, \\ (\Delta b_{i1}^n)/\overline{M}_n &\to p_{3i}, \quad (\Delta \sigma_{i1}^n)/[2\overline{M}_n] \to p_{4i}. \end{aligned}$$

Here, at least one of p_{1i}, p_{2i}, p_{3i} and p_{4i} for $i \in \mathcal{A}_1$ equals either 1 or -1. Next, we divide both the numerator and the denominator of the ratio in equation (26) by $\overline{\beta}_n \overline{M}_n^{\ell_1 + \ell_2}$, we achieve the following system of polynomial equations:

$$\sum_{i \in \mathcal{A}_1} \sum_{\alpha \in \mathcal{I}_{\ell_1,\ell_2}} \frac{1}{\alpha!} \cdot p_{5i}^2 p_{1i}^{\alpha_1} p_{2i}^{\alpha_2} p_{3i}^{\alpha_3} p_{4i}^{\alpha_4} = 0,$$

for all (ℓ_1, ℓ_2) such that $1 \leq |\ell_1| + \ell_2 \leq \bar{r}(|\mathcal{A}_1|)$. However, based on definition of $\bar{r}(|\mathcal{A}_1|)$, the above system do not have any non-trivial solutions, which is a contradiction. Thus, not all the quantities $T_{\ell_1,\ell_2}(j)/\mathcal{D}_2'(G_n, G_*)$ and $S_\gamma(j)/\mathcal{D}_2'(G_n, G_*)$ go to 0 as $n \to \infty$.

Step 3: Fatou's lemma involvement

Subsequently, we denote by m_n be the maximum of the absolute values of those quantities. Based on the result in Step 2, we know that $1/m_n \not\to \infty$. Then, by applying the Fatou's lemma as in equation (21), we get that $Q_n/[m_n \mathcal{D}_2'(G_n, G_*)] \to 0$ as $n \to \infty$ for almost surely (X, Y). It follows from the decomposition of Q_n in equation (25) that

$$\begin{aligned} &\sum_{j=1}^{k_*} \sum_{2|\ell_1| + \ell_2 = 0}^{2\bar{r}(|\mathcal{A}_j|)} \eta_{\ell_1,\ell_2}(j) \cdot X^{\ell_1} \exp((\beta_{1j}^*)^\top X) \frac{\partial^{\ell_2} f}{\partial h_1^{\ell_2}}(Y | (a_j^*)^\top X + b_j^*, \sigma_j^*) \\ &+ \sum_{j=1}^{k_*} \sum_{|\gamma| = 0}^{1 + \mathbf{1}_{\{|\mathcal{A}_j| > 1\}}} \omega_\gamma(j) \cdot X^\gamma \exp((\beta_{1j}^*)^\top X) g_{G_*}(Y | X) = 0, \end{aligned}$$

for almost surely (X, Y), where $\eta_{\ell_1,\ell_2}(j)$ and $\omega_\gamma(j)$ denote the limits of $T_{\ell_1,\ell_2}(j)/[m_n\mathcal{D}_2'(G_n, G_*)]$ and $S_\gamma(j)/[m_n\mathcal{D}_2'(G_n, G_*)]$ as $n \to \infty$, respectively, for all $j \in [k_*]$, $0 \leq 2|\ell_1| + \ell_2 \leq 2\bar{r}(|\mathcal{A}_j|)$ and $0 \leq |\gamma| \leq 1 + \mathbf{1}_{\{|\mathcal{A}_j|>1\}}$. By definition, at least one among $\eta_{\ell_1,\ell_2}(j)$ and $\omega_\gamma(j)$ is different from zero. Nevertheless, as the set

$$\mathcal{W}_2 := \left\{ X^{\ell_1} \exp((\beta_{1j}^*)^\top X) \frac{\partial^{\ell_2} f}{\partial h_1^{\ell_2}}(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) : j \in [k_*],\ 0 \leq 2|\ell_1| + \ell_2 \leq 2\bar{r}(|\mathcal{A}_j|) \right\}$$
$$\cup \left\{ X^\gamma \exp((\beta_{1j}^*)^\top X) g_{G_*}(Y|X) : j \in [k_*],\ 0 \leq |\gamma| \leq 1 + \mathbf{1}_{\{|\mathcal{A}_j|>1\}} \right\}, \tag{27}$$

is linearly independent (proof can be done similarly to Lemma 2), it follows that $\eta_{\ell_1,\ell_2}(j) = \omega_\gamma(j) = 0$ for all $j \in [k_*]$, $0 \leq 2|\ell_1| + \ell_2 \leq 2\bar{r}(|\mathcal{A}_j|)$ and $0 \leq |\gamma| \leq 1 + \mathbf{1}_{\{|\mathcal{A}_j|>1\}}$, which is a contradiction. Hence, we achieve the inequality in equation (24), and complete the proof.

B Proofs of Auxiliary Results

In this appendix, we provide proofs for the remaining results of the paper.

B.1 Proof of Proposition 1

Given the notations in Proposition 1, assume that the equation $g_G(Y|X) = g_{G'}(Y|X)$ holds true, that is,

$$\sum_{i=1}^{k} \frac{\exp((\beta_{1i})^\top X + \beta_{0i})}{\sum_{j=1}^{k} \exp((\beta_{1j})^\top X + \beta_{0j})} f(Y|(a_i)^\top X + b_i, \sigma_i)$$
$$= \sum_{i=1}^{k'} \frac{\exp((\beta_{1i}')^\top X + \beta_{0i}')}{\sum_{j=1}^{k} \exp((\beta_{1j}')^\top X + \beta_{0j}')} f(Y|(a_i')^\top X + b_i', \sigma_i'), \tag{28}$$

for almost surely (X, Y). Then, it follows from the identifiability of the location-scale Gaussian mixtures that the number of atoms and the weight set of the mixing measure G equal to those of its counterpart G', i.e. $k = k'$ and

$$\left\{ \frac{\exp((\beta_{1i})^\top X + \beta_{0i})}{\sum_{j=1}^{k} \exp((\beta_{1j})^\top X + \beta_{0j})} : i \in [k] \right\} \equiv \left\{ \frac{\exp((\beta_{1i}')^\top X + \beta_{0i}')}{\sum_{j=1}^{k} \exp((\beta_{1j}')^\top X + \beta_{0j}')} : i \in [k] \right\},$$

for almost surely X. For simplicity, we may assume that

$$\frac{\exp((\beta_{1i})^\top X + \beta_{0i})}{\sum_{j=1}^{k} \exp((\beta_{1j})^\top X + \beta_{0j})} = \frac{\exp((\beta_{1i}')^\top X + \beta_{0i}')}{\sum_{j=1}^{k} \exp((\beta_{1j}')^\top X + \beta_{0j}')},$$

for all $i \in [k]$. Since the softmax function is invariant to translation, we get that $\beta_{0i} = \beta_{0i}' + t_1$ and $\beta_{1i} = \beta_{1i}' + t_2$ for some $t_1 \in \mathbb{R}$ and $t_2 \in \mathbb{R}^d$. Therefore, equation (28) reduces to

$$\sum_{i=1}^{k} \exp(\beta_{0i}) u(X, Y|\beta_{1i}, a_i, b_i, \sigma_i) = \sum_{i=1}^{k} \exp(\beta_{0i}) u(X, Y|\beta_{1i}, a_i', b_i', \sigma_i')), \tag{29}$$

for almost surely (X, Y), where $u(X, Y|\beta_1, a, b, \sigma) := \exp(\beta_1^\top X) f(Y|a^\top X + b, \sigma)$ for all $i \in [k]$. Next, we will partition the index set $[k]$ into q subsets $U_1, U_2, \ldots, U_q$ such that for each $\ell \in [q]$, we have $\exp(\beta_{0i}) = \exp(\beta_{0i'})$ for any $i, i' \in U_\ell$. As a result, equation (29) can be rewritten as

$$\sum_{\ell=1}^{q} \sum_{i \in U_\ell} \exp(\beta_{0i}) u(X, Y|\beta_{1i}, a_i, b_i, \sigma_i) = \sum_{\ell=1}^{q} \sum_{i \in U_\ell} \exp(\beta_{0i}) u(X, Y|\beta_{1i}, a'_i, b'_i, \sigma'_i),$$

for almost surely (X, Y). Given the above equation, for each $\ell \in [q]$, we obtain that

$$\left\{((a_i)^\top X + b_i, \sigma_i) : i \in U_\ell\right\} \equiv \left\{((a'_i)^\top X + b'_i, \sigma'_i) : i \in U_\ell\right\},$$

for almost surely X, which directly leads to

$$\{(a_i, b_i, \sigma_i) : i \in U_\ell\} \equiv \left\{(a'_i, b'_i, \sigma'_i) : i \in U_\ell\right\}.$$

WLOG, we assume that $(a_i, b_i, \sigma_i) = (a'_i, b'_i, \sigma'_i)$ for all $i \in U_\ell$. Consequently,

$$\sum_{\ell=1}^{q} \sum_{i \in U_\ell} \exp(\beta_{0i}) \delta_{\{\beta_{1i}, a_i, b_i, \sigma_i\}} = \sum_{\ell=1}^{q} \sum_{i \in U_\ell} \exp(\beta'_{0i} + t_1) \delta_{\{\beta'_{1i} + t_2, a'_i, b'_i, \sigma'_i\}},$$

or equivalently, $G \equiv G'_{t_1, t_2}$. Hence, the proof is completed.

B.2 Proof of Proposition 2

Our proof will be based on the convergence rates of density estimation from MLE in Theorem 7.4 in [31]. Before stating this result here, let us introduce some necessary notations. Firstly, let $\mathcal{P}_k(\Theta)$ be the set of conditional densities of all mixing measures in $\mathcal{O}_k(\Theta)$, i.e., $\mathcal{P}_k(\Theta) := \{g_G(Y|X) : G \in \mathcal{O}_k(\Theta)\}$. Additionally, we define

$$\widetilde{\mathcal{P}}_k^{1/2}(\Theta) := \{g^{1/2}_{(G+G_*)/2}(Y|X) : G \in \mathcal{O}_k(\Theta)\}.$$

Next, for each $\delta > 0$, the Hellinger ball centered around the conditional density $g_{G_*}(Y|X)$ and intersected with the set $\widetilde{\mathcal{P}}_k^{1/2}(\Theta)$ is denoted by

$$\widetilde{\mathcal{P}}_k^{1/2}(\Theta, \delta) := \left\{g^{1/2} \in \widetilde{\mathcal{P}}_k^{1/2}(\Theta) : h(g, g_{G_*}) \leq \delta\right\}.$$

Finally, in order to measure the size of the above set, [31] proposes using the following quantity:

$$\mathcal{J}_B(\delta, \widetilde{\mathcal{P}}_k^{1/2}(\Theta, \delta)) := \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(t, \widetilde{\mathcal{P}}_k^{1/2}(\Theta, t), \|\cdot\|_2)\, \mathrm{d}t \vee \delta, \tag{30}$$

where $H_B(t, \widetilde{\mathcal{P}}_k^{1/2}(\Theta, t), \|\cdot\|_2)$ denotes the bracketing entropy [31] of $\widetilde{\mathcal{P}}_k^{1/2}(\Theta, u)$ under the ℓ_2-norm, and $t \vee \delta := \max\{t, \delta\}$. Now, we are ready to recall the statement of Theorem 7.4 in [31]:

Theorem 3 (Theorem 7.4, [31]). *Take $\Psi(\delta) \geq \mathcal{J}_B(\delta, \widetilde{\mathcal{P}}_k^{1/2}(\Theta, \delta))$ that satisfies $\Psi(\delta)/\delta^2$ is a non-increasing function of δ. Then, for some universal constant c and for some sequence (δ_n) such that $\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n)$, we achieve that*

$$\mathbb{P}\Big(h(g_{\widehat{G}_n}, g_{G_*}) > \delta\Big) \leq c \exp\left(-\frac{n\delta^2}{c^2}\right),$$

for all $\delta \geq \delta_n$.

The proof of this theorem can be seen in [31].

Proof of Proposition 2. Back to our main proof, since

$$H_B(t, \widetilde{\mathcal{P}}_k^{1/2}(\Theta, u), \|\cdot\|_2) \leq H_B(t, \mathcal{P}_k(\Theta, t), h)$$

for any $t > 0$, it follows from equation (30) that

$$\mathcal{J}_B(\delta, \widetilde{\mathcal{P}}_k^{1/2}(\Theta, \delta)) \leq \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(t, \mathcal{P}_k(\Theta, t), \|\cdot\|_2)\, \mathrm{d}t \vee \delta \lesssim \int_{\delta^2/2^{13}}^{\delta} \log(1/t) dt \vee \delta,$$

where we apply the upper bound of a bracketing entropy in Lemma 3 (cf. the end of this proof) in the second inequality. Let $\Psi(\delta) = \delta[\log(1/\delta)]^{1/2}$, we have $\Psi(\delta)/\delta^2$ is a non-increasing function of θ. Moreover, the above equation deduces that $\Psi(\delta) \geq \mathcal{J}_B(\delta, \widetilde{\mathcal{P}}_k^{1/2}(\Theta, \delta))$. Additionally, we also have that $\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n)$ for some universal constant c. As all the assumptions are met, Theorem 3 gives us that

$$\mathbb{P}(h(g_{\widehat{G}_n}, g_{G_*}) > C(\log(n)/n)^{1/2}) \lesssim \exp(-c\log(n)),$$

for some universal constant C that depends only on Θ. □

For completion, we will provide the result regarding the upper bound of a bracketing entropy in the following lemma:

Lemma 3. *Assume that Θ is a bounded set, then the following inequality holds true for any $0 \leq \varepsilon \leq 1/2$:*

$$H_B(\varepsilon, \mathcal{P}_k(\Theta), h) \lesssim \log(1/\varepsilon).$$

Proof of Lemma 3. Firstly, we will establish an upper bound for the univariate Gaussian density $f(Y|a^\top X + b, \sigma)$. Since both $\mathcal{X}$ and Θ are bounded sets, there exist positive constants κ, u, ℓ such that $-\kappa \leq a^\top X + b \leq \kappa$ and $\ell \leq \sigma \leq u$. As a result,

$$f(Y|a^\top X + b, \sigma) = \frac{1}{\sqrt{2\pi h_2}} \exp\Big(-\frac{(Y - h_1)^2}{2h_2}\Big) \leq \frac{1}{\sqrt{2\pi\ell}}.$$

For any $|Y| \geq 2\kappa$, we have that $\frac{(Y-h_1)^2}{2h_2} \geq \frac{Y^2}{8u}$, which leads to

$$f(Y|a^\top X + b, \sigma) \leq \frac{1}{\sqrt{2\pi\ell}} \exp\Big(-\frac{Y^2}{8u}\Big).$$

Putting the above results together, we obtain that $f(Y|a^\top X + b, \sigma) \leq K(Y|X)$, where we define $K(Y|X) := \frac{1}{\sqrt{2\pi\ell}} \exp\Big(-\frac{Y^2}{8u}\Big)$ if $|Y| \geq 2\kappa$, and $K(Y|X) := \frac{1}{\sqrt{2\pi\ell}}$ otherwise.

Subsequently, let $\eta \leq \varepsilon$, we assume that the set $\mathcal{P}_k(\Theta)$ has an η-cover (under ℓ_1-norm) denoted by $\{\pi_1, \ldots, \pi_N\}$, where $N := N(\eta, \mathcal{P}_k(\Theta), \|\cdot\|_1)$ is known as the η-covering number of $\mathcal{P}_k(\Theta)$. Then, we will build up the brackets of the form $[\nu_i(Y|X), \mu_i(Y|X)]$ for all $i \in [N]$ as follows:

$$\nu_i(Y|X) := \max\{\pi_i(Y|X) - \eta, 0\},$$
$$\mu_i(Y|X) := \max\{\pi_i(Y|X) + \eta, K(Y|X)\}.$$

Consequently, it can be checked that $\mathcal{P}_k(\Theta) \subset \bigcup_{i=1}^N [\nu_i(Y|X), \mu_i(Y|X)]$ with a note that $\mu_i(Y|X) - \nu_i(Y|X) \leq \min\{2\eta, K(Y|X)\}$. Next, for each $i \in [N]$, we attempt to give an upper bound for

$$\begin{aligned}\|\mu_i - \nu_i\|_1 &= \int_{|Y|<2\kappa} (\mu_i(Y|X) - \nu_i(Y|X))\, \mathrm{d}(X,Y) + \int_{|Y|\geq 2\kappa} (\mu_i(Y|X) - \nu_i(Y|X))\, \mathrm{d}(X,Y) \\ &\leq R\eta + \exp\Big(-\frac{R^2}{2u}\Big) \leq R'\eta,\end{aligned}$$

where $R := \max\{2\kappa, \sqrt{8u}\}\log(1/\eta)$ and R' is some positive constant. By definition of the bracketing entropy, since $H_B(R'\eta, \mathcal{P}_k(\Theta), \|\cdot\|_1)$ is the logarithm of the smallest number of brackets of size $R'\eta$ necessary to cover $\mathcal{P}_k(\Theta)$, we achieve that

$$\begin{aligned}H_B(R'\eta, \mathcal{P}_k(\Theta), \|\cdot\|_1) &\leq \log N = \log N(\eta, \mathcal{P}_k(\Theta), \|\cdot\|_1) \\ &\leq \log N(\eta, \mathcal{P}_k(\Theta), \|\cdot\|_\infty),\end{aligned}$$

where $\|\cdot\|_\infty$ denotes the sup-norm and the last inequality is due to the fact that $\|\cdot\|_\infty \leq \|\cdot\|_1$. Assume that the following upper bound for the covering number $\log N(\eta, \mathcal{P}_k(\Theta), \|\cdot\|_\infty) \lesssim \log(1/\eta)$ holds true (proof provided at the end), then the above result leads to

$$H_B(R'\eta, \mathcal{P}_k(\Theta), \|\cdot\|_1) \lesssim \log(1/\eta).$$

By selecting $\eta = \varepsilon/R'$, we receive that $H_B(\varepsilon, \mathcal{P}_k(\Theta), \|\cdot\|_1) \lesssim \log(1/\varepsilon)$. Furthermore, since the Hellinger distance is upper bounded by the ℓ_1-norm, we reach the desired conclusion:

$$H_B(\varepsilon, \mathcal{P}_k(\Theta), h) \lesssim \log(1/\varepsilon).$$

Upper bound of the covering number. For completion, we will establish the following upper bound for the covering number, i.e.,

$$\log N(\eta, \mathcal{P}_k(\Theta), \|\cdot\|_\infty) \lesssim \log(1/\eta).$$

Let us denote Δ_η as an η-cover of size M_1 for an k-dimensional simplex and $\Omega := \{(a, b, \sigma) : (\beta_0, \beta_1, a, b, \sigma) \in \Theta\}$. Since Θ is a compact set, Ω is also a compact set in $\mathbb{R}^{d+2}$. Thus, we can find an η-cover $\overline{\Omega}_\eta$ of Ω with the covering number M_2. It can be verified that $M_1 \leq (5/\eta)^k$ and $M_2 \lesssim \mathcal{O}((1/\eta)^{(d+2)k})$.

For each mixing measure $G = \sum_{i=1}^k \exp(\beta_{0i})\delta_{(\beta_{1i}, a_i, b_i, \sigma_i)} \in \mathcal{O}_k(\Theta)$, we consider another one denoted by $\widetilde{G} := \sum_{i=1}^k \exp(\beta_{0i})\delta_{(\beta_{1i}, \overline{a}_i, \overline{b}_i, \overline{\sigma}_i)}$, where $(\overline{a}_i, \overline{b}_i, \overline{\sigma}_i) \in \overline{\Omega}_\eta$ such that $(\overline{a}_i, \overline{b}_i, \overline{\sigma}_i)$ are the closest to (a_i, b_i, σ_i) for all $i \in [k]$. In addition, we also take into account the following mixing measure $\overline{G} := \sum_{i=1}^k \exp(\overline{\beta}_{0i})\delta_{(\overline{\beta}_{1i}, \overline{a}_i, \overline{b}_i, \overline{\sigma}_i)}$, where $(\overline{p}_i(X))_{i=1}^k := \left(\frac{\exp(\overline{\beta}_{1i}^\top X + \overline{\beta}_{0i})}{\sum_{j=1}^k \exp(\overline{\beta}_{1j}^\top X + \overline{\beta}_{0j})}\right)_{i=1}^k$ is the closest to $(p_i(X))_{i=1}^k := \left(\frac{\exp(\beta_{1i}^\top X + \beta_{0i})}{\sum_{j=1}^k \exp(\beta_{1j}^\top X + \beta_{0j})}\right)_{i=1}^k$. We can verify that the conditional density $g_{\overline{G}}$ belongs to the following set:

$$\mathcal{R} := \Big\{g_G \in \mathcal{P}_k(\Theta) : (p_i(X))_{i=1}^k \in \Delta_\eta,\ (a_i, b_i, \sigma_i)_{i=1}^k \in \Omega_\eta\Big\}.$$

By the triangle inequality, we have

$$\|g_G - g_{\overline{G}}\|_\infty \leq \|g_G - g_{\widetilde{G}}\|_\infty + \|g_{\widetilde{G}} - g_{\overline{G}}\|_\infty.$$

From the formulation of $\widetilde{G}$, we get the following bounds:

$$\begin{aligned}
\|g_G - g_{\widetilde{G}}\|_\infty &\leq \sum_{i=1}^{k} \|p_i(X)[f(Y|(a_i)^\top X + b_i, \sigma_i) - f(Y|(\overline{a}_i)^\top X + \overline{b}_i, \overline{\sigma}_i)]\|_\infty \\
&\lesssim \sum_{i=1}^{k} (\|a_i - \overline{a}_i\| + |b_i - \overline{b}_i| + |\sigma_i - \overline{\sigma}_i|) \\
&\lesssim \eta, \qquad (31)
\end{aligned}$$

where the second inequality follows from the facts that $\mathcal{X}$ is a bounded set. Additionally, we have

$$\|g_{\widetilde{G}} - g_{\overline{G}}\|_\infty \leq \sum_{i=1}^{k} \|[p_i(X) - \bar{p}_i(X)]f(Y|(\overline{a}_i)^\top X + \overline{b}_i, \overline{\sigma}_i)\|_\infty \lesssim \eta \qquad (32)$$

Combine the bounds in equations (31) and (32), we receive $\|g_G - g_{\overline{G}}\|_\infty \lesssim \eta$, which means that $\mathcal{R}$ is an η-cover (not necessarily smallest) of $\mathcal{P}_k(\Theta)$ under the sup-norm. By definition of the covering number, we know that

$$N(\eta, \mathcal{P}_k(\Theta), \|\cdot\|_\infty) \leq \mathcal{O}((5/\eta)^k) \cdot \mathcal{O}((1/\eta)^{(d+2)k}),$$

or equivalently,

$$\log N(\eta, \mathcal{P}_k(\Theta), \|\cdot\|_\infty) \lesssim \log(1/\eta).$$

Hence, the proof is completed. □

B.3 Proof of Lemma 1

First of all, let us recall the system of polynomial equations of interest here:

$$\sum_{j=1}^{m} \sum_{\alpha \in \mathcal{I}_{\ell_1,\ell_2}} \frac{1}{\alpha!} \cdot p_{5j}^2 p_{1j}^{\alpha_1} p_{2j}^{\alpha_2} p_{3j}^{\alpha_3} p_{4j}^{\alpha_4} = 0, \qquad (33)$$

where $\mathcal{I}_{\ell_1,\ell_2} = \{\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \mathbb{N}^d \times \mathbb{N}^d \times \mathbb{N} \times \mathbb{N} : \ \alpha_1 + \alpha_2 = \ell_1, \ |\alpha_2| + \alpha_3 + 2\alpha_4 = \ell_2\}$ for any $(\ell_1, \ell_2) \in \mathbb{N}^d \times \mathbb{N}$ such that $0 \leq |\ell_1| \leq r$, $0 \leq \ell_2 \leq r - |\ell_1|$ and $|\ell_1| + \ell_2 \geq 1$.

In this proof, we denote $p_{1j} = (p_{1j1}, p_{1j2}, \ldots, p_{1jd})$ and $p_{2j} = (p_{2j1}, p_{2j2}, \ldots, p_{2jd})$.

When $m = 2$:

By observing a portion of the above system when $\ell_1 = \mathbf{0}_d$, which is,

$$\sum_{j=1}^{m} \sum_{\alpha_3 + 2\alpha_4 = \ell_2} \frac{p_{5j}^2 p_{3j}^{\alpha_3} p_{4j}^{\alpha_4}}{\alpha_3! \alpha_4!} = 0, \qquad (34)$$

for all $1 \leq \ell_2 \leq r$, we deduce that $\bar{r}(m) \leq 4$ based on the Proposition 2.1 in [17]. Moreover, from the discussion in Section 3.2 about the system in equation (33), we know that $\bar{r}(m) \geq 3$. As a result, it is sufficient to show that $\bar{r}(m) > 3$. Indeed, when $r = 3$, the system in equation (33) can be written as follows:

$$\sum_{j=1}^{m} p_{5j}^2 p_{1jl} = 0 \ \forall l \in [d], \quad \sum_{j=1}^{m} p_{5j}^2 p_{3j} = 0, \quad \sum_{j=1}^{m} p_{5j}^2 (p_{2ju} + p_{1jv} p_{3j}) = 0 \ \forall u, v \in [d],$$
$$\sum_{j=1}^{m} p_{5j}^2 p_{1ju} p_{1jv} = 0 \ \forall u, v \in [d], \quad \sum_{j=1}^{m} p_{5j}^2 \Big(\frac{1}{2} p_{3j}^2 + p_{4j} \Big) = 0, \quad \sum_{j=1}^{m} p_{5j}^2 \Big(\frac{1}{3!} p_{3j}^3 + p_{3j} p_{4j} \Big) = 0,$$
$$\sum_{j=1}^{m} p_{5j}^2 p_{1ju} p_{1jv} p_{1jl} = 0 \ \forall u, v, l \in [d], \quad \sum_{j=1}^{m} p_{5j}^2 \Big(\frac{1}{2} p_{1ju} p_{1jv} p_{3j} + p_{1jl} p_{2j\tau} \Big) = 0 \ \forall u, v, l, \tau \in [d],$$
$$\sum_{j=1}^{m} p_{5j}^2 \Big(\frac{1}{2} p_{1ju} \cdot p_{3j}^2 + p_{1jv} p_{4j} + p_{2jl} p_{3j} \Big) = 0 \ \forall u, v, l, \tau \in [d]. \quad (35)$$

It can be seen that the following is a non-trivial solution of the above system: $p_{5j} = 1$, $p_{1j} = p_{2j} = \mathbf{0}_d$ for all $j \in [m]$, $p_{31} = \frac{\sqrt{3}}{3}$, $p_{32} = -\frac{\sqrt{3}}{3}$, $p_{41} = p_{42} = -\frac{1}{6}$. Therefore, we obtain that $\bar{r}(m) > 3$, which leads to $\bar{r}(m) = 4$.

When $m = 3$:

Note that $\bar{r}(m)$ is a monotonically increasing function of m, the previous result implies that $\bar{r}(m) > \bar{r}(2) = 4$, or equivalently, $\bar{r}(m) \geq 5$ when $m = 3$. Additionally, according to the Proposition 2.1 in [17], we have that $\bar{r}(m) \leq 6$ based on the reduced system in equation (34). Thus, we only need to show that $\bar{r}(m) > 5$. The system in equation (33) when $r = 5$ is a combination of the system in equation (35) and the following system:

$$\sum_{j=1}^{m} p_{5j}^2 p_{1ju} p_{1jv} p_{1jl} p_{1j\tau} = 0 \ \forall u, v, l, \tau \in [d], \quad \sum_{j=1}^{m} p_{5j}^2 \Big(\frac{1}{4!} p_{3j}^4 + \frac{1}{2!} p_{3j}^2 p_{4j} + \frac{1}{2!} p_{4j}^2 \Big) = 0,$$
$$\sum_{j=1}^{m} p_{5j}^2 \Big(\frac{1}{3!} p_{1ju} p_{3j}^3 + p_{1jv} p_{3j} p_{4j} + \frac{1}{2!} p_{2jl} p_{3j}^2 + p_{2j\tau} p_{4j} \Big) = 0 \ \forall u, v, l, \tau \in [d],$$
$$\sum_{j=1}^{m} p_{5j}^2 \Big(\frac{1}{3!} p_{1ju_1} p_{1ju_2} p_{1ju_3} p_{3j} + \frac{1}{2!} p_{1jv_1} p_{1jv_2} p_{2jv_3} \Big) = 0 \ \forall \{u_i\}_{i=1}^3, \{v_i\}_{i=1}^3 \in [d],$$
$$\sum_{j=1}^{m} p_{5j}^2 \Big(\frac{1}{2!2!} p_{1ju_1} p_{1ju_2} p_{3j}^2 + \frac{1}{2!} p_{1ju_3} p_{1ju_4} p_{4j} + p_{1ju_5} p_{1ju_6} p_{3j} \Big) = 0 \ \forall \{u_i\}_{i=1}^6 \in [d],$$
$$\sum_{j=1}^{m} p_{5j}^2 \prod_{i=1}^{5} p_{1ju_i} = 0 \ \forall \{u_i\}_{i=1}^5 \in [d],$$

$$\sum_{j=1}^{m} p_{5j}^2 \Big(\frac{1}{5!} p_{3j}^5 + \frac{1}{3!} p_{3j}^3 p_{4j} + \frac{1}{2!} p_{3j} p_{4j}^2 \Big) = 0,$$

$$\sum_{j=1}^{m} p_{5j}^2 \Big(\frac{1}{4!} p_{1ju_1} p_{3j}^4 + \frac{1}{2!} p_{1ju_2} p_{3j}^2 p_{4j} + \frac{1}{2!} p_{1ju_3} p_{4j}^2 + \frac{1}{3!} p_{2ju_4} p_{3j}^3 + p_{2ju_5} p_{3j} p_{4j} \Big) = 0 \ \forall \{u_i\}_{i=1}^5 \in [d],$$

$$\sum_{j=1}^{m} p_{5j}^2 \Big(\frac{1}{4!} \prod_{i=1}^{4} p_{1ju_i} p_{3j} + \frac{1}{3!} \prod_{i=5}^{7} p_{1ju_i} p_{2ju_8} \Big) = 0 \ \forall \{u_i\}_{i=1}^8 \in [d],$$

$$\sum_{j=1}^{m} p_{5j}^2 \Big(\frac{1}{2!3!} \prod_{i=1}^{2} p_{1ju_i} p_{3j}^3 + \frac{1}{2!} \prod_{i=3}^{4} p_{1ju_i} p_{3j} p_{4j} + p_{1ju_5} p_{2ju_6} (\frac{1}{2} p_{3j}^2 + p_{4j}) + \frac{1}{2!} \prod_{i=7}^{8} p_{2ju_i} p_{3j} \Big) = 0 \ \forall \{u_i\}_{i=1}^8 \in [d],$$

$$\sum_{j=1}^{m} p_{5j}^2 \Big(\frac{1}{3!2!} \prod_{i=1}^{3} p_{1ju_i} p_{3j}^2 + \frac{1}{3!} \prod_{i=4}^{6} p_{1ju_i} p_{4j} + \frac{1}{2!} p_{1ju_7} p_{2ju_8} p_{3j} + \frac{1}{2!} p_{1ju_9} \prod_{i=10}^{11} p_{2ju_i} \Big) = 0 \forall \{u_i\}_{i=1}^{11} \in [d].$$

We can verify that the non-trivial solution mentioned in the previous setting also satisfies this system. Hence, we conclude that $\bar{r}(m) = 6$.

References

[1] A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden markov models. In *COLT*, 2012. (Cited on page 2.)

[2] S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, 45:77–120, 2017. (Cited on page 2.)

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. (Cited on page 6.)

[4] J. H. Chen. Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23(1):221–233, 1995. (Cited on page 1.)

[5] Y. Chen, X. Yi, and C. Caramanis. A convex formulation for mixed regression with two components: Minimax optimal rates. In *COLT*, 2014. (Cited on page 2.)

[6] D. Do, L. Do, and X. Nguyen. Strong identifiability and parameter learning in regression with heterogeneous response. *arXiv preprint arXiv:2212.04091*, 2022. (Cited on page 2.)

[7] N. Doss, Y. Wu, P. Yang, and H. H. Zhou. Optimal estimation of high-dimensional Gaussian location mixtures. *The Annals of Statistics*, 51(1):62 – 95, 2023. Publisher: Institute of Mathematical Statistics. (Cited on page 2.)

[8] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. Yu, O. Firat, B. Zoph, L. Fedus, M. Bosma, Z. Zhou, T. Wang, E. Wang, K. Webster, M. Pellat, K. Robinson, K. Meier-Hellstern, T. Duke, L. Dixon, K. Zhang, Q. Le, Y. Wu, Z. Chen, and C. Cui. Glam: Efficient scaling of language models with mixture-of-experts. In *ICML*, 2022. (Cited on page 1.)

[9] R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu. Sharp analysis of expectation-maximization for weakly identifiable models. *AISTATS*, 2020. (Cited on page 2.)

[10] R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu. Singularity, misspecification, and the convergence rate of EM. *Annals of Statistics*, 44:2726–2755, 2020. (Cited on page 2.)

[11] D. Eigen, M. Ranzato, and I. Sutskever. Learning factored representations in a deep mixture of experts. In *ICLR Workshops*, 2014. (Cited on page 1.)

[12] W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23:1–39, 2022. (Cited on page 1.)

[13] A. Guha, N. Ho, and X. Nguyen. On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli*, 27(4):2159–2188, 2021. (Cited on pages 4 and 8.)

[14] M. Hardt and E. Price. Tight bounds for learning a mixture of two gaussians. In *STOC*, 2015. (Cited on page 2.)

[15] H. Hazimeh, Z. Zhao, A. Chowdhery, M. Sathiamoorthy, Y. Chen, R. Mazumder, L. Hong, and E. H. Chi. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. In *NeurIPS*, 2021. (Cited on page 1.)

[16] P. Heinrich and J. Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6):2844–2870, 2018. (Cited on page 1.)

[17] N. Ho and X. Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44:2726–2755, 2016. (Cited on pages 1, 3, and 24.)

[18] N. Ho and X. Nguyen. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10:271–307, 2016. (Cited on page 1.)

[19] N. Ho, C.-Y. Yang, and M. I. Jordan. Convergence rates for Gaussian mixtures of experts. *Journal of Machine Learning Research*, 23(323):1–81, 2022. (Cited on pages 2 and 3.)

[20] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3, 1991. (Cited on page 1.)

[21] W. Jiang and M. A. Tanner. On the identifiability of mixtures-of-experts. *Neural Networks*, 9:1253–1258, 1999. (Cited on page 5.)

[22] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994. (Cited on page 1.)

[23] J. Y. Kwon, N. Ho, and C. Caramanis. On the minimax optimality of the EM algorithm for learning two-component mixed linear regression. In *AISTATS*, 2021. (Cited on page 2.)

[24] T. Manole and N. Ho. Uniform convergence rates for maximum likelihood estimation under two-component gaussian mixture models. *arXiv preprint arXiv:2006.00704*, 2020. (Cited on page 1.)

[25] B. Mustafa, C. Ruiz, J. Puigcerver, R. Jenatton, and N. Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. In *NeurIPS*, 2022. (Cited on page 1.)

[26] X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 4(1):370–400, 2013. (Cited on pages 1 and 3.)

[27] F. Peng, R. A. Jacobs, and M. A. Tanner. Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association*, 91:953–960, 1996. (Cited on page 1.)

[28] C. Ruiz, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. Pinto, D. Keysers, and N. Houlsby. Scaling vision with sparse mixture of experts. In *NeurIPS*, 2021. (Cited on page 1.)

[29] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017. (Cited on page 1.)

[30] B. Sturmfels. *Solving Systems of Polynomial Equations*. Providence, R.I, 2002. (Cited on page 7.)

[31] S. van de Geer. *Empirical processes in M-estimation*. Cambridge University Press, 2000. (Cited on pages 20 and 21.)

[32] Y. Wu and P. Yang. Optimal estimation of Gaussian mixtures via denoised method of moments. *The Annals of Statistics*, 48:1987–2007, 2020. (Cited on page 2.)

[33] Y. Wu and H. H. Zhou. Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $o(\sqrt{n})$ iterations. *Mathematical Statistics and Learning*, 4:143–220, 2021. (Cited on page 2.)

[34] X. Yi, C. Caramanis, and S. Sanghavi. Alternating minimization for mixed linear regression. In *ICML*, 2014. (Cited on page 2.)

[35] Z. You, S. Feng, D. Su, and D. Yu. Speechmoe: Scaling to large acoustic models with dynamic routing mixture of experts. In *Interspeech*, 2021. (Cited on page 1.)

[36] Z. You, S. Feng, D. Su, and D. Yu. Speechmoe2: Mixture-of-experts model with improved routing. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7217–7221, 2022. (Cited on page 1.)

[37] K. Zhong, P. Jain, and I. S. Dhillon. Mixed linear regression with multiple components. In *NeurIPS*, 2016. (Cited on page 2.)