On Parameter Estimation in Deviated Gaussian Mixture of Experts

Huy Nguyen **Khai Nguyen** **Nhat Ho**

Department of Statistics and Data Sciences, The University of Texas at Austin

Abstract

We consider the parameter estimation problem in the *deviated Gaussian mixture of experts* in which the data are generated from $(1-\lambda^*)g_0(Y|X) + \lambda^* \sum_{i=1}^{k_*} p_i^* f(Y|(a_i^*)^\top X + b_i^*, \sigma_i^*)$, where X, Y are respectively a covariate vector and a response variable, $g_0(Y|X)$ is a known function, $\lambda^* \in [0,1]$ is true but unknown mixing proportion, and $(p_i^*, a_i^*, b_i^*, \sigma_i^*)$ for $1 \leq i \leq k^*$ are unknown parameters of the Gaussian mixture of experts. This problem arises from the goodness-of-fit test when we would like to test whether the data are generated from $g_0(Y|X)$ (null hypothesis) or they are generated from the whole mixture (alternative hypothesis). Based on the algebraic structure of the expert functions and the distinguishability between g_0 and the mixture part, we construct novel Voronoi-based loss functions to capture the convergence rates of maximum likelihood estimation (MLE) for our models. We further demonstrate that our proposed loss functions characterize the local convergence rates of parameter estimation more accurately than the generalized Wasserstein, a loss function being commonly used for estimating parameters in the Gaussian mixture of experts.

1 INTRODUCTION

Assume that $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$ are i.i.d samples from a joint distribution with density function $p_{\lambda^*, G_*}(X, Y) := p_{\lambda^*, G_*}(Y|X)\bar{f}(X)$, where $\bar{f}(X)$ is a prior density function of the explanatory variable X and the conditional density function $p_{\lambda^*, G_*}(Y|X)$ is a *deviated Gaussian mixture of experts* of order k_*, which takes the following form:

$$p_{\lambda^*, G_*}(Y|X) := (1-\lambda^*)g_0(Y|X) + \lambda^* \sum_{i=1}^{k_*} p_i^* f(Y|(a_i^*)^\top X + b_i, \sigma_i^*). \quad (1)$$

where $f(\cdot|\mu, \sigma)$ denotes the univariate Gaussian density function with mean μ and variance σ. Here, $g_0(Y|X)$ is a known function and $p_{G_*}(Y|X) := \sum_{i=1}^{k_*} p_i^* f(Y|(a_i^*)^\top X + b_i, \sigma_i^*)$ denotes the mixture of experts part with respect to G_*. Next, $\lambda^* \in [0,1]$ represents a true mixing proportion, whereas $G_* := \sum_{i=1}^{k_*} p_i^* \delta_{(a_i^*, b_i^*, \sigma_i^*)}$ is a true but unknown *mixing measure*, that is, a linear combination of Dirac measures δ associated with positive weights $(p_i^*)_{i=1}^{k_*}$ which sum up to one, i.e., $\sum_{i=1}^{k_*} p_i^* = 1$. Additionally, $(a_i^*, b_i^*, \sigma_i^*) \in \Theta \subset \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+$, for all $1 \leq i \leq k_*$, are called atoms or components of the true mixing measure G_*. Meanwhile, $h_1(X, a, b) := a^\top X + b$ and $h_2(X, \sigma) := \sigma$ are referred to as mean and variance expert functions.

Universal assumptions. For the sake of theory, we assume the distribution of X to be continuous so that the deviated Gaussian mixture of experts is identifiable (see Proposition 1). Moreover, we also assume that the parameter space Θ is compact and the covariate space $\mathcal{X}$ is bounded in order to guarantee the convergence of parameter estimation. Finally, we let $(a_1^*, b_1^*, \sigma_1^*), \ldots, (a_{k_*}^*, b_{k_*}^*, \sigma_{k_*}^*)$ be pairwise distinct to ensure the difference of Gaussian experts.

The deviated Gaussian mixture of experts (1) arises from the goodness-of-fit test (Jitkrittum et al., 2020; del Barrio et al., 1999; Hunter et al., 2008) when the null hypothesis says that the data are generated from the known joint distribution $g_0(Y|X)\bar{f}(X)$ while the alternative hypothesis corresponds to the assumption that the data indeed follow the joint distribution $p_{\lambda^*, G_*}(X, Y)$. Several settings of this testing problem had been considered in the literature; namely the problem of detection of sparse homogeneous mixtures (Donoho and Jin, 2004; Cai et al., 2007; Cai and Wu, 2014; Verzelen and Arias-Castro, 2017), the prob-

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

lem of testing the number of components (Chen et al., 2001; Kasahara and Shimotsu, 2014, 2015), and multiple testing problems (Patra and Sen, 2016; Deb et al., 2022). Moreover, the deviated Gaussian mixture of experts is also a generalization of the Gaussian mixture of experts (Jacobs et al., 1991; Jordan and Jacobs, 1994; Jordan and Xu, 1995), which have been used in various fields, namely speech recognition (Peng et al., 1996; Gaur et al., 2021; You et al., 2022), multi-task learning (Liang et al., 2022; Ma et al., 2018; Hazimeh et al., 2021), computer vision (Puigcerver et al., 2021; Lathuilière et al., 2017; Xia et al., 2022), medical images (Han et al., 2024) and natural language processing (Eigen et al., 2014; Shazeer et al., 2017; Fedus et al., 2022; Du et al., 2022; Zuo et al., 2023; Pham et al., 2024).

Maximum likelihood estimation. An important application of the deviated Gaussian mixture of experts to the hypothesis testing problem is parameter estimation, namely, the problem of estimating unknown mixing proportion λ^* and mixing measure G_*. It is worth noting that the number of experts k_* is also unknown in practice. Therefore, we fit the true model (1) with a deviated Gaussian mixture of k experts, where $k > k_*$, and then use the maximum likelihood estimation (MLE) method to find the estimates of λ^* and G_* as follows:

$$(\hat{\lambda}_n, \widehat{G}_n) \in \underset{(\lambda,G)\in[0,1]\times\mathcal{G}_k(\Theta)}{\arg\max} \sum_{i=1}^{n} \log(p_{\lambda,G}(Y_i|X_i)). \quad (2)$$

Here, we denote $\mathcal{G}_k(\Theta) := \{G = \sum_{i=1}^{k'} p_i \delta_{(a_i,b_i,\sigma_i)} : 1 \leq k' \leq k,\ (a_i, b_i, \sigma_i) \in \Theta\}$ as the set of all discrete probability measures with at most k components.

Challenge discussion. When $\lambda^* = 1$ is known, the conditional density $p_{\lambda^*,G_*}(Y|X)$ reduces to the mixture part $p_{G_*}(Y|X)$. Thus, the problem of estimating $\widehat{G}_n$ becomes a parameter estimation problem in Gaussian mixture of experts, which had been studied in Theorem 2 of Ho et al. (2022). Ho et al. (2022) demonstrated that the convergence rates of parameter estimation in the Gaussian mixture of experts were determined by the solvability of a system of polynomial equations induced the algebraic structures between the expert functions. These convergence rates ranged from order $\widetilde{\mathcal{O}}(n^{-1/4})$ to order $\widetilde{\mathcal{O}}(n^{-1/2r})$ for some $r \geq 4$.

However, when $\lambda^* \in [0, 1]$ is unknown, the theoretical understanding of the MLE $(\hat{\lambda}_n, \widehat{G}_n)$ in the deviated Gaussian mixture of experts becomes more challenging than those in the standard Gaussian mixture of experts. The main challenge comes from the interaction between the known function $g_0(Y|X)$ and the mixture part $p_{G_*}(Y|X)$ with respect to the mixing measure G_* via some partial differential equations (PDEs). This interaction influences not only the identifiability of the model but also the convergence rate of the MLE.

Another issue comes from the suboptimality of the generalized Wasserstein loss function (Villani, 2003, 2008) used in learning parameters. The idea of leveraging that loss function in analyzing the convergence behavior of MLE in mixture models was initialized by Nguyen (2013), and then extended to mixture of experts by Ho et al. (2022). An important property of this divergence is that the convergence of the MLE $\widehat{G}_n$ is able to imply those of its atoms. For example, it can be seen from Theorem 1 in Ho et al. (2022) that the convergence rate $\widetilde{\mathcal{O}}(n^{-1/4})$ of $\widehat{G}_n$ to G_* under the generalized Wasserstein indicates that the rates of estimating individual components are also $\widetilde{\mathcal{O}}(n^{-1/4})$. On the other hand, the generalized Wasserstein are unable to capture those rates accurately. In particular, while the estimation rates for those components should vary with the number of fitted components approximating them, that loss function always leads to the same rates.

Contribution. In the paper, we first establish the parametric convergence rate of density estimation $p_{\hat{\lambda}_n,\widehat{G}_n}$ to the true density p_{λ^*,G_*} of order $\widetilde{\mathcal{O}}(n^{-1/2})$ under the Total Variation distance V. Next, to address the above challenges of the parameter estimation problem in the deviated Gaussian mixture of experts, we first develop a distinguishability condition between the function $g_0(Y|X)$ and the mixture part $p_{G_*}(Y|X)$ in the deviated Gaussian mixture of experts in order to isolate the effect of function g_0 on the convergence behaviors of parameter estimation of p_{G_*}. Then, we conduct the convergence analysis of parameter estimation under *distinguishable settings*, namely when the distinguishability condition holds true, and *non-distinguishable settings*, i.e. when that condition does not hold. In each scenario, we construct a novel Voronoi loss function to precisely capture distinct convergence rates of parameter estimation in the deviated Gaussian mixture of experts (see also Table 1). Our theory can be summarized as follows:

1. Distinguishable settings: When the distinguishability condition holds, there is no impact of the function g_0 on the mixture of experts part p_{G_*}. Therefore, we design a novel Voronoi loss function $D_1((\hat{\lambda}_n, \widehat{G}_n), (\lambda^*, G_*))$ in equation (8), and then demonstrate that it is lower bounded by the Total Variation distance $V(p_{\hat{\lambda}_n,\widehat{G}_n}, p_{\lambda^*,G_*})$, which vanishes at the rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$. It follows from this bound that the estimation rate for (λ^*, G_*) is of order $\widetilde{\mathcal{O}}(n^{-1/2})$. Moreover, parameters $(a_i^*, b_i^*, \sigma_i^*)$ which are fitted by exactly one component enjoy the same estimation rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$. By contrast, if $(a_i^*, b_i^*, \sigma_i^*)$ are approached by more than one component, then the rates for estimating a_i^* become slower at $\widetilde{\mathcal{O}}(n^{-1/4})$,

Setting	Bound of k	Loss	Exact-fitted a_j^*, b_j^*, σ_j^*	Over-fitted a_j^*, b_j^*, σ_j^*
Distinguishable	$k \geq k_*$	D_1	$n^{-1/2}$	$n^{-1/4}, n^{-1/2\bar{r}(\lvert\mathcal{A}_j\rvert)}, n^{-1/\bar{r}(\lvert\mathcal{A}_j\rvert)}$
Non-distinguishable	$k \geq k_* + k_0 - \bar{k}$, $\hat{\lambda}_n > \lambda^*$	D_2	$n^{-1/2}$	$n^{-1/4}, n^{-1/2\bar{r}(\lvert\mathcal{B}_j\rvert)}, n^{-1/\bar{r}(\lvert\mathcal{B}_j\rvert)}$
	Otherwise			$n^{-1/4}, n^{-1/2\bar{r}(\lvert\mathcal{A}_j\rvert)}, n^{-1/\bar{r}(\lvert\mathcal{A}_j\rvert)}$
Theorem 2 Ho et al. (2022)	$k \geq k_*$	$\widetilde{W}$	$n^{-1/4}$	$n^{-1/4}, n^{-1/2\bar{r}(k-k_*+1)}, n^{-1/\bar{r}(k-k_*+1)}$

Table 1: Summary of parameter estimation rates in the (deviated) Gaussian mixture of experts. Here, exact-fitted parameters are those approximated by one fitted component, while their over-fitted counterparts are approached by at least two fitted components. Additionally, the value of function $\bar{r}(\cdot) \geq 4$ is determined by the solvability of the system of polynomial equations (6). Meanwhile, the cardinalities of Voronoi cells $\mathcal{A}_j$ and $\mathcal{B}_j$, which are respectively defined in equations (7) and (11), indicate the number of components fitting parameters a_j^*, b_j^*, σ_j^*. Lastly, the notation $\widetilde{W}$ stands for the generalized Wasserstein loss function used in Ho et al. (2022).

while those for b_i^* and σ_i^* are of orders $\widetilde{\mathcal{O}}(n^{-1/2\bar{r}_i})$ and $\widetilde{\mathcal{O}}(n^{-1/\bar{r}_i})$, respectively, where $\bar{r}_i \geq 4$ is determined by the solvability of the system of polynomial equations defined in equation (6).

2. Non-distinguishable settings: When the distinguishability condition fails, we consider the function g_0 as a Gaussian mixture of k_0 experts, where $1 \leq k_0 \leq k_*$, whose parameters interact with those of the mixture part p_{G_*}. Notably, the convergence behaviors of parameter estimation in the deviated Gaussian mixture of experts strictly depend on the interaction level determined by the number of overlapped components $\bar{k}$ that g_0 and p_{G_*} share. Therefore, we propose a novel Voronoi loss function $D_2((\hat{\lambda}_n, \widehat{G}_n), (\lambda^*, G_*))$ in equation (12) to capture this property, and then derive the Total Variation lower bound $D_2((\hat{\lambda}_n, \widehat{G}_n), (\lambda^*, G_*)) \,.\, V(p_{\hat{\lambda}_n, \mathfrak{b}_n}, p_{\lambda^*, G_*}) = \widetilde{\mathcal{O}}(n^{-1/2})$. Consequently, the rates for estimating (λ^*, G_*) and exact-fitted parameters a_j^*, b_j^*, σ_j^* are of order $\widetilde{\mathcal{O}}(n^{-1/2})$. On the other hand, for over-fitted parameters a_j^*, b_j^*, σ_j^*, while the estimation rate for a_j^* remains unchanged of order $\widetilde{\mathcal{O}}(n^{-1/4})$, those for b_j^*, σ_j^* not only depend on the solvability of the system (6) but also vary with the relation between k and $k_* + k_0 - \bar{k}$.

Organization. The paper is organized as follows. Firstly, we introduce a novel distinguishability condition and a notion of Voronoi cells as well as establish the density estimation rate in Section 2. In Section 3.1, we analyze the convergence behavior of parameter estimation under both the distinguishable and non-distinguishable settings and provide a proof sketch for main results. Then, we conduct a simulation study in Section 4 to empirically verify our theoretical results before concluding the paper in Section 5. Additional results and detailed proofs are all deferred to the supplementary material.

Notation. Let $[n]$ stand for the set $\{1, 2, \ldots, n\}$ for any $n \in \mathbb{N}$. Given two sequences $\{s_n\}$ and $\{t_n\}$, we write $s_n \,.\, t_n$ or $s_n = \mathcal{O}(t_n)$ if there exists a constant $C > 0$ independent of n such that $s_n \leq C t_n$ for all $n \in \mathbb{N}$ (similar for $s_n \,\&\, t_n$). Next, the notation $s_n = \widetilde{\mathcal{O}}(t_n)$ indicates that the previous bound occurs up to some logarithmic factor. Let $\|\cdot\|_p$ represents for the usual p-norm in $\mathbb{R}^d$ with a convention that $\|\cdot\|$ being the 2-norm. Finally, for any two probability density functions p and q (with respect to the Lebesgue measure μ), we define the Total Variation distance between them as $V(p, q) := \frac{1}{2}\int |p - q| d\mu$, while the Hellinger distance is given by $h(p, q) := \left(\frac{1}{2}\int (\sqrt{p} - \sqrt{q})^2 d\mu\right)^{1/2}$.

2 BACKGROUND

In this section, we first introduce a distinguishability condition, and then validate the identifiability of the deviated Gaussian mixture of experts as well as characterize the density estimation rate.

Recall that h_1 and h_2 are mean and variance expert functions in the true model (1). Then, we begin this section with the following distinguishability condition between the function g_0 and the mixture part p_{G_*}:

Definition 1 (Distinguishability Condition)**.** *We say that p_{G_*} is distinguishable from g_0 with respect to vector $r = (r_1, \ldots, r_{k_*}) \in \mathbb{N}^{k_*}$ if the following holds: assume that there exist real coefficients $\alpha^{(0)}$ and $\alpha^{(i)}_{\ell_1,\ell_2}$, for $i \in [k_*]$ and $0 \leq \ell_1 + \ell_2 \leq r_i$ that satisfy*

$$\sum_{i=1}^{k_*} \sum_{\ell_1+\ell_2=0}^{r_i} \alpha^{(i)}_{\ell_1,\ell_2} \frac{\partial^{\ell_1+\ell_2} f}{\partial h_1^{\ell_1} \partial h_2^{\ell_2}}(Y|(a_i^*)^\top X + b_i^*, \sigma_i^*) + \alpha^{(0)} g_0(Y|X) = 0,$$

for almost surely (X, Y), then $\alpha^{(0)} = \alpha^{(i)}_{\ell_1,\ell_2} = 0$ for any $i \in [k_*]$ and $0 \leq \ell_1 + \ell_2 \leq r_i$.

For better understanding, we provide below a scenario when p_{G_*} is distinguishable from g_0.

Example 1. *Let* $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{(\theta_{1i}^0, \theta_{2i}^0)} \in \mathcal{E}_{k_0}(\Theta) := \mathcal{O}_{k_0}(\Theta) \setminus \mathcal{O}_{k_0-1}(\Theta)$*, where* $k_0 \in \mathbb{N}$*. If we set*

$$g_0(Y|X) = p_{G_0}(Y|X) := \sum_{i=1}^{k_0} p_i^0 f(Y|(a_i^0)^\top X + b_i^0, \sigma_i^0), \quad (3)$$

then p_{G_*} *is distinguishable from* g_0 *whenever* $k_0 > k_*$*.*

In high level, the purpose of the distinguishability condition is to control the interaction level between the function g_0 and the mixture part p_{G_*}. From the perspective of the parameter estimation problem, if there is no effect of the function g_0 on the mixture p_{G_*}, then the convergence behaviors of parameter estimation in the deviated Gaussian mixture of experts will be similar to those in the standard Gaussian mixture of experts previously studied in Ho et al. (2022). On the other hand, when the distinguishability condition fails, i.e., there are interactions between g_0 and p_{G_*}, the parameter estimation rates will strictly depend on the interaction level among these two functions. In our paper, we illustrate that point by considering g_0 as a Gaussian mixture of k_0 expert given in equation (3), where $1 \leq k_0 \leq k_*$. This choice of function g_0 allows us to determine the level of the interaction between g_0 and p_{G_*} explicitly via the number of overlapped components that these functions share, which will be discussed further in Section 3.2.

Subsequently, we figure out in the following proposition that if the distinguishability condition is satisfied, then the deviated Gaussian mixture of experts in equation (1) is identifiable.

Proposition 1 (Identifiability)**.** *Let* G, G' *be two mixing measures in* $\mathcal{G}_k(\Theta)$ *and* λ, λ' *be two mixing proportions in* $[0,1]$*. Assume that* p_{G_*} *is distinguishable from* g_0*, then if the identifiability equation* $p_{\lambda,G}(X,Y) = p_{\lambda',G'}(X,Y)$ *holds for almost surely* $(X,Y) \in \mathcal{X} \times \mathcal{Y}$*, then we achieve that* $(\lambda, G) \equiv (\lambda', G')$*.*

Proof of Proposition 1 is in Appendix B.1. Given that p_{G_*} is distinguishable from h_0, this result ensures the convergence of the MLE $(\hat{\lambda}_n, \widehat{G}_n)$ to the true pair of mixing proportion and mixing measure (λ^*, G_*) when the density estimation $p_{\hat{\lambda}_n, \hat{\mathfrak{G}}_n}(X,Y)$ converges to the true density $p_{\lambda^*,G_*}(X,Y)$ for almost surely (X,Y). Thus, it is natural to explore the density estimation rate in the following proposition:

Proposition 2 (Density estimation rate)**.** *Suppose that the function* g_0 *is bounded with tail* $\mathbb{E}_X[-\log g_0(Y|X)]$ & Y^q *for almost surely* $Y \in \mathcal{Y}$ *for some* $q > 0$*. Then, the following inequality holds true:*

$$\mathbb{P}\Big(V(p_{\hat{\lambda}_n, \hat{\mathfrak{G}}_n}, p_{\lambda^*,G_*}) > C\sqrt{\log(n)/n}\Big) \ . \ n^{-c},$$

where $C > 0$ *is a constant that depends on* g_0*,* λ^**,* G_* *and* Θ*, while the constant* $c > 0$ *depends only on* Θ*.*

Proof of Proposition 2 is in Appendix B.2. The above bound indicates that the density estimation $p_{\hat{\lambda}_n, \hat{\mathfrak{G}}_n}$ converges to the true density p_{λ^*,G_*} under the Total Variation distance at the parametric rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$. In order to leverage this result, we assume that the function g_0 is bounded with tail $\mathbb{E}_X[-\log g_0(Y|X)]$ & Y^q for almost surely $Y \in \mathcal{Y}$ for some $q > 0$ throughout the paper unless stating otherwise.

3 CONVERGENCE RATES OF PARAMETER ESTIMATION

In this section, we aim to establish the convergence rates of maximum likelihood estimation in the deviated Gaussian mixture of experts under both the distinguishable and non-distinguishable settings.

3.1 Distinguishable Settings

Under this setting, the mixture part p_{G_*} is distinguishable from the function g_0 w.r.t vector $r = (r_1, \ldots, r_{k_*})$ that we will choose later. In other words, there is no interaction between p_{G_*} and g_0, and the following set is linearly independent for almost surely X:

$$\Big\{ \frac{\partial^{\ell_1+\ell_2} f}{\partial h_1^{\ell_1} \partial h_2^{\ell_2}}(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*), \ g_0(Y|X) : j \in [k_*], \ 0 \leq \ell_1 + \ell_2 \leq r_j \Big\}. \quad (4)$$

Given the parametric density estimation rate $V(p_{\hat{\lambda}_n, \hat{\mathfrak{G}}_n}, p_{\lambda^*,G_*}) = \widetilde{\mathcal{O}}(n^{-1/2})$ in Proposition 2, our main goal is to establish the Total Variation lower bound $V(p_{\hat{\lambda}_n, \hat{\mathfrak{G}}_n}, p_{\lambda^*,G_*})$ & $D_1((\hat{\lambda}_n, \widehat{G}_n), (\lambda^*, G_*))$, where D_1 will be defined in equation (8), in order to achieve the parametric convergence rate of the MLE $D_1((\hat{\lambda}_n, \widehat{G}_n), (\lambda^*, G_*)) = \widetilde{\mathcal{O}}(n^{-1/2})$. For that purpose, we first rewrite the density discrepancy $p_{\hat{\lambda}_n, \hat{\mathfrak{G}}_n}(X,Y) - p_{\lambda^*,G_*}(X,Y)$ in terms of $f(Y|(\hat{a}_i^n)^\top X + \hat{b}_i^n, \hat{\sigma}_i^n) - f(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*)$, where $(\hat{a}_i^n, \hat{b}_i^n, \hat{\sigma}_i^n)$ is a component of $\widehat{G}_n$. Next, we apply a Taylor expansion to the function $f(Y|(\hat{a}_i^n)^\top X + \hat{b}_i^n, \hat{\sigma}_i^n)$ about the point $(a_j^*, b_j^*, \sigma_j^*)$ to decompose the density discrepancy into a linear combination of linearly independent elements associated with coefficients involving the parameter discrepancies, namely $\hat{a}_i^n - a_j^*$, $\hat{b}_i^n - b_j^*$ and $\hat{\sigma}_i^n - \sigma_j^*$. As a result, when $p_{\hat{\lambda}_n, \hat{\mathfrak{G}}_n}$ converges to p_{λ^*,G_*}, the previous parameter discrepancies also go to zero and we

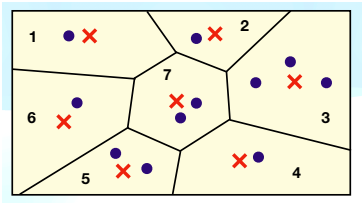


Figure 1: Illustration of the Voronoi cells generated by the components of G_* (red crosses) and the fitted components of the MLE $\widehat{G}_n$ (blue points). Under the distinguishable settings, Theorem 1 indicates that the rates for estimating true components $(a_j^*, b_j^*, \sigma_j^*)$ in cells $1, 2, 4, 6$, which are fitted by one component, are of order $\widetilde{\mathcal{O}}(n^{-1/2})$. Meanwhile, those for true components $(a_3^*, b_3^*, \sigma_3^*)$ in cell 3, which are fitted by three components, are drastically slow at $\widetilde{\mathcal{O}}(n^{-1/4})$, $\widetilde{\mathcal{O}}(n^{-1/12})$ and $\widetilde{\mathcal{O}}(n^{-1/6})$, respectively.

achieved our desired estimation rates. Note that such decomposition cannot be done if the set in equation (4) is linearly dependent. However, we observe an interaction among parameters of the Gaussian density f via the following partial differential equation (PDE):

$$\frac{\partial^2 f}{\partial b^2} = 2\,\frac{\partial f}{\partial \sigma}. \tag{5}$$

In high level, in Step 1 of our proofs, we need to decompose the density discrepancy $p_{\lambda_n, G_n}(Y|X) - p_{\lambda^*, G_*}(Y|X)$ into a combination of elements from some linearly independent set. To this end, we apply Taylor expansions to $f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n)$ around the true parameters $(a_j^*, b_j^*, \sigma_j^*)$. Unfortunately, there are many linearly dependent derivative terms arising from the above PDE. Thus, we have to group these terms together by taking the summation of their coefficients. Consequently, when the resulting coefficients tend to zero, we arrive at a system of polynomial equations which was previously studied in Ho and Nguyen (2016).

System of polynomial equations. Let $\bar{r}(m)$ be the smallest natural number r such that the following system of polynomial equations does not admit any non-trivial solutions for the unknown variables: $(s_l, t_{1l}, t_{2l})_{l=1}^m \subseteq \mathbb{R}^3$

$$\sum_{l=1}^{m} \sum_{\substack{n_1, n_2 \in \mathbb{N} \\ n_1 + 2n_2 = \beta}} \frac{s_l^2\, t_{1l}^{n_1}\, t_{2l}^{n_2}}{n_1! n_2!} = 0, \quad \beta = 1, 2, \ldots, r, \tag{6}$$

A solution to the above system is regarded as non-trivial if all variables s_l are non-zero, whereas at least one of the t_{1l} is different from zero. As shown in [Proposition 2.1, Ho and Nguyen (2016)], we have $\bar{r}(2) = 4$, $\bar{r}(3) = 6$ and $\bar{r}(m) \geq 7$ when $m \geq 4$.

Voronoi loss function: Intuitively, true parameters a_j^*, b_j^*, σ_j^* which are fitted by one component should admit faster estimation rates than those approximated by more than one component. To capture this convergence behavior of parameter estimation, let us introduce a class of Voronoi cells $\mathcal{A}_j \equiv \mathcal{A}_j(G)$ w.r.t an arbitrary mixing measure G, which are generated by the components $\theta_j^* := (a_j^*, b_j^*, \sigma_j^*)$ of G_* as follows:

$$\mathcal{A}_j := \{i \in [k] : \|\theta_i - \theta_j^*\| \leq \|\theta_i - \theta_\ell^*\|,\ \forall \ell \neq j\}, \tag{7}$$

where $\theta_i := (a_i, b_i, \sigma_i)$ for any $i \in [k]$. Notably, the cardinality of Voronoi cell $\mathcal{A}_j$ is exactly the number of components fitting θ_j^*. An instance of Voronoi cells is illustrated in Figure 1. Based on those cells, the Voronoi loss function used for this setting is defined as

$$\begin{aligned} &D_1((\lambda, G), (\lambda^*, G_*)) := |\lambda - \lambda^*| + (\lambda + \lambda^*) \\ &\times \Big[\sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} p_i \Big(\|\Delta a_{ij}\| + |\Delta b_{ij}| + |\Delta \sigma_{ij}| \Big) \\ &+ \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} p_i \Big(\|\Delta a_{ij}\|^2 + |\Delta b_{ij}|^{\bar{r}(|\mathcal{A}_j|)} \\ &\quad + |\Delta \sigma_{ij}|^{\bar{r}(|\mathcal{A}_j|)/2} \Big) + \sum_{j=1}^{k_*} \Big| \sum_{i \in \mathcal{A}_j} \lambda p_i - \lambda^* p_j^* \Big| \Big]. \end{aligned} \tag{8}$$

where $\Delta a_{ij} := a_i - a_j^*$, $\Delta b_{ij} := b_i - b_j^*$ and $\Delta \sigma_{ij} := \sigma_i - \sigma_j^*$. It is obvious that $D_1((\lambda, G), (\lambda^*, G_*)) = 0$ if and only if $(\lambda, G) = (\lambda^*, G_*)$. It is worth emphasizing that we design the above loss function to merely determine parameter estimation rates, so we do not attempt to optimize that loss. Now, we derive the convergence rate of the MLE $(\hat{\lambda}_n, \widehat{G}_n)$ under the challenging scenario when $\lambda^* \in (0, 1]$ in Theorem 1, while a discussion on the scenario when $\lambda^* = 0$ is relegated to Appendix D due to the space limit.

Theorem 1. *Assume that the distinguishability condition holds and $\lambda^* \in (0, 1]$ is unknown. Then, we achieve the Total Variation lower bound $V(p_{\lambda, G}, p_{\lambda^*, G_*})$ & $D_1((\lambda, G), (\lambda^*, G_*))$ for any $(\lambda, G) \in [0, 1] \times \mathcal{G}_k(\Theta)$. This bound together with Proposition 2 imply that*

$$\mathbb{P}\Big(D_1((\hat{\lambda}_n, \widehat{G}_n), (\lambda^*, G_*)) > C_1 \sqrt{\log(n)/n} \Big) \ .\ n^{-c_1},$$

where $C_1 > 0$ is a constant depending on $g_0, \lambda^*, G_*, \Theta$, while the constant $c_1 > 0$ depends only on Θ.

Proof of Theorem 1 is in Appendix A.1. When $\lambda^* \in (0, 1]$, the result that $D_1((\hat{\lambda}_n, \widehat{G}_n), (\lambda^*, G_*))$ van-

ishes at a rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$ implies the following observations (which are illustrated in Figure 1 as well):

(i) Firstly, for any $j \in [k_*]$ such that $|\mathcal{A}_j^n| = 1$, where $\mathcal{A}_j^n = \mathcal{A}_j(\widehat{G}_n)$, it follows that all the true parameters a_j^*, b_j^*, σ_j^*, which are fitted by a single component, share the same parametric rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$. On the other hand, [Theorem 2, Ho et al. (2022)], which used the generalized Wasserstein as a loss function, indicated that the rates for estimating those parameters were of orders $\widetilde{\mathcal{O}}(n^{-1/4})$, $\widetilde{\mathcal{O}}(n^{-1/2\bar{r}(k-k_*+1)})$ and $\widetilde{\mathcal{O}}(n^{-1/\bar{r}(k-k_*+1)})$, respectively. When $k - k_* + 1 = 3$, these rates become $\widetilde{\mathcal{O}}(n^{-1/4})$, $\widetilde{\mathcal{O}}(n^{-1/12})$ and $\widetilde{\mathcal{O}}(n^{-1/6})$, which are substantially slower than our parametric rate. This highlights the benefits of using the Voronoi loss function over the generalized Wasserstein in the convergence analysis of the MLE.

(ii) Secondly, for any $j \in [k_*]$ such that $|\mathcal{A}_j^n| > 1$, the rates for estimating true parameters a_j^*, b_j^*, σ_j^*, which are fitted by more than one component, are not uniform. More specifically, the estimation rates for b_j^* and σ_j^* are significantly slow, standing at orders $\widetilde{\mathcal{O}}(n^{-1/2\bar{r}(|\mathcal{A}_j^n|)})$ and $\widetilde{\mathcal{O}}(n^{-1/\bar{r}(|\mathcal{A}_j^n|)})$, respectively. This is due to the interaction between them via the PDE in equation (5). By contrast, since a_j^* does not interact with those parameters, their estimation rates are much faster of order $\widetilde{\mathcal{O}}(n^{-1/4})$.

(iii) Finally, we point out a scenario when true parameters b_j^*, σ_j^* attain the slowest estimation rates. In particular, assume that the MLE $\widehat{G}_n$ has $\hat{k}_n$ components. When $\widehat{G}_n$ converges to G_*, each Voronoi cell $\mathcal{A}_j^n$ contains at least one element for any $j \in [k_*]$, which implies that $|\mathcal{A}_j^n| \leq \hat{k}_n - k_* + 1$. The equality is achieved if, for example, $|\mathcal{A}_1^n| = \hat{k}_n - k_* + 1$ and $|\mathcal{A}_j^n| = 1$ for any $j \in [k_*] \setminus \{1\}$. Then, the rates for estimating b_1^*, σ_1^* reach the slowest orders of $\widetilde{\mathcal{O}}(n^{-1/2\bar{r}(|\mathcal{A}_j^n|)})$ and $\widetilde{\mathcal{O}}(n^{-1/\bar{r}(|\mathcal{A}_j^n|)})$, respectively, which match those of their counterparts in [Theorem 2, Ho et al. (2022)]. Conversely, we achieve the fastest estimation rates for other parameters b_j^*, σ_j^*, which are of order $\widetilde{\mathcal{O}}(n^{-1/2})$.

3.2 Non-distinguishable Settings

Under this setting, the mixture part p_{G_*} is not distinguishable from the function g_0, that is, the set in equation (4) is no longer linearly independent for almost surely X. There are several scenarios under which this phenomenon occurs, and one of them is when g_0 being a Gaussian mixture of k_0 experts as follows:

$$g_0(Y|X) = p_{G_0}(Y|X) := \sum_{j=1}^{k_0} p_j^0 f(Y|(a_j^0)^\top X + b_j^0, \sigma_j^0), \quad (9)$$

where $G_0 := \sum_{i=1}^{k_0} p_i^0 \delta_{(a_i^0, b_i^0, \sigma_i^0)} \in \mathcal{E}_{k_0}(\Theta)$ with $k_0 \in [k_*]$ such that G_0 and G_* share some common atoms. Without loss of generality, we assume that $(a_i^0, b_i^0, \sigma_i^0) = (a_j^*, b_j^*, \sigma_j^*)$ for any $j \in [\bar{k}]$, where $\bar{k} \in [k_0]$. We consider this choice of function g_0 as we can control the level of the interaction between g_0 and p_{G_*} explicitly via the number of overlapped components $\bar{k}$ of two mixing measures G_0 and G_*. In particular, we consider two separate regimes of the value of $\bar{k}$. The first one is when $1 \leq \bar{k} < k_0$, which is referred to as the *partial overlap* regime, and the second one is when $\bar{k} = k_0$, which is termed the *full overlap* regime. Due to the space limit, we will present only results for the partial overlap regime in this section, while those for the full overlap regime are deferred to Appendix C. Furthermore, similar to Section 3.1, we will also focus on only the scenario when $\lambda^* \in (0, 1]$, and relegate the details for the scenario when $\lambda^* = 0$ to Appendix D.

Partial overlap. Given the formulation of function g_0 in equation (9), the deviated Gaussian mixture of experts is no longer identifiable, that is, the equation $p_{\lambda,G}(X,Y) = p_{\lambda^*,G_*}(X,Y)$ for almost surely (X,Y) does not merely lead to $(\lambda, G) \equiv (\lambda^*, G_*)$ anymore, which causes a significant issue compared to the distinguishable settings. Therefore, it is necessary to find all the solutions (λ, G) of the equation $p_{\lambda,G}(X,Y) = p_{\lambda^*,G_*}(X,Y)$ for almost surely (X,Y). For that purpose, let us consider a new mixing measure. In particular, for any mixing proportion $\lambda > \lambda^*$, we define

$$\overline{G}_*(\lambda) := \left(1 - \frac{\lambda^*}{\lambda}\right) G_0 + \frac{\lambda^*}{\lambda} G_*, \quad (10)$$

as a mixing measure having a total of $k_* + k_0 - \bar{k}$ components in Θ. Note that the previous equation can be rewritten as $\lambda[p_G(X,Y) - p_{\overline{G}_*(\lambda)}(X,Y)] = 0$ for almost surely (X,Y). Thus, when $k \geq k_* + k_0 - \bar{k}$ and $\lambda > \lambda^*$, we admit $(\lambda, G) \equiv (\lambda, \overline{G}_*(\lambda))$ as a solution. On the other hand, when either $k < k_* + k_0 - \bar{k}$ or $\lambda \leq \lambda^*$, we obtain an obvious solution $(\lambda, G) \equiv (\lambda^*, G_*)$. For those reasons, we have to design a Voronoi loss function which is able to capture the aforementioned solutions.

Voronoi loss function. To facilitate our presentation, we assume that $\overline{G}_*(\lambda) = \sum_{j=1}^{k_*+k_0-\bar{k}} p_j'(\lambda) \delta_{(a_j', b_j', \sigma_j')}$. When $k \geq k_* + k_0 - \bar{k}$ and $\lambda > \lambda^*$, we introduce another set of Voronoi cells $\mathcal{B}_j \equiv \mathcal{B}_j(G)$ w.r.t an arbitrary mixing measure $G \in \mathcal{G}_k(\Theta)$ generated by the support of $\overline{G}_*(\lambda)$, denoted by $\theta_j' := (a_j', b_j', \sigma_j')$, as follows:

$$\mathcal{B}_j = \{i \in [k] : \|\theta_i - \theta_j'\| \leq \|\theta_i - \theta_\ell'\|, \ \forall \ell \neq j\}, \quad (11)$$

for any $j \in [k_* + k_0 - \bar{k}]$, where $\theta_i := (a_i, b_i, \sigma_i)$. Let us denote $\Delta' a_{ij} := a_i - a_j'$, $\Delta' b_{ij} := b_i - b_j'$ and $\Delta' \sigma_{ij} := \sigma_i - \sigma_j'$, then the discrepancy between two

mixing measures G and $\overline{G}_*(\lambda)$ can be characterized by

$$\begin{aligned}
&D_3(G, \overline{G}_*(\lambda)) \\
&:= \sum_{j:|\mathcal{B}_j|=1} \sum_{i\in\mathcal{B}_j} p_i \left(\|\Delta' a_{ij}\| + |\Delta' b_{ij}| + |\Delta' \sigma_{ij}|\right) \\
&+ \sum_{j:|\mathcal{B}_j|>1} \sum_{i\in\mathcal{B}_j} p_i \left(\|\Delta' a_{ij}\|^2 + |\Delta' b_{ij}|^{\bar{r}(|\mathcal{B}_j|)}\right. \\
&\left. + |\Delta' \sigma_{ij}|^{\bar{r}(|\mathcal{B}_j|)/2}\right) + \sum_{j=1}^{k_*+k_0-\bar{k}} \left|\sum_{i\in\mathcal{B}_j} p_i - p'_j(\lambda)\right|.
\end{aligned}$$

When either $k < k_* + k_0 - \bar{k}$ or $\lambda \leq \lambda^*$, we reuse the loss function D_1 in equation (8) to capture the solution $(\lambda, G) \equiv (\lambda^*, G_*)$. Thus, we combine the two loss functions D_1 and D_3 together to construct the following general Voronoi loss function used for any settings of k and λ under the partial overlap regime:

$$\begin{aligned}
&D_2((\lambda, G), (\lambda^*, G_*)) \\
&:= \begin{cases}
D_1((\lambda, G), (\lambda^*, G_*)), & \forall k < k_* + k_0 - \bar{k}; \\
\mathbf{1}_{\{\lambda\leq\lambda^*\}} D_1((\lambda, G), (\lambda^*, G_*)) & \\
+\mathbf{1}_{\{\lambda>\lambda^*\}} D_3(G, \overline{G}_*(\lambda)), & \forall k \geq k_* + k_0 - \bar{k}.
\end{cases}
\end{aligned} \tag{12}$$

Now, we are ready to present the main result for the partial overlap regime in the following theorem.

Theorem 2. *Assume that $\lambda^* \in (0, 1]$ is unknown, and let g_0 take the form in equation (9) with $1 \leq \bar{k} < k_0$. Then, we obtain that $V(p_{\lambda,G}, p_{\lambda^*,G_*})$ & $D_2((\lambda, G), (\lambda^*, G_*))$ for any $(\lambda, G) \in [0, 1] \times \mathcal{G}_k(\Theta)$. This bound together with Proposition 2 indicate that*

$$\mathbb{P}(D_2((\hat{\lambda}_n, \widehat{G}_n), (\lambda^*, G_*)) > C_2\sqrt{\log(n)/n}) \ . \ n^{-c_2},$$

where $C_2 > 0$ is a constant depending on $g_0, \lambda^*, G_*, \Theta$, while the constant $c_2 > 0$ depends only on Θ.

Proof of Theorem 2 is in Appendix A.2. A few comments regarding Theorem 2 are in order. Firstly, when either $k < k_* + k_0 - \bar{k}$ or $\hat{\lambda}_n \leq \lambda^*$, the loss function D_2 reduces to D_1. Therefore, the convergence rates of parameter estimation remain the same as those in Theorem 1. Secondly, when $k \geq k_* + k_0 - \bar{k}$ and $\hat{\lambda}_n > \lambda^*$, the loss function D_2 turns into D_3. As a result, for true components (a'_j, b'_j, σ'_j) which are approximated by more than one fitted components, the rates of estimating b'_j and σ'_j are reported to be $\widetilde{\mathcal{O}}(n^{-1/2\bar{r}(|\mathcal{B}^n_j|)})$ and $\widetilde{\mathcal{O}}(n^{-1/\bar{r}(|\mathcal{B}^n_j|)})$, respectively, where $\mathcal{B}^n_j := \mathcal{B}_j(\widehat{G}_n)$. Meanwhile, those for a'_j are of order $\widetilde{\mathcal{O}}(n^{-1/4})$. However, for true components (a'_j, b'_j, σ'_j) approximated by a single fitted component, their estimation rates are uniform, standing at $\widetilde{\mathcal{O}}(n^{-1/2})$. This again confirms the ability to capture distinct estimation rates accurately of the proposed Voronoi loss functions in comparison with the generalized Wasserstein.

3.3 Proof Sketch

In this section, we provide a generic proof sketch for Theorems 1 and 2, and distinguish our proof techniques from those used in the most related work (Ho et al., 2022). More details of these proofs are deferred to Appendix A. For simplicity, the metric $\mathcal{D}$ used in this sketch is implicitly understood as one among the Voronoi loss functions D_1 and D_2 in Sections 3.1 and 3.2. Generally, our goal is to establish the Total Variation lower bound $V(p_{\lambda,G}, p_{\lambda^*,G_*})$ & $\mathcal{D}((\lambda, G), (\lambda^*, G_*))$ for any $(\lambda, G) \in [0, 1] \times \mathcal{G}_k(\Theta)$, which together with Proposition 2 give us our desired conclusions in those theorems.

Local inequality. First, we prove the following local bound by contradiction in three main steps:

$$\lim_{\varepsilon\to 0} \inf_{\substack{(\lambda,G)\in[0,1]\times\mathcal{G}_k(\Theta):\\ \mathcal{D}((\lambda,G),(\lambda^*,G_*))\leq\varepsilon}} \frac{V(p_{\lambda,G}, p_{\lambda^*,G_*})}{\mathcal{D}((\lambda, G), (\lambda^*, G_*))} > 0. \tag{13}$$

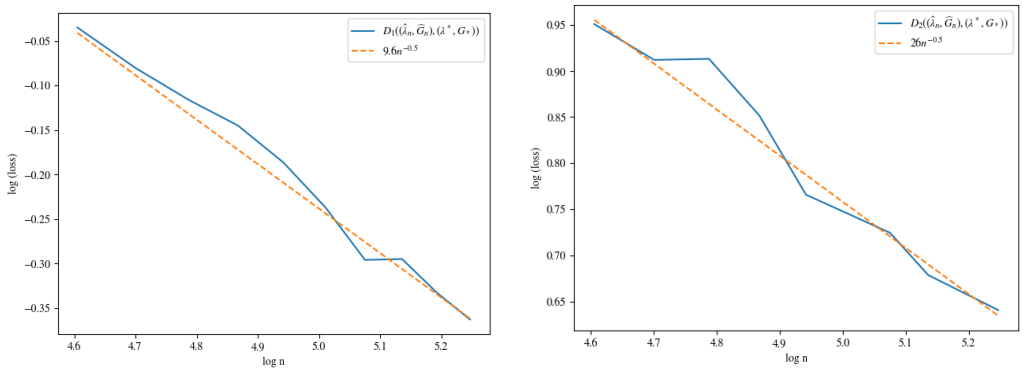
Step 1. Assume that the local inequality does not hold, then there exists a sequence (λ_n, G_n) such that both $\mathcal{D}_n := \mathcal{D}((\lambda_n, G_n), (\lambda^*, G_*))$ and $V(p_{\lambda_n,G_n}, p_{\lambda^*,G_*})/\mathcal{D}_n$ vanish as $n \to \infty$. Different from Ho et al. (2022), we need to invoke the Taylor expansion twice in this step due to the sophisticated structure of our metric $\mathcal{D}$. Firstly, for each $j \in [k_*] : |\mathcal{A}_j| = 1$, we apply the first-order Taylor expansion to the quantity $U_n := [p_{\lambda_n,G_n}(X, Y) - p_{\lambda^*,G_*}(X, Y)]/\mathcal{D}_n$, whereas for each $j \in [k_*] : |\mathcal{A}_j| > 1$, we use the Taylor expansion up to order r_j that will be chosen later in Step 2. Then, we show that U_n can be written as a linear combination of elements of some linearly independent set $\mathcal{H}$.

Step 2. We attempt to show by contradiction that at least one among the coefficients in that combination does not converge to zero. Assume that all of them go to zero. Then, by some algebraic transformations of those limits, we arrive at the following system of polynomial equations:

$$\sum_{i\in\mathcal{A}_j} \sum_{\substack{n_1,n_2\in\mathbb{N},\\ n_1+2n_2=\beta}} \frac{s_i^2\, t_{1i}^{n_1}\, t_{2i}^{n_2}}{n_1! n_2!} = 0, \quad \beta = 1, 2, \ldots, r_j.$$

By construction, this system necessarily has at least one non-trivial solution. Therefore, in order to point out a contradiction, we set $r_j = \bar{r}(|\mathcal{A}_j|)$ so that the above system does not have any non-trivial solutions.

Step 3. On the other hand, by means of Fatou's lemma and the limit $V(p_{\lambda_n,G_n}, p_{\lambda^*,G_*})/\mathcal{D}_n \to 0$, we show that $U_n \to 0$ for almost surely (X, Y). Since $\mathcal{H}$ is a linearly independent set, we deduce that all the coefficients of elements of $\mathcal{H}$ in the representation of U_n vanish as $n \to \infty$, which contradicts the result in Step 2. Hence, we obtain the local inequality in equation (13).



(a) Distinguishable setting (b) Non-distinguishable setting

Figure 2: Convergence rate of the maximum likelihood estimation $(\hat{\lambda}_n, \widehat{G}_n)$ under the Voronoi loss functions.

Global inequality. Given the above local inequality, it suffices to prove the following global bound:

$$\inf_{\substack{(\lambda,G)\in[0,1]\times\mathcal{G}_k(\Theta):\\ \mathcal{D}((\lambda,G),(\lambda^*,G_*))>\varepsilon}} \frac{V(p_{\lambda,G}, p_{\lambda^*,G_*})}{\mathcal{D}((\lambda,G),(\lambda^*,G_*))} > 0. \quad (14)$$

A key step for establishing this bound is to solve the equation $p_{\lambda,G}(X,Y) = p_{\lambda^*,G_*}(X,Y)$ for almost surely (X,Y). If p_{G_*} is distinguishable from g_0, then this equation has the unique solution $(\lambda, G) \equiv (\lambda^*, G_*)$. However, this property does not hold when the distinguishability condition fails, which induces a huge challenge compared to Ho et al. (2022). Thus, we solve the previous equation under the partial overlap regime in Section 3.1. Hence, the proof sketch is completed.

4 SIMULATION STUDY

In this section, we carry out a simulation study to empirically verify our theoretical results regarding the convergence rate of the MLE $(\hat{\lambda}_n, \widehat{G}_n)$ under both the distinguishable and non-distinguishable settings.

Distinguishable setting. We first generate the covariates $X_1, \ldots, X_n \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, and then set

$$g_0(Y|X) = \sum_{j=1}^{2} p_j^0 f(Y|a_j^0 X + b_j^0, \sigma_j^0),$$

where $(a_1^0, b_1^0, \sigma_1^0) = (.2, .1, .01)$, $(a_2^0, b_2^0, \sigma_2^0) = (.1, 0, .01)$ and $p_1^0 = p_2^0 = \frac{1}{2}$. Next, we consider the following true conditional density function:

$$p_{\lambda^*,G_*}(Y|X) := (1-\lambda^*)g_0(Y|X) + \lambda^* f(Y|a^*X + b^*, \sigma^*),$$

in which $\lambda^* = 0.5$ and $(a^*, b^*, \sigma^*) = (1, 1, 1)$. Here, we have $k_0 = 2 > 1 = k_*$, therefore, the distinguishability condition is satisfied according to Example 1. Subsequently, we draw a sample $Y_1, Y_2 \ldots, Y_n$ of size $n \in \{100, 110, 120, \ldots, 200\}$ from $p_{\lambda^*,G_*}(Y|X)$. Then, we overfit the true model by a deviated Gaussian mixture of $k = 2$ experts, and run the EM algorithm (Dempster et al., 1977) in 1000 iterations to find the estimators $\hat{\lambda}_n$, $(\hat{p}_1^n, \hat{a}_1^n, \hat{b}_1^n, \hat{\sigma}_1^n)$ and $(\hat{p}_2^n, \hat{a}_2^n, \hat{b}_2^n, \hat{\sigma}_2^n)$. Finally, we compute the discrepancy $D_1((\hat{\lambda}_n, \widehat{G}_n), (\lambda^*, G_*))$, where $\widehat{G}_n = \sum_{i=1}^2 \hat{p}_i^n \delta_{(\hat{a}_i^n, \hat{b}_i^n, \hat{\sigma}_i^n)}$ and $G_* = \delta_{(a^*,b^*,\sigma^*)}$, and plot its values in Figure 2(a). From the figure, it is obvious that the convergence rate of the MLE $(\hat{\lambda}_n, \widehat{G}_n)$ under the Voronoi loss D_1 is at the order of $\widetilde{\mathcal{O}}(n^{-1/2})$, which is consistent with our theoretical result in Theorem 1.

Non-distinguishable setting. For this setting, we also generate $X_1, \ldots, X_n \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, but set

$$g_0(Y|X) = \sum_{j=1}^{3} p_j^0 f(Y|a_j^0 X + b_j^0, \sigma_j^0),$$

with $(a_1^0, b_1^0, \sigma_1^0) = (.2, .1, .01)$, $(a_2^0, b_2^0, \sigma_2^0) = (.1, .1, .01)$, $(a_3^0, b_3^0, \sigma_3^0) = (.21, .11, .01215)$ and $p_1^0 = p_2^0 = p_0^3 = \frac{1}{3}$. Next, we consider the following true conditional density

$$p_{\lambda^*,G_*}(Y|X) := (1-\lambda^*)g_0(Y|X) + \lambda^* \sum_{j=1}^{3} p_j^* f(Y|a_j^* X + b_1^*, \sigma_j^*),$$

with $(a_1^*, b_1^*, \sigma_1^*) = (.2, .1, .01)$, $(a_2^*, b_2^*, \sigma_2^*) = (.1, .4, .25)$, $(a_3^*, b_3^*, \sigma_3^*) = (1.1, .3, .25)$ $p_1^* = p_2^* = p_3^* = \frac{1}{3}$ and $\lambda^* = 0.5$. Here, each of G_0 and G_* has 3 components and they share one common component among them, specifically $(a_1^0, b_1^0, \sigma_1^0) = (a_1^*, b_1^*, \sigma_1^*)$. Thus, p_{G_*} is not distinguishable from g_0, and this setting belongs to the partial overlap regime in Section 3.2 with $\bar{k} = 1$. We then sample $Y_1, \ldots, Y_n$ with $n \in \{100, 110, 120, \ldots, 200\}$ from $p_{\lambda^*,G_*}(Y|X)$. Next, we overfit the true model by a deviated Gaussian of $k = 4$ experts and use the EM algorithm to find the estimators $\hat{\lambda}_n$, $\hat{p}_i^n$, $\hat{a}_i^n$, $\hat{b}_i^n$ and $\hat{\sigma}_i^n$ for $i \in [4]$ in 1000 iterations. After that, we calculate the distance $D_2((\hat{\lambda}_n, \widehat{G}_n), (\lambda^*, G_*))$, and plot its values

in Figure 2(b). It can be seen from the figure that the convergence rate of the MLE $(\hat{\lambda}_n, \widehat{G}_n)$ is at the order of $\widetilde{\mathcal{O}}(n^{-1/2})$, which aligns with our claim in Theorem 2.

5 CONCLUSION AND DISCUSSION

In this paper, we characterize the convergence behaviors of maximum likelihood estimation in the deviated Gaussian mixture of experts, which is motivated by the goodness-of-fit test. We first show that the convergence rate of density estimation to the true density is parametric on the sample size. Regarding the parameter estimation problem, we consider two separate settings based on the level of distinguishability between the known function $g_0(Y|X)$ and the mixture of experts part $p_{G_*}(Y|X)$ in the proposed model. In each setting, we design a novel Voronoi loss function to capture the interaction among the parameters of expert functions, and the distinguishability of p_{G_*} from g_0. We then theoretically and empirically demonstrate that our proposed loss functions outperform the generalized Wasserstein studied in previous work in terms of precisely characterizing distinct parameter estimation rates, which are determined by the solvability of a system of polynomial equations.

Technical novelty. The most related work to our paper is Ho et al. (2022) which characterize the convergence behavior of parameter estimation in Gaussian mixture of experts. Compared to Ho et al. (2022), our paper is technically novel in three major aspects:

1. Distinguishability condition: In this work, we cope with not only the interaction (5) among parameters of the Gaussian density f as in Ho et al. (2022) but also another interaction between the known density g_0 and the mixture part p_{G_*}. Thus, we introduce a novel distinguishability condition to isolate the effect of g_0 on the convergence behaviors of parameter estimation of p_{G_*}. To the best of our knowledge, such condition for mixture-of-experts model had never been studied in prior work.

2. Loss functions: Ho et al. (2022) propose using the generalized Wasserstein loss function among parameters to determine the convergence rates of their estimation. However, this loss leads to the same estimation rates for true parameters fitted by multiple components, which should be distinct. To capture those rates precisely, we construct novel Voronoi loss functions in equations (8) and (12) based on Voronoi cells. For instance, it follows from our Theorem 1 that the estimation rate for exact-fitted parameter b_j^* is $\mathcal{O}(n^{-1/2})$, while that for its over-fitted counterpart is $\mathcal{O}(n^{-1/2\bar{r}(|\mathcal{A}_j|)})$. By contrast, Theorem 2 in Ho et al. (2022) indicates that the rates for estimating exact-fitted and over-fitted parameters b_j^* are the same of order $\mathcal{O}(n^{-1/2\bar{r}(k-k_*+1)})$. When $k - k_* + 1 = 3$, these rates become $\mathcal{O}(n^{-1/12})$, which are substantially slower than our parametric rate $\mathcal{O}(n^{-1/2})$ (see our Table 1).

3. Model identifiability: Although the deviated Gaussian mixture of experts is identifiable under the distinguishable settings (see Proposition 1), this property does not hold under the non-distinguishable settings. Therefore, we have to solve the non-trivial equation of variable (λ, G): $p_{\lambda,G}(X,Y) = p_{\lambda^*,G_*}(X,Y)$ for almost surely (X,Y) under both the partial overlap and full overlap regimes. This accounts for the involvement of mixing measures $\overline{G}_*(\lambda)$ and $\widetilde{G}_*(\lambda)$ in the loss functions D_2 (see equation (12)) and D_4 (see equation (34)), respectively, which are unprecedented in the literature, including Ho et al. (2022).

Future directions. There are a few potential directions which are beyond the scope of our work and can be developed in the future:

(i) Firstly, in the deviated Gaussian mixture of experts (1), we may consider a closer setting to practice by assuming that the true parameters $(a_i^*, b_i^*, \sigma_i^*)$ and the true mixing proportion λ^* vary with the sample size n (see (Do et al., 2023)). Under that setting, we would achieve uniform convergence rates of parameter estimation rather than point-wise rates as in the current work.

(ii) Secondly, we can adopt the current techniques to characterize the convergence behavior of parameter estimation under the deviated Gaussian mixture of experts with covariate-dependent gating functions, namely Gaussian gate (Nguyen et al., 2024d), softmax gate (Nguyen et al., 2023, 2024c), Top-K sparse softmax gate (Nguyen et al., 2024b) and dense-to-sparse gate (Nguyen et al., 2024a).

(iii) Finally, the theory in our paper relies on the assumption that the data are generated from a deviated Gaussian mixture of experts, i.e., the model is well-specified. When the model is misspecified, which resembles real-world applications, the data are sampled from some unknown distribution associated with a conditional density $q(Y|X)$ (not necessarily a deviated Gaussian mixture of experts). Then, the MLE $(\hat{\lambda}_n, \widehat{G}_n)$ converges to a pair

$$(\breve{\lambda}, \breve{G}) \in \operatorname*{arg\,min}_{(\lambda,G)\in[0,1]\times\mathcal{G}_k(\Theta)} \mathrm{KL}(q(Y|X)\|p_{\lambda,G}(Y|X)),$$

where KL denotes the Kullback-Leibler divergence. The insights from our theories indicate that the Voronoi losses can be used to obtain the precise rates of individual parameters of the MLE under the misspecified setting. See (van de Geer, 2000) for further details.

Acknowledgements

NH acknowledges support from the NSF IFML 2019844 and the NSF AI Institute for Foundations of Machine Learning.

References

T. Cai and Y. Wu. Optimal detection of sparse mixtures against a given null distribution. *IEEE Transactions on Information Theory*, 60(4):2217 – 2232, 2014. (Cited on page 1.)

T. Cai, J. Jin, and M. G. Low. Estimation and confidence sets for sparse normal mixtures. *Annals of Statistics*, 35(6):2421–2449, 2007. (Cited on page 1.)

H. Chen, J. Chen, and J. D. Kalbfleisch. A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(1):19–29, 2001. (Cited on page 2.)

N. Deb, S. Saha, A. Guntuboyina, and B. Sen. Two-component mixture model in the presence of covariates. *Journal of the American Statistical Association*, 117(540):1820–1834, 2022. (Cited on page 2.)

E. del Barrio, J. Cuesta-Albertos, C. Matrán, and J. Rodríguez-Rodríguez. Tests of goodness of fit based on the l_2-Wasserstein distance. *Annals of Statistics*, 27(4):1230–1239, 1999. (Cited on page 1.)

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, Sept. 1977. ISSN 0035-9246. (Cited on page 8.)

D. Do, H. Nguyen, K. Nguyen, and N. Ho. Minimax optimal rate for parameter estimation in multivariate deviated models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. (Cited on page 9.)

D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32(3):962–994, 2004. (Cited on page 1.)

N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, B. Zoph, L. Fedus, M. P. 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 *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5547–5569. PMLR, 17–23 Jul 2022. (Cited on page 2.)

D. Eigen, M. Ranzato, and I. Sutskever. Learning factored representations in a deep mixture of experts. In *ICLR Workshops*, 2014. (Cited on page 2.)

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 2.)

N. Gaur, B. Farris, P. Haghani, I. Leal, P. J. M. Mengibar, M. Prasad, B. Ramabhadran, and Y. Zhu. Mixture of informed experts for multilingual speech recognition. In *ICASSP 2021, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021. (Cited on page 2.)

X. Han, H. Nguyen, C. Harris, N. Ho, and S. Saria. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. *arXiv preprint arXiv:2402.03226*, 2024. (Cited on page 2.)

H. Hazimeh, Z. Zhao, A. Chowdhery, M. Sathiamoorthy, Y. Chen, R. Mazumder, L. Hong, and E. Chi. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. *Advances in Neural Information Processing Systems*, 34:29335–29347, 2021. (Cited on page 2.)

N. Ho and X. Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6):2726 – 2755, 2016. doi: 10.1214/16-AOS1444. (Cited on page 5.)

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, 3, 4, 6, 7, 8, 9, 23, 32, and 34.)

D. R. Hunter, S. M. Goodreau, and M. S. Handcock. Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481): 248–258, 2008. (Cited on page 1.)

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 2.)

W. Jitkrittum, H. Kanagawa, and B. Schölkopf. Testing goodness of fit of conditional density models with kernels. In J. Peters and D. Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 221–230. PMLR, 03–06 Aug 2020. (Cited on page 1.)

M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994. (Cited on page 2.)

M. I. Jordan and L. Xu. Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks*, 8, 1995. (Cited on page 2.)

H. Kasahara and K. Shimotsu. Non-parametric identification and estimation of the number of components in multivariate mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):97–111, 2014. (Cited on page 2.)

H. Kasahara and K. Shimotsu. Testing the number of components in normal mixture regression models. *Journal of the American Statistical Association*, 110 (512):1632–1645, 2015. (Cited on page 2.)

S. Lathuilière, R. Juge, P. Mesejo, R. Muñoz-Salinas, and R. Horaud. Deep mixture of linear inverse regressions applied to head-pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4817–4825, 2017. (Cited on page 2.)

H. Liang, Z. Fan, R. Sarkar, Z. Jiang, T. Chen, K. Zou, Y. Cheng, C. Hao, and Z. Wang. M^3ViT: Mixture-of-Experts Vision Transformer for Efficient Multi-task Learning with Model-Accelerator Co-design. In *NeurIPS*, 2022. (Cited on page 2.)

J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi. Modeling Task Relationships in Multi-Task Learning with Multi-Gate Mixture-of-Experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 1930–1939, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 978-1-4503-5552-0. doi: 10.1145/3219819.3220007. event-place: London, United Kingdom. (Cited on page 2.)

H. Nguyen, T. Nguyen, and N. Ho. Demystifying softmax gating function in Gaussian mixture of experts. In *Advances in Neural Information Processing Systems*, 2023. (Cited on page 9.)

H. Nguyen, P. Akbarian, and N. Ho. Is temperature sample efficient for softmax Gaussian mixture of experts? *arXiv preprint arXiv:2401.13875*, 2024a. (Cited on page 9.)

H. Nguyen, P. Akbarian, F. Yan, and N. Ho. Statistical perspective of top-k sparse softmax gating mixture of experts. In *International Conference on Learning Representations*, 2024b. (Cited on page 9.)

H. Nguyen, N. Ho, and A. Rinaldo. On least squares estimation in softmax gating mixture of experts. *arXiv preprint arXiv:2402.02952*, 2024c. (Cited on page 9.)

H. Nguyen, T. Nguyen, K. Nguyen, and N. Ho. Towards convergence rates for parameter estimation in Gaussian-gated mixture of experts. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, 2024d. (Cited on page 9.)

X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370 – 400, 2013. doi: 10.1214/12-AOS1065. (Cited on page 2.)

R. Patra and B. Sen. Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):869–893, 2016. (Cited on page 2.)

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 2.)

Q. Pham, G. Do, H. Nguyen, T. Nguyen, C. Liu, M. Sartipi, B. T. Nguyen, S. Ramasamy, X. Li, S. Hoi, and N. Ho. Competesmoe – effective training of sparse mixture of experts via competition, 2024. (Cited on page 2.)

J. Puigcerver, C. R. Ruiz, B. Mustafa, C. Renggli, A. S. Pinto, S. Gelly, D. Keysers, and N. Houlsby. Scalable Transfer Learning with Expert Models. In *International Conference on Learning Representations*, 2021. (Cited on page 2.)

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 2.)

S. van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000. (Cited on pages 9, 21, 22, and 23.)

N. Verzelen and E. Arias-Castro. Detection and feature selection in sparse mixture models. *Annals of Statistics*, 45(5):1920–1950, 2017. (Cited on page 1.)

C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003. (Cited on page 2.)

C. Villani. *Optimal transport: Old and New*. Springer, 2008. (Cited on page 2.)

X. Xia, W. Yang, J. Ren, Y. Li, Y. Zhan, B. Han, and T. Liu. Pluralistic image completion with gaussian mixture models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24087–24100. Curran Associates, Inc., 2022. (Cited on page 2.)

Z. You, S. Feng, D. Su, and D. Yu. Speechmoe2: Mixture-of-experts model with improved routing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7217–7221, 2022. (Cited on page 2.)

S. Zuo, Q. Zhang, C. Liang, P. He, T. Zhao, and W. Chen. Moebert: from bert to mixture-of-experts

via importance-guided adaptation. *The North American Chapter of the Association for Computational Linguistics*, 2023. (Cited on page 2.)

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, Section 1]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes, Section 1]
   (b) Complete proofs of all theoretical results. [Yes, they are in the supplementary material]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes, Section 4]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes, Section 4]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials for "On Parameter Estimation in Deviated Gaussian Mixture of Experts"

In this supplementary material, we first provide proofs for main results in Appendix A, while leaving those for auxiliary results in Appendix B. Then, we present the convergence behavior of parameter estimation under the full overlap regime in Appendix C. Finally, we study the parameter estimation problem in the deviated Gaussian mixture of experts when the true mixing proportion vanishes, i.e. $\lambda^* = 0$, in Appendix D.

A PROOF OF MAIN RESULTS

In this appendix, we provide the proof of Theorem 1 in Appendix A.1, and then present that for Theorem 2 in Appendix A.2.

To begin with, let us define some essential notations that will be used in our arguments. For any vectors $u = (u_1, u_2, \ldots, u_d) \in \mathbb{R}^d$ and $q = (q_1, q_2, \ldots, q_d) \in \mathbb{N}^d$, we denote $u^q := u_1^{q_1} u_2^{q_2} \ldots u_d^{q_d}$, $|u| := u_1 + u_2 + \ldots + u_d$ and $q! := q_1! q_2! \ldots q_d!$. Next, the two expert functions considered in this work are denoted by $h_1(X, a, b) = a^\top X + b$ and $h_2(X, \sigma) = \sigma$, for any $(a, b, \sigma) \in \Theta$ and $X \in \mathcal{X}$. Finally, for any set A, we denote A^c as its complement.

A.1 Proof of Theorem 1

According to the result in Proposition 2, in order to reach the desired conclusion, we need to demonstrate that $V(p_{\hat{\lambda}_n, \hat{\mathfrak{b}}_n}, p_{\lambda^*, G_*}) \,\&\, D_1((\hat{\lambda}_n, \widehat{G}_n), (\lambda^*, G_*))$, which follows from the following inequality:

$$\inf_{\lambda \in [0,1], G \in \mathcal{G}_{k,\xi}(\Theta)} \frac{V(p_{\lambda,G}, p_{\lambda^*,G_*})}{D_1((\lambda, G), (\lambda^*, G_*))} > 0. \tag{15}$$

For that purpose, we split the above inequality into two parts, which we referred to local inequality and global inequality. Note that in the above infimum is subject to mixing measures in the set $\mathcal{G}_{k,\xi}(\Theta) := \{G = \sum_{i=1}^{k'} p_i \delta_{(a_i, b_i, \sigma_i)} : 1 \leq k' \leq k,\ p_i \geq \xi,\ (a_i, b_i, \sigma_i) \in \Theta\}$ for some $\xi > 0$ for simplicity.

Local inequality. Firstly, we will show the following local inequality:

$$\lim_{\varepsilon \to 0} \inf_{\substack{\lambda \in [0,1], G \in \mathcal{G}_{k,\xi}(\Theta): \\ D_1((\lambda, G), (\lambda^*, G_*)) \leq \varepsilon}} \frac{V(p_{\lambda,G}, p_{\lambda^*,G_*})}{D_1((\lambda, G), (\lambda^*, G_*))} > 0. \tag{16}$$

Assume by contrary that the above claim does not hold, then there exist a sequence of mixing measures $G_n = \sum_{i=1}^{k_n} p_i^n \delta_{(a_i^n, b_i^n, \sigma_i^n)} \in \mathcal{G}_{k,\xi}(\Theta)$ and a sequence of mixing proportions $\lambda_n \in [0, 1]$ that satisfy

$$\begin{cases} D_{1n} := D_1((\lambda_n, G_n), (\lambda^*, G_*)) \to 0, \\ V(p_{\lambda_n, G_n}, p_{\lambda^*, G_*}) / D_{1n} \to 0, \end{cases}$$

as $n \to \infty$. Next, the following Voronoi cells with respect to $\widehat{G}_n$ is defined as:

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

where $\theta_i^n := (a_i^n, b_i^n, \sigma_i^n)$ and $\theta_j^* := (a_j^*, b_j^*, \sigma_j^*)$. Since $k_n \leq k$ for all n, we can find a subsequence of G_n such that k_n does not change with n. Then, by replacing G_n with this subsequence, we may assume that $k_n = k$ for all $n \in \mathbb{N}$. Additionally, $\mathcal{A}_j = \mathcal{A}_j^n$ does not change with n for all $j \in [k_*]$, either. As $D_{1n} \to 0$, we can represent $G_n = \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} p_i^n \delta_{(a_i^n, b_i^n, \sigma_i^n)}$ such that $|\mathcal{A}_j| \geq 1$ for all $j \in [k_*]$ and $\sum_{j=1}^{k_*} |\mathcal{A}_j| = k$. Furthermore, it follows from the formulation of metric D_1 in equation (8) that $\lambda_n \to \lambda^*$, $(a_i^n, b_i^n, \sigma_i^n) \to (a_j^*, b_j^*, \sigma_j^*)$ for any $i \in \mathcal{A}_j$ and $\sum_{i \in \mathcal{A}_j} p_i^n \to p_j^*$ for any $j \in [k_*]$ when n tends to infinity.

Step 1 - Application of Taylor expansion. Now, we consider the following quantity:

$$\begin{aligned}
&p_{\lambda_n,G_n}(X,Y) - p_{\lambda^*,G_*}(X,Y)\\
&= \sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\lambda_n p_i^n[f(Y|(a_i^n)^\top X + b_i^n,\sigma_i^n) - f(Y|(a_j^*)^\top X + b_j^*,\sigma_j^*)]\bar{f}(X)\\
&+ \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\lambda_n p_i^n[f(Y|(a_i^n)^\top X + b_i^n,\sigma_i^n) - f(Y|(a_j^*)^\top X + b_j^*,\sigma_j^*)]\bar{f}(X)\\
&+ \sum_{j=1}^{k_*}\left(\sum_{i\in\mathcal{A}_j}\lambda_n p_i^n - \lambda^* p_j^*\right) f(Y|(a_j^*)^\top X + b_j^*,\sigma_j^*)\bar{f}(X) + (\lambda^* - \lambda_n)g_0(Y|X)\bar{f}(X)\\
&:= A_{n,1} + A_{n,2} + B_n + C_n. \qquad (17)
\end{aligned}$$

For the sake of presentation, let us denote $\Delta a_{ij}^n := a_i^n - a_j^*$, $\Delta b_{ij}^n := b_i^n - b_j^*$ and $\Delta\sigma_{ij}^n := \sigma_i^n - \sigma_j^*$ for all $i \in [k_n]$ and $j \in [k_*]$. For each $j \in [k_*]$ such that $|\mathcal{A}_j| > 1$, by applying the Taylor expansion up to order $\bar{r}(|\mathcal{A}_j|)$, we can rewrite $f(Y|(a_i^n)^\top X + b_i^n,\sigma_i^n) - f(Y|(a_j^*)^\top X + b_j^*,\sigma_j^*)$ as

$$\begin{aligned}
&\sum_{|\alpha|=1}^{\bar{r}(|\mathcal{A}_j|)}\frac{1}{\alpha!}(\Delta a_{ij}^n)^{\alpha_1}(\Delta b_{ij}^n)^{\alpha_2}(\Delta\sigma_{ij}^n)^{\alpha_3}\frac{\partial^{|\alpha_1|+\alpha_2+\alpha_3}f}{\partial a^{\alpha_1}\partial b^{\alpha_2}\partial\sigma^{\alpha_3}}(Y|(a_j^*)^\top X + b_j^*,\sigma_j^*) + R_{1ij}(Y|X)\\
&= \sum_{|\alpha|=1}^{\bar{r}(|\mathcal{A}_j|)}\frac{1}{\alpha!}(\Delta a_{ij}^n)^{\alpha_1}(\Delta b_{ij}^n)^{\alpha_2}(\Delta\sigma_{ij}^n)^{\alpha_3}\frac{X^{\alpha_1}}{2^{\alpha_3}}\frac{\partial^{|\alpha_1|+\alpha_2+2\alpha_3}f}{\partial h_1^{|\alpha_1|+\alpha_2+2\alpha_3}}(Y|(a_j^*)^\top X + b_j^*,\sigma_j^*) + R_{1ij}(Y|X),
\end{aligned}$$

where $R_{1ij}(Y|X)$ is a Taylor remainder term such that $R_{1ij}(X,Y)/D_{1n} \to 0$, and the first equality comes from the following partial differential equation (PDE):

$$\frac{\partial^{\alpha_3}f}{\partial h_2^{\alpha_3}}(Y|(a_j^*)^\top X + b_j^*,\sigma_j^*) = \frac{1}{2^{\alpha_3}}\cdot\frac{\partial^{2\alpha_3}f}{\partial h_1^{2\alpha_3}}(Y|(a_j^*)^\top X + b_j^*,\sigma_j^*).$$

As a result, let $\ell = \alpha_2 + 2\alpha_3$, then $A_{n,1}$ can be represented as

$$A_{n,1} = \sum_{j:|\mathcal{A}_j|>1}\sum_{|\alpha_1|=0}^{\bar{r}(|\mathcal{A}_j|)}\sum_{\ell=0}^{2(\bar{r}(|\mathcal{A}_j|)-|\alpha_1|)} E_{\alpha_1,\ell}^n(j)X^{\alpha_1}\cdot\frac{\partial^{|\alpha_1|+\ell}f}{\partial h_1^{|\alpha_1|+\ell}}(Y|(a_j^*)^\top X + b_j^*,\sigma_j^*)\bar{f}(X) + R_1(X,Y),$$

where $R_1(X,Y) := \sum_{j=1}^{k_*}\sum_{i\in\mathcal{A}_j}R_{1ij}(Y|X)\bar{f}(X)$, which leads to $R_1(X,Y)/D_{1n} \to 0$ as $n \to \infty$. In addition, the coefficients $E_{\alpha_1,\ell}^n(j)$ in this representation are defined as

$$E_{\alpha_1,\ell}^n(j) := \sum_{i\in\mathcal{A}_j}\sum_{\substack{\alpha_2+2\alpha_3=\ell\\ \alpha_2+\alpha_3\geq 1-|\alpha_1|}}\frac{\lambda_n p_i^n}{2^{\alpha_3}\alpha!}\cdot(\Delta a_{ij}^n)^{\alpha_1}(\Delta b_{ij}^n)^{\alpha_2}(\Delta\sigma_{ij}^n)^{\alpha_3},$$

for any $j \in [k_*]$, $0 \leq |\alpha_1| \leq \bar{r}(|\mathcal{A}_j|)$ and $0 \leq \ell \leq 2(\bar{r}(|\mathcal{A}_j|) - |\alpha_1|)$.

Similarly, by means of Taylor expansion up to the first order, $A_{n,2}$ is decomposed as

$$A_{n,2} = \sum_{j:|\mathcal{A}_j|=1}\sum_{|\alpha_1|=0}^{1}\sum_{\ell=0}^{2(1-|\alpha_1|)} E_{\alpha_1,\ell}^n(j)X^{\alpha_1}\cdot\frac{\partial^{|\alpha_1|+\ell}f}{\partial h_1^{|\alpha_1|+\ell}}(Y|(a_j^*)^\top X + b_j^*,\sigma_j^*)\bar{f}(X) + R_2(X,Y),$$

where $R_2(X,Y)$ is a Taylor remainder term such that $R_2(X,Y)/D_{1n} \to 0$ as $n \to \infty$.

Note that three terms $A_{n,1}$, $A_{n,2}$, B_n and C_n can be viewed as linear combinations of elements of the set $\mathcal{H}_1$ defined as

$$\begin{aligned}
\mathcal{H}_1 := \Bigg\{&X^{\alpha_1}\cdot\frac{\partial^{|\alpha_1|+\ell}f}{\partial h_1^{|\alpha_1|+\ell}}(Y|(a_j^*)^\top X + b_j^*,\sigma_j^*)\bar{f}(X),\ g_0(Y|X)\bar{f}(X):\ j\in[k_*],\\
&0\leq|\alpha_1|\leq\bar{r}(|\mathcal{A}_j|),\ 0\leq\ell\leq 2(\bar{r}(|\mathcal{A}_j|)-|\alpha_1|)\Bigg\}. \qquad (18)
\end{aligned}$$

Step 2 - Non-vanishing coefficients. In this step, we will show by contradiction that not all the coefficients in the representations of $A_{n,1}/D_{1n}$, $A_{n,2}/D_{1n}$, B_n/D_{1n} and C_n/D_{1n} vanish as $n \to \infty$. In particular, assume that all of them vanish converge to zero. Given this hypothesis, it can be seen from the definitions of C_n and B_n in equation (17) that

$$\frac{(\lambda^* - \lambda_n)}{D_{1n}} \to 0, \qquad \frac{1}{D_{1n}} \cdot \sum_{j=1}^{k_*} \Big| \sum_{i \in \mathcal{A}_j} \lambda_n p_i^n - \lambda^* p_j^* \Big| \to 0. \tag{19}$$

Regarding the coefficients in $A_{n,2}$, by considering the limits of $E^n_{\mathbf{0}_d,1}(j)/D_{1n}$ and $E^n_{\alpha_1,0}(j)/D_{1n}$ for $j \in [k_*] : |\mathcal{A}_j| = 1$ and $\alpha_1 \in \{e_1, e_2, \ldots, e_d\}$, where $e_u := (0, \ldots, 0, \underbrace{1}_{u\text{-}th}, 0, \ldots, 0)$ being a one-hot vector in $\mathbb{R}^d$ for any $u \in [d]$, we obtain that

$$\frac{1}{D_{1n}} \cdot \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \lambda_n p_i^n \Big(\|\Delta a_{ij}^n\|_1 + |\Delta b_{ij}^n| + |\Delta \sigma_{ij}^n| \Big) \to 0.$$

Due to the topological equivalence of 1-norm and 2-norm, the above limit is equivalent to

$$\frac{1}{D_{1n}} \cdot \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \lambda_n p_i^n \Big(\|\Delta a_{ij}^n\| + |\Delta b_{ij}^n| + |\Delta \sigma_{ij}^n| \Big) \to 0. \tag{20}$$

Regarding the coefficients in $A_{n,1}$, it follows from the limits of $E^n_{\alpha_1,0}(j)/D_{1n}$ for any $\alpha_1 \in \{2e_1, 2e_2, \ldots, 2e_d\}$ and $j \in [k_*] : |\mathcal{A}_j| > 1$ that

$$\frac{1}{D_{1n}} \cdot \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \lambda_n p_i^n \|\Delta a_{ij}^n\|^2 \to 0.$$

Putting the results in equations (19) and (20) together with the formulation of D_{1n} that

$$\frac{1}{D_{1n}} \cdot \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \lambda_n p_i^n \Big(|\Delta b_{ij}^n|^{\bar{r}(|\mathcal{A}_j|)} + |\Delta \sigma_{ij}^n|^{\bar{r}(|\mathcal{A}_j|)/2} \Big) \to 1.$$

Therefore, we can find an index $j^* \in [k_*] : |\mathcal{A}_{j^*}| > 1$ such that

$$\frac{1}{D_{1n}} \cdot \sum_{i \in \mathcal{A}_{j^*}} \lambda_n p_i^n \Big(|\Delta b_{ij^*}^n|^{\bar{r}(|\mathcal{A}_{j^*}|)} + |\Delta \sigma_{ij^*}^n|^{\bar{r}(|\mathcal{A}_{j^*}|)/2} \Big) \not\to 0.$$

WLOG, we assume that $j^* = 1$ throughout this proof. Moreover, since $E^n_{\mathbf{0}_d,\ell}(1)/D_{1n} \to 0$ as $n \to \infty$ for any $1 \leq \ell \leq \bar{r}(|\mathcal{A}_1|)$, we deduce that

$$\frac{\sum_{i \in \mathcal{A}_1} \sum_{\alpha_2 + 2\alpha_3 = \ell} p_i^n \cdot \frac{(\Delta b_{i1}^n)^{\alpha_2} (\Delta \sigma_{i1}^n)^{\alpha_3}}{2^{\alpha_3} \alpha_2! \alpha_3!}}{\sum_{i \in \mathcal{A}_1} p_i^n \Big(|\Delta b_{i1}^n|^{\bar{r}(|\mathcal{A}_1|)} + |\Delta \sigma_{i1}^n|^{\bar{r}(|\mathcal{A}_1|)/2} \Big)} \to 0, \tag{21}$$

for any $1 \leq \ell \leq \bar{r}(|\mathcal{A}_1|)$. Subsequently, we denote

$$\overline{M}_n = \max\{|\Delta b_{i1}^n|, \ |\Delta \sigma_{i1}^n|^{1/2} : i \in \mathcal{A}_1\}, \quad \overline{p}_n = \max_{i \in \mathcal{A}_1} p_i^n.$$

Since the sequence $p_i^n/\overline{p}_n$ is bounded, we can substitute it by its subsequence which admits a non-negative limit $s_i^2 = \lim_{n \to \infty} p_i^n/\overline{p}_n$. Furthermore, as $p_i^n \geq \xi > 0$ for all $i \in \mathcal{A}_1$, at least one among the limit s_i^2 is equal to 1. Similarly, let $(\Delta b_{i1}^n)/\overline{M}_n \to t_{1i}$ and $(\Delta \sigma_{i1}^n)/(2\overline{M}_n^2) \to t_{2i}$ as $n \to \infty$ for any $i \in \mathcal{A}_1$. Then, at least one among t_{1i} and t_{2i} for $i \in \mathcal{A}_1$ is equal to either 1 or -1.

Then, we divide both the numerator and the denominator of the ratio in equation (21) by $\overline{p}_n \overline{M}_n^{\ell}$, and obtain the following system of polynomial equations:

$$\sum_{i \in \mathcal{A}_1} \sum_{\alpha_2 + 2\alpha_3 = \ell} \frac{s_i^2 \ t_{1i}^{\alpha_2} \ t_{2i}^{\alpha_3}}{\alpha_2! \ \alpha_3!} = 0, \quad \forall \ell = 1, 2, \ldots, \bar{r}(|\mathcal{A}_1|).$$

From the definition of $\bar{r}(|\mathcal{A}_1|)$, this system does not have any non-trivial solutions, which contradicts to the aforementioned properties of s_i, t_{1i} and t_{2i}. Consequently, not all the coefficients in the representations of $A_{n,1}/D_{1n}$, $A_{n,2}/D_{1n}$, B_n/D_{1n} and C_n/D_{1n} go to 0 as $n \to \infty$.

Step 3 - Collapse of coefficients by Fatou's lemma. In this step, we will point out a contradiction to the result in Step 2 by using Fatou's lemma. In particular, let us denote by m_n the maximum of the absolute values of the coefficients in the representations of $A_{n,1}/D_{1n}$, $A_{n,2}/D_{1n}$, B_n/D_{1n} and C_n/D_{1n}, i.e.

$$m_n = \max_{\substack{j\in[k_*],\ 0\leq|\alpha_1|\leq\bar{r}(|\mathcal{A}_j|),\\ 0\leq\ell\leq 2(\bar{r}(|\mathcal{A}_j|)-|\alpha_1|)}} \left\{\frac{|E^n_{\alpha_1,\ell}(j)|}{D_{1n}}, \frac{|\lambda_n-\lambda^*|}{D_{1n}}\right\},$$

with a note that $E^n_{\mathbf{0}_d,0}(j) := \sum_{i\in\mathcal{A}_j}\lambda_n p_i^n - \lambda^* p_j^*$. Since $|E^n_{\alpha_1,\ell}(j)|/(m_n D_{1n})$ and $|\lambda_n-\lambda^*|/(m_n D_{1n})$ are bounded, we can replace them by their subsequences such that

$$\frac{|E^n_{\alpha_1,\ell}(j)|}{m_n D_{1n}} \to \tau_{\alpha_1,\ell}(j), \qquad \frac{|\lambda_n-\lambda^*|}{m_n D_{1n}} \to \tau,$$

as $n\to\infty$ for all $j\in[k_*]$, $0\leq|\alpha_1|\leq\bar{r}(|\mathcal{A}_j|)$ and $0\leq\ell\leq 2(\bar{r}(|\mathcal{A}_j|)-|\alpha_1|)$. Here, at least one among $\tau_{\alpha_1,\ell}(j)$ and τ is different from zero. By applying the Fatou's lemma, we get

$$0 = \lim_{n\to\infty}\frac{2V(p_{\lambda_n,G_n},p_{\lambda^*,G_*})}{m_n D_{1n}} \geq \int \liminf_{n\to\infty}\frac{|p_{\lambda_n,G_n}(X,Y)-p_{\lambda^*,G_*}(X,Y)|}{m_n D_{1n}}\mathrm{d}(X,Y)\geq 0,$$

which implies that

$$\frac{|p_{\lambda_n,G_n}(X,Y)-p_{\lambda^*,G_*}(X,Y)|}{m_n D_{1n}} \to 0,$$

for almost surely (X,Y). Recall that the left hand side in the above equation converges to

$$\sum_{j,\alpha_1,\ell}\tau_{\alpha_1,\ell}(j)X^{\alpha_1}\cdot\frac{\partial^{|\alpha_1|+\ell}f}{\partial h_1^{|\alpha_1|+\ell}}(Y|(a_j^*)^\top X+b_j^*,\sigma_j^*)\bar{f}(X)+\tau g_0(Y|X)\bar{f}(X),$$

where the ranges of (j,α_1,ℓ) in the summation are $j\in[k_*]$, $0\leq|\alpha_1|\leq\bar{r}(|\mathcal{A}_j|)$ and $0\leq\ell\leq 2(\bar{r}(|\mathcal{A}_j|)-|\alpha_1|)$. As a result, we get

$$\sum_{j,\alpha_1,\ell}\tau_{\alpha_1,\ell}(j)X^{\alpha_1}\cdot\frac{\partial^{|\alpha_1|+\ell}f}{\partial h_1^{|\alpha_1|+\ell}}(Y|(a_j^*)^\top X+b_j^*,\sigma_j^*)+\tau g_0(Y|X)=0, \tag{22}$$

for almost surely (X,Y). Since p_{G_*} is distinguishable from g_0, it follows from Definition 1 that $\tau_{\alpha_1,\ell}(j)X^{\alpha_1}=\tau=0$, for any $j\in[k_*]$, $0\leq|\alpha_1|\leq\bar{r}(|\mathcal{A}_j|)$ and $0\leq\ell\leq 2(\bar{r}(|\mathcal{A}_j|)-|\alpha_1|)$ for almost surely X. This result indicates that $\tau_{\alpha_1,\ell}(j)=\tau=0$, which contradicts to the fact that at least one among $\tau_{\alpha_1,\ell}(j),\tau$ is non-zero. Hence, we obtain the local inequality in equation (16).

As a consequence, there exists some $\varepsilon'>0$ such that

$$\inf_{\substack{\lambda\in[0,1],G\in\mathcal{G}_{k,\xi}(\Theta):\\ D_1((\lambda,G),(\lambda^*,G_*))\leq\varepsilon'}} V(p_{\lambda,G},p_{\lambda^*,G_*})/D_1((\lambda,G),(\lambda^*,G_*))>0.$$

Global inequality: Thus, it remains to prove the following global inequality:

$$\inf_{\substack{\lambda\in[0,1],G\in\mathcal{G}_{k,\xi}(\Theta):\\ D_1((\lambda,G),(\lambda^*,G_*))>\varepsilon'}} V(p_{\lambda,G},p_{\lambda^*,G_*})/D_1((\lambda,G),(\lambda^*,G_*))>0.$$

Assume by contrary that it is not the case. Then, there exist some sequences $G'_n\in\mathcal{G}_{k,\xi}(\Theta)$ and $\lambda'_n\in[0,1]$ such that

$$V(p_{\lambda'_n,G'_n},p_{\lambda^*,G_*})/D_1((\lambda'_n,G'_n),(\lambda^*,G_*))\to 0,$$
$$D_1((\lambda'_n,G'_n),(\lambda^*,G_*))>\varepsilon'.$$

As a result, we get $V(p_{\lambda'_n,G'_n}, p_{\lambda^*,G_*}) \to 0$. Note that Θ and $[0,1]$ are bounded sets, then we can find a subsequence of G'_n and a subsequence of λ'_n such that $G'_n \to G'$ and $\lambda'_n \to \lambda'$, where $G' \in \mathcal{G}_{k,\xi}(\Theta)$ and $\lambda' \in [0,1]$. By replacing G'_n and λ'_n with these subsequences, we get that $D_1((\lambda'_n, G'_n), (\lambda^*, G_*)) > \varepsilon'$. Moreover, by the Fatou's lemma, we obtain that

$$\begin{aligned}0 = \lim_{n\to\infty} 2V(p_{\lambda'_n,G'_n}, p_{\lambda^*,G_*}) &\geq \int \liminf_{n\to\infty} \left|p_{\lambda'_n,G'_n}(X,Y) - p_{\lambda^*,G_*}(X,Y)\right| \mathrm{d}(X,Y)\\ &= \int |p_{\lambda',G'}(X,Y) - p_{\lambda^*,G_*}(X,Y)| \,\mathrm{d}(X,Y) \geq 0,\end{aligned}$$

which indicates that $p_{\lambda',G'}(X,Y) = p_{\lambda^*,G_*}(X,Y)$ for almost surely (X,Y). According to Proposition 1, the deviated Gaussian mixture of experts is identifiable when p_{G_*} is distinguishable from g_0. Thus, it follows that $(\lambda', G') \equiv (\lambda^*, G_*)$. This contradicts to the previous claim that $D_1((\lambda', G'), (\lambda^*, G_*)) > \varepsilon' > 0$. Hence, the proof is completed.

A.2 Proof of Theorem 2

Similar to the proof of Theorem 1, we need to prove the following claim:

$$\inf_{\lambda\in[0,1], G\in\mathcal{G}_{k,\xi}(\Theta)} \frac{V(p_{\lambda,G}, p_{\lambda^*,G_*})}{D_2((\lambda,G),(\lambda^*,G_*))} > 0.$$

Local inequality. Firstly, we will demonstrate the local version of the above inequality:

$$\lim_{\varepsilon\to 0} \inf_{\substack{\lambda\in[0,1], G\in\mathcal{G}_{k,\xi}(\Theta),\\ D_2((\lambda,G),(\lambda^*,G_*))\leq\varepsilon}} \frac{V(p_{\lambda,G}, p_{\lambda^*,G_*})}{D_2((\lambda,G),(\lambda^*,G_*))} > 0. \tag{23}$$

Assume by contrary that the above inequality does not hold, then there exist sequences $\lambda_n \in [0,1]$ and $G_n = \sum_{i=1}^{k_n} p_i^n \delta_{(a_i^n, b_i^n, \sigma_i^n)} \in \mathcal{G}_{k,\xi}(\Theta)$ such that

$$\begin{cases} D_{2n} := D_2((\lambda_n, G_n), (\lambda^*, G_*)) \to 0,\\ V(p_{\lambda_n,G_n}, p_{\lambda^*,G_*})/D_{2n} \to 0.\end{cases}$$

Case 1: $\lambda_n \leq \lambda^*$ for infinitely $n \in \mathbb{N}$. WLOG, we assume that $\lambda_n \leq \lambda^*$ for all $n \in \mathbb{N}$.

In this case, we have $D_{2n} = D_1((\lambda_n, G_n), (\lambda^*, G_*))$, for any $n \in \mathbb{N}$. Note that $k_n \leq k$, thus, we can replace G_n with one of its subsequences such that k_n does not vary with n. Therefore, we assume that $k_n = k$ for all n. In addition, the Voronoi cells $\mathcal{A}_j = \mathcal{A}_j^n$ does not change with n for all $j \in [k_*]$. Next, we decompose the quantity $p_{\lambda_n,G_n}(X,Y) - p_{\lambda^*,G_*}(X,Y)$ as follows:

$$\begin{aligned}p_{\lambda_n,G_n}(X,Y) - p_{\lambda^*,G_*}(X,Y) &= (\lambda^* - \lambda_n) \sum_{j=\bar{k}+1}^{k_0} p_j^0 f(Y|(a_j^0)^\top X + b_j^0, \sigma_j^0) \bar{f}(X)\\ &+ \sum_{j:|\mathcal{A}_j|>1} \sum_{i\in\mathcal{A}_j} \lambda_n p_i^n [f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n) - f(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*)] \bar{f}(X)\\ &+ \sum_{j:|\mathcal{A}_j|=1} \sum_{i\in\mathcal{A}_j} \lambda_n p_i^n [f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n) - f(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*)] \bar{f}(X)\\ &+ \sum_{j=1}^{k_*} \left(\sum_{i\in\mathcal{A}_j} \lambda_n p_i^n - \bar{p}_j^*(\lambda_n)\right) f(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) \bar{f}(X)\\ &:= C_n + A_{n,1} + A_{n,2} + B_n,\end{aligned}$$

where we define $\bar{p}_j^*(\lambda_n) := \begin{cases} \lambda^* p_j^* + (\lambda_n - \lambda^*) p_j^0, & j \in [\bar{k}]\\ \lambda^* p_j^*, & j \in [k_*] \setminus [\bar{k}]\end{cases}$.

By applying the Taylor expansions as in Appendix A.1, we are able to show that $A_{n,1}/D_{2n}$, $A_{n,2}/D_{2n}$, B_n/D_{2n} and C_n/D_{2n} can be written as linear combinations of elements of the following set

$$\mathcal{H}_2 := \Bigg\{ X^{\alpha_1} \cdot \frac{\partial^{|\alpha_1|+\ell} f}{\partial h_1^{|\alpha_1|+\ell}}(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*)\bar{f}(X),\ f(Y|(a_{j'}^0)^\top X + b_{j'}^0, \sigma_{j'}^0)\bar{f}(X) : j \in [k_*], \\ j' \in [k_0] \setminus [\bar{k}],\ 0 \leq |\alpha_1| \leq \bar{r}(|\mathcal{A}_j|),\ 0 \leq \ell \leq 2(\bar{r}(|\mathcal{A}_j|) - |\alpha_1|) \Bigg\}. \tag{24}$$

Furthermore, not all the coefficients in these representations go to zero as n tends to infinity.

Subsequently, by following the same arguments for deriving equation (22), we can find some constants $\tau_{\alpha_1,\ell}(j)$ and $\tau(j')$, where $j \in [k_*]$, $0 \leq |\alpha_1| \leq \bar{r}(|\mathcal{A}_j|)$, $0 \leq \ell \leq 2(\bar{r}(|\mathcal{A}_j|) - |\alpha_1|)$ and $j' \in [k_0] \setminus [\bar{k}]$, such that at least one among them is non-zero and

$$\sum_{j,\alpha_1,\ell} \tau_{\alpha_1,\ell}(j) X^{\alpha_1} \cdot \frac{\partial^{|\alpha_1|+\ell} f}{\partial h_1^{|\alpha_1|+\ell}}(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) + \sum_{j'=\bar{k}+1}^{k_0} \tau(j') f(Y|(a_{j'}^0)^\top X + b_{j'}^0, \sigma_{j'}^0) = 0, \tag{25}$$

for almost surely (X, Y). Now, we demonstrate that the set $\mathcal{H}_2$ is linearly independent with respect to X and Y, or equivalently, $\tau_{\alpha_1,\ell}(j) = \tau(j') = 0$, for any $j \in [k_*]$, $0 \leq |\alpha_1| \leq \bar{r}(|\mathcal{A}_j|)$, $0 \leq \ell \leq 2(\bar{r}(|\mathcal{A}_j|) - |\alpha_1|)$ and $j' \in [k_0] \setminus [\bar{k}]$. Indeed, equation (25) can be rewritten as

$$\sum_{j=1}^{k_*} \sum_{v=0}^{2\bar{r}(|\mathcal{A}_j|)} \Bigg(\sum_{|\alpha_1|+\ell=v} \tau_{\alpha_1,\ell}(j) X^{\alpha_1} \Bigg) \frac{\partial^v f}{\partial h_1^v}(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) + \sum_{j'=\bar{k}+1}^{k_0} \tau(j') f(Y|(a_{j'}^0)^\top X + b_{j'}^0, \sigma_{j'}^0) = 0, \tag{26}$$

for almost surely X and Y. It is worth noting that $(a_j^*, b_j^*, \sigma_j^*)$ and $(a_{j'}^0, b_{j'}^0, \sigma_{j'}^0)$ are distinct components for $j \in [k_*]$ and $j \in [k_0] \setminus [\bar{k}]$, therefore, $((a_j^*)^\top X + b_j^*, \sigma_j^*)$ and $((a_{j'}^0)^\top X + b_{j'}^0, \sigma_{j'}^0)$ are distinct pairs for almost surely $X \in \mathcal{X}$. This implies that $\frac{\partial^v f}{\partial h_1^v}(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*)$ and $f(Y|(a_{j'}^0)^\top X + b_{j'}^0, \sigma_{j'}^0)$ are linearly independent with respect to Y for $0 \leq v \leq 2\bar{r}(|\mathcal{A}_j|)$ for any $j \in [k_*]$ and $j' \in [k_0] \setminus [\bar{k}]$. Then, it follows from equation (26) that $\tau(j') = 0$ for any $j' \in [k_0] \setminus [\bar{k}]$ and $\sum_{|\alpha_1|+\ell=u} \tau_{\alpha_1,\ell}(j) X^{\alpha_1} = 0$ for any $j \in [k_*]$, $0 \leq v \leq 2\bar{r}(|\mathcal{A}_j|)$ for almost surely X. Note that this is a polynomial of $X \in \mathcal{X}$, which is a bounded subset of $\mathbb{R}^d$, we deduce that $\tau_{\alpha_1,\ell}(j) = 0$ for all $|\alpha_1| + \ell = v$, $j \in [k_*]$ and $0 \leq v \leq 2\bar{r}(|\mathcal{A}_j|)$. This contradicts the previous claim that at least one among $\tau_{\alpha_1,\ell}(j), \tau(j')$ is different from zero.

Thus, we obtain the local inequality in equation (23) for this case.

Case 2: $\lambda_n > \lambda^*$ for infinitely $n \in \mathbb{N}$. WLOG, we assume that $\lambda_n > \lambda^*$ for all $n \in \mathbb{N}$.

Case 2.1: $k \leq k_* + k_0 - \bar{k} - 1$

In this case, the discrepancy D_{2n} reduces to $D_1((\lambda_n, G_n), (\lambda^*, G_*))$. Therefore, the local inequality for this case can be achieved analogously to that for Case 1.

Case 2.2: $k \geq k_* + k_0 - \bar{k}$

In this case, the discrepancy D_{2n} equals to $D_3(G_n, \overline{G}_*(\lambda_n))$, which was defined in equation (12). Additionally, we have

$$\begin{aligned} p_{\lambda_n,G_n}(X,Y) - p_{\lambda^*,G_*}(X,Y) &= \lambda_n \Bigg\{ \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} p_i^n f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n) \\ &- \Big[\Big(1 - \frac{\lambda^*}{\lambda_n}\Big) \sum_{j=1}^{k_0} p_j^0 f(Y|(a_j^0)^\top X + b_j^0, \sigma_j^0) + \frac{\lambda^*}{\lambda_n} \sum_{j=1}^{k_*} p_j^* f(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) \Big] \Bigg\} \bar{f}(X) \\ &= \lambda_n \Big[p_{G_n}(X,Y) - p_{\overline{G}_*(\lambda_n)}(X,Y) \Big]. \end{aligned}$$

Recall that

$$0 = \lim_{n\to\infty} \frac{2V(p_{\lambda_n,G_n}, p_{\lambda^*,G_*})}{D_{2n}} = \lim_{n\to\infty} \frac{\int |p_{\lambda_n,G_n}(X,Y) - p_{\lambda^*,G_*}(X,Y)|d(X,Y)}{D_3(G_n, \overline{G}_*(\lambda_n))}$$
$$= \lim_{n\to\infty} \lambda_n \cdot \frac{\int |p_{G_n}(X,Y) - p_{\overline{G}_*(\lambda_n)}(X,Y)|d(X,Y)}{D_3(G_n, \overline{G}_*(\lambda_n))}$$
$$= \lim_{n\to\infty} \lambda_n \cdot \frac{2V(p_{G_n}, p_{\overline{G}_*(\lambda_n)})}{D_3(G_n, \overline{G}_*(\lambda_n))}$$

Since $\lambda_n > \lambda^* > 0$ for all $n \in \mathbb{N}$, we get $V(p_{G_n}, p_{\overline{G}_*(\lambda_n)})/D_3(G_n, \overline{G}_*(\lambda_n)) \to 0$ as $n \to \infty$. For the sake of presentation, we represent the mixing measure $\overline{G}_*(\lambda_n) = \Big(1 - \frac{\lambda^*}{\lambda_n}\Big)G_0 + \frac{\lambda^*}{\lambda_n}G_*$ as

$$\overline{G}_*(\lambda_n) = \sum_{j=1}^{k_*+k_0-\bar{k}} (p_j^n)'\delta_{(a_j',b_j',\sigma_j')} \in \mathcal{E}_{k_*+k_0-\bar{k}}(\Theta).$$

Next, let us define Voronoi cells used for this case as follows:

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

where $\theta_i^n = (a_i^n, b_i^n, \sigma_i^n)$ and $\theta_j' = (a_i', b_i', \sigma_i')$ for any $j \in [k_* + k_0 - \bar{k}]$. Since $k_n \leq k$, there exists a subsequence of G_n such that k_n does not vary with n. Thus, by replacing G_n with this subsequence, we can assume that $k_n = k$ for all n. In addition, $\mathcal{B}_j = \mathcal{B}_j^n$ does not change with n for all $j \in [k_* + k_0 - \bar{k}]$. Then, we can rewrite the difference $p_{G_n}(X,Y) - p_{\overline{G}_*(\lambda_n)}(X,Y)$ as follows:

$$p_{G_n}(X,Y) - p_{\overline{G}_*(\lambda_n)}(X,Y) = \sum_{j:|\mathcal{B}_j|>1|} \sum_{i\in\mathcal{B}_j} p_i^n \Big[f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n) - f(Y|(a_j')^\top X + b_j', \sigma_j')\Big]$$
$$+ \sum_{j:|\mathcal{B}_j|=1|} \sum_{i\in\mathcal{B}_j} p_i^n \Big[f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n) - f(Y|(a_j')^\top X + b_j', \sigma_j')\Big]$$
$$+ \sum_{j=1}^{k_*+k_0-\bar{k}} \Big(\sum_{i\in\mathcal{B}_j} p_i^n - p_j'\Big) f(Y|(a_j')^\top X + b_j', \sigma_j')$$

By abuse of notation, we denote three terms in the above summation as $A_{n,1}$, $A_{n,2}$ and B_n, respectively. By invoking the Taylor expansions as in Appendix A.1, we get that $A_{n,1}$, $A_{n,2}$ and B_n can be treated as linear combinations of elements of the following set:

$$\mathcal{H} := \Bigg\{ X^{\alpha_1} \cdot \frac{\partial^{|\alpha_1|+\ell} f}{\partial h_1^{|\alpha_1|+\ell}}(Y|(a_j')^\top X + b_j', \sigma_j')\bar{f}(X) : \ j \in [k_* + k_0 - \bar{k}], \ 0 \leq |\alpha_1| \leq \bar{r}(|\mathcal{B}_j|),$$
$$0 \leq \ell \leq 2(\bar{r}(|\mathcal{B}_j|) - |\alpha_1|)\Bigg\}.$$

Moreover, not all the coefficients in the representations of $A_{n,1}/D_{2n}$, $A_{n,2}/D_{2n}$ and B_n/D_{2n} approach zero as $n \to \infty$. Additionally, we can utilize the same arguments for deriving equation (22) to deduce that there exist some constants $\tau_{\alpha_1,\ell}(j)$, where $j \in [k_* + k_0 - \bar{k}]$, $0 \leq |\alpha_1| \leq \bar{r}(|\mathcal{B}_j|)$ and $0 \leq \ell \leq 2(\bar{r}(|\mathcal{B}_j|) - |\alpha_1|)$ that satisfy at least one among them is different from zero and

$$\sum_{j,\alpha_1,\ell} \tau_{\alpha_1,\ell}(j) X^{\alpha_1} \cdot \frac{\partial^{|\alpha_1|+\ell} f}{\partial h_1^{|\alpha_1|+\ell}}(Y|(a_j')^\top X + b_j', \sigma_j') = 0,$$

for almost surely (X,Y). Since $\mathcal{H}$ is a linearly independent set, which can be proved in a similar way as for the set $\mathcal{H}_2$ in Case 1, the above equation leads to $\tau_{\alpha_1,\ell}(j)$ for any $j \in [k_* + k_0 - \bar{k}]$, $0 \leq |\alpha_1| \leq \bar{r}(|\mathcal{B}_j|)$ and $0 \leq \ell \leq 2(\bar{r}(|\mathcal{B}_j|) - |\alpha_1|)$. This contradicts with the result that at least one among $\tau_{\alpha_1,\ell}(j)$ is non-zero. Hence, we achieve the local inequality in equation (23).

As a consequence, there exists a positive constant ε' that satisfies

$$\inf_{\substack{\lambda\in[0,1],G\in\mathcal{G}_{k,\xi}(\Theta),\\ D_2((\lambda,G),(\lambda^*,G_*))\leq\varepsilon}} \frac{V(p_{\lambda,G},p_{\lambda^*,G_*})}{D_2((\lambda,G),(\lambda^*,G_*))}>0.$$

Global inequality. Now, it suffices to show that

$$\inf_{\substack{\lambda\in[0,1],G\in\mathcal{G}_{k,\xi}(\Theta),\\ D_2((\lambda,G),(\lambda^*,G_*))>\varepsilon'}} \frac{V(p_{\lambda,G},p_{\lambda^*,G_*})}{D_2((\lambda,G),(\lambda^*,G_*))}>0. \tag{27}$$

Assume by contrary that the above claim does not hold, then there exist sequences $(\lambda'_n)\subset[0,1]$ and $(G'_n)\subset\mathcal{G}_{k,\xi}(\Theta)$ that satisfy

$$\begin{cases} D_2((\lambda'_n,G'_n),(\lambda^*,G_*))>\varepsilon', \\ V(p_{\lambda'_n,G'_n},p_{\lambda^*,G_*})/D_2((\lambda'_n,G'_n),(\lambda^*,G_*))\to 0, \end{cases}$$

which leads to the fact that $V(p_{\lambda'_n,G'_n},p_{\lambda^*,G_*})\to 0$ as $n\to\infty$.

Case 1: $\lambda'_n\leq\lambda^*$ for infinitely $n\in\mathbb{N}$. WLOG, we assume that $\lambda'_n\leq\lambda^*$ for all $n\in\mathbb{N}$.

In this case, we have $D_2((\lambda'_n,G'_n),(\lambda^*,G_*))=D_1((\lambda'_n,G'_n),(\lambda^*,G_*))>\varepsilon'$. Since the sets Θ and $[0,1]$ are bounded, we can find a subsequence of G'_n and a subsequence of λ'_n such that $G'_n\to G'$ and $\lambda'_n\to\lambda'$, where $G'\in\mathcal{G}_{k,\xi}(\Theta)$ and $\lambda'\in[0,1]$. By replacing G'_n and λ'_n with those subsequences, we get that $D_1((\lambda',G'),(\lambda^*,G_*))>\varepsilon'$. On the other hand, the result that $V(p_{\lambda'_n,G'_n},p_{\lambda^*,G_*})\to 0$ as $n\to\infty$ implies that $V(p_{\lambda',G'},p_{\lambda^*,G_*})=0$, which leads to

$$p_{\lambda',G'}(X,Y)=p_{\lambda^*,G_*}(X,Y),$$

for almost surely (X,Y). Since $\lambda'_n\leq\lambda^*$, we get $\lambda'\leq\lambda^*$. It is worth noting that if $\lambda'<\lambda^*$, $\overline{G}_*(\lambda')$ is not valid mixing measure. Therefore, we obtain $(\lambda',G')\equiv(\lambda^*,G_*)$ in this scenario. If $\lambda'=\lambda^*$, then $\overline{G}_*(\lambda')\equiv G_*$ and the above identifiability equation also admits $(\lambda',G')\equiv(\lambda^*,G_*)$ as a solution. Thus, it follows that $D_1((\lambda',G'),(\lambda^*,G_*))=0$, which contradicts the result that $D_1((\lambda',G'),(\lambda^*,G_*))>\varepsilon'>0$. Hence, the global inequality (27) holds true in this case.

Case 2: $\lambda'_n>\lambda^*$ for infinitely $n\in\mathbb{N}$. WLOG, we assume that $\lambda'_n>\lambda^*$ for all $n\in\mathbb{N}$.

Case 2.1: $k\leq k_*+k_0-\bar{k}-1$

In this case, we also have $D_2((\lambda'_n,G'_n),(\lambda^*,G_*))=D_1((\lambda'_n,G'_n),(\lambda^*,G_*))>\varepsilon'$. Similar to Case 1, we get $p_{\lambda',G'}(X,Y)=p_{\lambda^*,G_*}(X,Y)$ for almost surely (X,Y), where $\lambda'\in[0,1]$ and $G'\in\mathcal{G}_{k,\xi}(\Theta)$ are the limits of λ'_n and G'_n as n goes to infinity, respectively.

Under this setting, the previous identifiability equation admits either $(\lambda',G')\equiv(\lambda^*,G_*)$ or $(\lambda',G')\equiv(\lambda',\overline{G}_*(\lambda'))$ for any $\lambda'\in[0,1]$ as a solution as mentioned in Section 3.1. However, as $\overline{G}_*(\lambda')$ has $k_*+k_0-\bar{k}$ components, which is higher than that of G' which has no more than k components. As a result, we obtain that $(\lambda',G')\equiv(\lambda^*,G_*)$, leading to $D_1((\lambda',G'),(\lambda^*,G_*))=0$, which is a contradiction to the fact that $D_1((\lambda',G'),(\lambda^*,G_*))>\varepsilon'>0$.

Case 2.2: $k\geq k_*+k_0-\bar{k}$

In this case, we have $D_2((\lambda'_n,G'_n),(\lambda^*,G_*))=D_3(G'_n,\overline{G}(\lambda'_n))>\varepsilon'$. Similar to Case 1, we may replace G'_n and λ'_n with their subsequences whose limits are $G'\in\mathcal{G}_{k,\xi}(\Theta)$ and $\lambda'\in[0,1]$, respectively. Then, we get $D_3(G',\overline{G}(\lambda'))>\varepsilon'$. Additionally, we also achieve the identifiability equation $p_{\lambda',G'}(X,Y)=p_{\lambda^*,G_*}(X,Y)$ for almost surely (X,Y). Note that in this case, G' has more components than G_*, then the previous equation admits only $(\lambda',\overline{G}(\lambda'))$ as a solution. Therefore, we obtain that $D_3(G',\overline{G}(\lambda'))=0$, which contradicts the result that $D_3(G',\overline{G}(\lambda'))>\varepsilon'>0$.

Hence, we reach the global inequality in equation (27), and the proof is totally completed.

B PROOF OF AUXILIARY RESULTS

In this appendix, we provide proofs for Proposition 1 and Proposition 2 in Appendix B.1 and Appendix B.2, respectively.

B.1 Proof of Proposition 1

Firstly, we suppose that $G = \sum_{i=1}^{k} p_i \delta_{(a_i, b_i, \sigma_i)}$ and $G' = \sum_{i=1}^{k'} p'_i \delta_{(a'_i, b'_i, \sigma'_i)}$ are two mixing measures such that the equation $p_{\lambda,G}(X,Y) = p_{\lambda',G'}(X,Y)$ holds for almost surely $(X,Y) \in \mathcal{X} \times \mathcal{Y}$. This equation can be expanded as

$$(\lambda' - \lambda) g_0(Y|X) + \sum_{i=1}^{k} \lambda p_i f(Y|a_i^\top X + b_i, \sigma_i) - \sum_{i=1}^{k'} \lambda' p'_i f(Y|(a'_i)^\top X + b'_i, \sigma'_i) = 0,$$

for almost surely (X,Y). Now, assume that G and G' share only ℓ components in common, where $0 \leq \ell \leq \min\{k, k'\}$, e.g. $(a_i, b_i, \sigma_i) = (a'_i, b'_i, \sigma'_i)$ for any $i \in [\ell]$. Then, we rewrite the above equation as

$$\begin{aligned}(\lambda' - \lambda) g_0(Y|X) + \sum_{i=1}^{\ell} (\lambda p_i - \lambda' p'_i) f(Y|a_i^\top X + b_i, \sigma_i) + \sum_{i=\ell+1}^{k} \lambda p_i f(Y|a_i^\top X + b_i, \sigma_i) \\ - \sum_{i=\ell+1}^{k'} \lambda' p'_i f(Y|(a'_i)^\top X + b'_i, \sigma'_i) = 0,\end{aligned}$$

for almost surely (X,Y). Next, we consider a mixing measure G_ℓ which has $k + k' - \ell$ components $(a_1, b_1, \sigma_1), \ldots, (a_k, b_k, \sigma_k)$, $(a'_{\ell+1}, b'_{\ell+1}, \sigma'_{\ell+1}), \ldots, (a'_{k'}, b'_{k'}, \sigma'_{k'})$. Since p_{G_ℓ} is distinguishable from g_0, we see that if either $k \neq k'$ or $0 \leq \ell < \min\{k, k'\}$, then there exists an index i such that $p_i = 0$ and/or $p'_i = 0$, which contradicts to the fact that $p_i, p'_i > 0$. As a result, we must have $k = k'$ and $\ell = k$, i.e. $(a_i, b_i, \sigma_i) = (a'_i, b'_i, \sigma'_i)$ for all $i \in [k]$. Given this result, the above equation is equivalent to

$$(\lambda' - \lambda) g_0(Y|X) + \sum_{i=1}^{k} (\lambda p_i - \lambda' p'_i) f(Y|a_i^\top X + b_i, \sigma_i) = 0,$$

for almost surely (X,Y). Note that p_G is distinguishable from g_0, then we obtain that $\lambda - \lambda' = \lambda p_i - \lambda' p'_i = 0$, which implies that $\lambda = \lambda'$ and $p_i = p'_i$ for any $i \in [k]$. As a result, it follows that $(\lambda, G) \equiv (\lambda', G')$.

Hence, the proof is completed.

B.2 Proof of Proposition 2

First of all, let us introduce some necessary notations used throughout this appendix. In particular, we denote $\mathcal{P}_K([0,1] \times \Theta) := \{p_{\lambda,G}(X,Y) : \lambda \in [0,1],\ G \in \mathcal{G}_{K,\xi}(\Theta)\}$. Additionally, we define

$$\begin{aligned}\overline{\mathcal{P}}_K([0,1] \times \Theta) &= \{\bar{p}_{\lambda,G} = (p_{\lambda,G} + p_{\lambda^*,G_*})/2 : (\lambda, G) \in [0,1] \times \mathcal{G}_{k,\xi}(\Theta)\}, \\ \overline{\mathcal{P}}_K^{1/2}([0,1] \times \Theta) &= \{\bar{p}_{\lambda,G}^{1/2} : \bar{p}_{\lambda,G} \in \overline{\mathcal{P}}_K([0,1] \times \Theta)\}.\end{aligned}$$

To derive a convergence rate for the joint density estimators under the Hellinger distance, we require a condition on the complexity of the following class introduced in van de Geer (2000):

$$\overline{\mathcal{P}}_K^{1/2}([0,1] \times \Theta, \varepsilon) := \{\bar{p}_{\lambda,G}^{1/2} \in \overline{\mathcal{P}}_K^{1/2}(\Theta) :\ h(\bar{p}_{\lambda,G}, p_{\lambda^*,G_*}) \leq \varepsilon\},$$

The complexity of this class can be captured by the following bracketing entropy integral

$$\mathcal{J}_B(\varepsilon, \overline{\mathcal{P}}_K^{1/2}([0,1] \times \Theta, \varepsilon), m) = \int_{\varepsilon^2/2^{13}}^{\varepsilon} \sqrt{\log H_B(u, \overline{\mathcal{P}}_K^{1/2}([0,1] \times \Theta, \varepsilon), m)} du \vee \varepsilon,$$

where $u \vee \varepsilon = \max\{u, \varepsilon\}$ and $H_B(\varepsilon, \mathcal{P}, m)$ represents the ε-bracketing entropy of a set $\mathcal{P}$ under the Lebesgue measure m (readers are referred to van de Geer (2000) for more detail about the definition of this term). Interestingly, we discover a connection between this quantity and the convergence of the density estimators as follows:

Lemma 1. *Given a universal constant $J > 0$, assume that we can find a natural number N, possibly depending on Θ and k, such that for all $n \geq N$ and $\varepsilon > \sqrt{\log n / n}$, the following holds:*

$$\mathcal{J}_B(\varepsilon, \overline{\mathcal{P}}_K^{1/2}([0,1] \times \Theta, \varepsilon), m) \leq J \sqrt{n} \varepsilon^2. \tag{28}$$

Then, there exists a positive constant c that depends only on Θ such that for all $n \in \mathbb{N}$, we have

$$\mathbb{P}\left(h(p_{\hat{\lambda}_n,\hat{\mathfrak{G}}_n}, p_{\lambda^*,G_*}) > \delta\right) \leq c\exp\left(-\frac{n\delta^2}{c^2}\right).$$

Proof of Lemma 1 is in Appendix B.2.3. As a consequence, in order to guarantee that our estimators will converge, it is sufficient to satisfy the condition in equation (28). For that purpose, we need to introduce a result regarding the upper bounds of the covering number $N(\varepsilon, \mathcal{P}_k(\Theta,[0,1]), \|\cdot\|_\infty)$ and the bracketing entropy $H_B(\varepsilon, \mathcal{P}_k(\Theta,[0,1]), h)$ of the metric space $\mathcal{P}_k(\Theta,[0,1])$ in the following lemma:

Lemma 2. *Suppose that Θ_1 and Θ_2 are respectively two bounded subsets of $\mathbb{R}^{q_1}$ and $\mathbb{R}^{q_2}$. Then, for any $0 < \varepsilon < 1/2$, the following results hold*

(i) $\log N\left(\varepsilon, \mathcal{P}_k(\Theta\times[0,1]), \|.\|_\infty\right) \;.\; \log(1/\varepsilon)$,

(ii) $H_B(\varepsilon, \mathcal{P}_k([0,1]\times\Theta), h) \;.\; \log(1/\varepsilon)$.

Proof of Lemma 2 is in Appendix B.2.2. Now, we are ready to provide the proof of Proposition 2 in Appendix B.2.1.

B.2.1 Main Proof

Note that $\overline{\mathcal{P}}_k^{1/2}([0,1]\times\Theta,\delta) \subseteq \overline{\mathcal{P}}_k^{1/2}([0,1]\times\Theta)$, it follows from the definition of Hellinger distance that

$$\begin{aligned} H_B(\delta, \overline{\mathcal{P}}_k^{1/2}([0,1]\times\Theta,\delta), \|\cdot\|_2) &\leq H_B(\delta, \overline{\mathcal{P}}_k^{1/2}([0,1]\times\Theta), \|\cdot\|_2) \\ &= H_B\Big(\frac{\delta}{\sqrt{2}}, \overline{\mathcal{P}}_k([0,1]\times\Theta), h\Big) \\ &\leq H_B(\delta, \mathcal{P}_k([0,1]\times\Theta), h). \end{aligned}$$

According to part (ii) of Lemma 2, we find that

$$H_B(\delta, \overline{\mathcal{P}}_k^{1/2}([0,1]\times\Theta,\delta), \|\cdot\|_2) \;.\; \log(1/\delta).$$

Consequently, we deduce that

$$\mathcal{J}_B(\varepsilon, \overline{\mathcal{P}}_k^{1/2}([0,1]\times\Theta,\delta), u) \;.\; \varepsilon[\log(2^{13}\varepsilon^2)]^{1/2} < n\varepsilon^2,$$

for all $\varepsilon > \sqrt{\log n/n}$. By applying Lemma 1 with $\delta = \sqrt{\log n/n}$, we obtain the desired result.

B.2.2 Proof of Lemma 1

It follows from Lemma 4.1 and Lemma 4.2 in van de Geer (2000) that

$$\frac{1}{16}h^2(p_{\hat{\lambda}_n,\hat{\mathfrak{G}}_n}, p_{\lambda^*,G_*}) \leq h^2(\bar{p}_{\hat{\lambda}_n,\hat{\mathfrak{G}}_n}, p_{\lambda^*,G_*}) \leq \frac{1}{\sqrt{n}}\nu_n(\hat{\lambda}_n, \widehat{G}_n),$$

where $\nu(\hat{\lambda}_n, \widehat{G}_n)$ is an empirical process defined as

$$\nu_n(\hat{\lambda}_n, \widehat{G}_n) := \sqrt{n}\int_{p_{\lambda^*,G_*}>0} \frac{1}{2}\log\Big(\frac{\overline{p}_{\hat{\lambda}_n,\hat{\mathfrak{G}}_n}}{p_{\lambda^*,G_*}}\Big)\cdot[\overline{p}_{\hat{\lambda}_n,\hat{\mathfrak{G}}_n} - p_{\lambda^*,G_*}]\mathrm{d}(X,Y).$$

Thus, for any $\delta > \delta_n := \sqrt{\log n/n}$, we obtain

$$\begin{aligned} &\mathbb{P}_{\lambda^*,G_*}(h(p_{\hat{\lambda}_n,\hat{\mathfrak{G}}_n}, p_{\lambda^*,G_*}) \geq \delta) \\ &\leq \mathbb{P}_{\lambda^*,G_*}\Big(\nu_n(\hat{\lambda}_n, \widehat{G}_n) - \sqrt{n}h^2(p_{\hat{\lambda}_n,\hat{\mathfrak{G}}_n,p_{\lambda^*,G_*}}) \geq 0, h(p_{\hat{\lambda}_n,\hat{\mathfrak{G}}_n}, p_{\lambda^*,G_*}) \geq \delta/4\Big) \\ &\leq \mathbb{P}_{\lambda^*,G_*}\Big(\sup_{\lambda,G:h(\bar{p}_{\lambda,G},p_{\lambda^*,G_*})\geq\delta/4}[\nu_n(\lambda G) - \sqrt{n}h^2(\bar{p}_{\lambda,G}, p_{\lambda^*,G_*})] \geq 0\Big) \\ &\leq \sum_{s=0}^{S}\mathbb{P}_{\lambda^*,G_*}\Big(\sup_{\lambda,G:2^s\delta/4\leq h(\bar{p}_{\lambda,G},p_{\lambda^*,G_*})\leq 2^{s+1}\delta/4}|\nu_n(\lambda G)| \geq \sqrt{n}2^{2s}(\delta/4)^2\Big) \\ &\leq \sum_{s=0}^{S}\mathbb{P}_{\lambda^*,G_*}\Big(\sup_{\lambda,G:h(\bar{p}_{\lambda,G},p_{\lambda^*,G_*})\leq 2^{s+1}\delta/4}|\nu_n(\lambda G)| \geq \sqrt{n}2^{2s}(\delta/4)^2\Big), \end{aligned}$$

where S is the smallest number such that $2^S\delta/4 > 1$. Now, we proceed by recalling Theorem 5.11 in van de Geer (2000) with adapted notations to our setting as follows:

Lemma 3 (Theorem 5.11 in van de Geer (2000)). *Let $R > 0$, $k \geq 1$, and $\mathcal{G}_{k,\xi}(\Theta) \subset \mathcal{G}_{k,\xi}(\Theta)$ containing G_*. Let C be sufficiently large, then for all $n \in \mathbb{N}$ and $C_0, C_1, t > 0$ that satisfy*

(i) $t \leq 8\sqrt{n}R \vee C_1\sqrt{n}R^2/K$;

(ii) $t \geq C^2(C_1+1)\left(R \wedge \int_{t/(2^6\sqrt{n})}^{R} H_B^{1/2}\Big(\frac{u}{\sqrt{2}}, \overline{\mathcal{P}}_K^{1/2}([0,1]\times\Theta, R), \nu\Big)du\right)$,

we get

$$\mathbb{P}_{\lambda^*,G_*}\left(\sup_{G\in\mathcal{G}_{k,\xi}(\Theta), h(\bar{p}_{\lambda,G}, p_{\lambda^*,G_*})\leq R} |\nu_n(\lambda G)| \geq t\right) \leq C\exp\left[-\frac{t^2}{C^2(C_1+1)R^2}\right].$$

Proof of Lemma 3 can be found in van de Geer (2000).

Back to our proof, by choosing $R = 2^{s+1}\delta$, $C_1 = 15$ and $t = \sqrt{n}2^{2s}(\delta/4)^2$, we can verify that condition (i) in Lemma 3 is satisfied as $2^{s-1}\delta/4 \leq 1$ for all $s \leq S$. Meanwhile, the condition (ii) is met as

$$\begin{aligned}\int_{t/(2^6\sqrt{n})}^{R} H_B^{1/2}\left(\frac{u}{\sqrt{2}}, \overline{\mathcal{P}}_K^{1/2}([0,1]\times\Theta, R), \nu\right)du \vee 2^{s+1}\delta &= \sqrt{2}\int_{R^2/2^{13}}^{R/\sqrt{2}} H_B^{1/2}\left(u, \overline{\mathcal{P}}_K^{1/2}([0,1]\times\Theta, R), \nu\right)du \vee 2^{s+1}\delta \\ &\leq 2\mathcal{J}_B(R, \overline{\mathcal{P}}_K^{1/2}([0,1]\times\Theta, R), \nu) \\ &\leq 2J\sqrt{n}2^{2s+1}\delta^2 = 2^6 Jt.\end{aligned}$$

Applying Lemma 3, we get that

$$\mathbb{P}_{\lambda^*,G_*}(h(p_{\tilde{\lambda}_n,\tilde{G}_n}, p_{\lambda^*,G_*}) > \delta) \leq C\sum_{s=0}^{\infty}\exp\left(\frac{2^{2s}n\delta^2}{J^2 2^{14}}\right) \leq c\exp\left(-\frac{n\delta^2}{c}\right).$$

Hence, the proof is completed.

B.2.3 Proof of Lemma 2

Part (i). For any set S, we denote $\mathcal{E}_\varepsilon(S)$ an ε-net of S if each element of S is within ε distance from some elements of $\mathcal{E}_\varepsilon(S)$. By definition of the covering number, we get $|\mathcal{E}_\varepsilon(S)| = N(\varepsilon, S, \|\cdot\|_\infty)$. Let $\mathcal{P}_k(\Theta) := \{p_G : G \in \mathcal{G}_k(\Theta)\}$, where $p_G(X,Y) := \sum_{i=1}^k p_i f(Y|a_i^\top X + b_i, \sigma_i)$. According to [Lemma 6, Ho et al. (2022)], we have

$$\log|\mathcal{E}_\varepsilon(\mathcal{P}_k(\Theta))| = N(\varepsilon, \mathcal{P}_k(\Theta), \|\cdot\|_\infty) \;.\; \log(1/\varepsilon).$$

Let $\mathcal{G}_{k,\xi}(\Theta) := \{\widetilde{G} : p_{\tilde{G}} \in \mathcal{E}_\varepsilon(\mathcal{P}_k(\Theta))\}$ be the set of all latent mixing measures G in the net $\mathcal{E}_\varepsilon(\mathcal{P}_k(\Theta))$. We will show that

$$\mathcal{E}_\varepsilon(\mathcal{P}_k([0,1]\times\Theta)) \subseteq \{p_{\lambda,G} : \tilde{\lambda} \in \mathcal{E}_\varepsilon([0,1]), \widetilde{G} \in \mathcal{G}_{k,\xi}(\Theta)\}. \tag{29}$$

Indeed, for any $\lambda \in [0,1], G \in \mathcal{G}_k(\Theta)$, there exist $\tilde{\lambda} \in \mathcal{E}_\varepsilon([0,1]), \widetilde{G} \in \mathcal{G}_{k,\xi}(\Theta)$ such that $|\lambda - \tilde{\lambda}| \leq \varepsilon$ and $\|p_G - p_{\widetilde{G}}\|_\infty \leq \varepsilon$, which leads to

$$\begin{aligned}\|p_{\lambda,G} - p_{\tilde{\lambda}\widetilde{G}}\|_\infty &\leq \|p_{\lambda,G} - p_{\tilde{\lambda}G}\|_\infty + \|p_{\tilde{\lambda}G} - p_{\tilde{\lambda}\widetilde{G}}\|_\infty \\ &= |\lambda - \tilde{\lambda}|\|g_0 - p_G\|_\infty + \tilde{\lambda}\|p_G - p_{\widetilde{G}}\|_\infty \\ &\leq \varepsilon(\|g_0\|_\infty + \|p_G\|_\infty) + \varepsilon \\ &\;.\; \varepsilon.\end{aligned}$$

Therefore, we obtain equation (29). Putting the above results together with a note that $\log(|\mathcal{E}_\varepsilon([0,1])|) \leq \log(1/\varepsilon)$, we have

$$\log(N(\varepsilon, \mathcal{P}_k(\Theta\times[0,1]), \|\cdot\|_\infty)) \leq \log(|\mathcal{E}_\varepsilon([0,1])|) + \log(|\mathcal{E}_\varepsilon(\mathcal{P}_k(\Theta))|) \;.\; \log(1/\varepsilon).$$

Hence, we reach the conclusion in part (i).

Part (ii). Firstly, let $\eta \leq \varepsilon$ be some positive number that we will chose later. Since f is the density function of an univariate location-scale Gaussian distribution, we can verify for any $|Y| \geq 2a$ and $X \in \mathcal{X}$ that

$$f(Y|h_1(X,\theta_1), h_2(X,\theta_2)) \leq \frac{1}{\sqrt{2\pi}\ell} \exp\left(-Y^2/(8u^2)\right).$$

Recall that $\log g_0(Y|X) \ . \ -Y^p$ and $g_0(Y|X) \leq M$ for some positive constants $M, p > 0$. Let $q = \min\{p, 2\}$, $C_2 = \max\left\{M, \frac{1}{\sqrt{2\pi}\ell}\right\}$ and

$$H(X,Y) = \begin{cases} C_1 \exp\left(-Y^q\right) \overline{f}(X), & \text{for } |Y| \geq 2a \\ C_2 \overline{f}(X), & \text{for } |Y| < 2a, \end{cases} \tag{30}$$

where $C_1 > 0$ is a constant depending on ℓ, g_0. Thus, it can be shown that $H(X,Y)$ is an envelope of $\mathcal{P}_k([0,1] \times \Theta)$. Subsequently, we denote by $g_1, \ldots, g_N$ an η-net over $\mathcal{P}_k([0,1] \times \Theta)$. Then, we construct the brackets $[p_i^L(X,Y), p_i^U(X,Y)]$ as follows:

$$p_i^L(X,Y) := \max\{g_i(X,Y) - \eta, 0\}, \ p_i^U(X,Y) := \max\{g_i(X,Y) + \eta, H(X,Y)\}$$

for $1 \leq i \leq N$. As a result, $\mathcal{P}_k([0,1] \times \Theta) \subset \cup_{i=1}^N [p_i^L(X,Y), p_i^U(X,Y)]$ and $p_i^U(X,Y) - p_i^L(X,Y) \leq \min\{2\eta, H(X,Y)\}$. It follows that ,

$$\begin{aligned} &\int \left(p_i^U(X,Y) - p_i^L(X,Y)\right) d(X,Y) \\ &\leq \int\limits_{|Y|<2a} \left(p_i^U(X,Y) - p_i^L(X,Y)\right) d(X,Y) + \int\limits_{|Y|\geq 2a} \left(p_i^U(X,Y) - p_i^L(X,Y)\right) d(X,Y) \\ &\leq \int\limits_{|Y|<2a} 2\eta d(X,Y) + \int\limits_{|Y|\geq 2a} H(X,Y) d(X,Y) \leq c\eta, \end{aligned} \tag{31}$$

where c is some positive universal constant. This implies that

$$H_B(c\eta, \mathcal{P}_k([0,1] \times \Theta), \|.\|_1) \leq N \ . \ \log(1/\eta).$$

By choosing $\eta = \varepsilon/c$, we have

$$H_B(\varepsilon, \mathcal{P}_k([0,1] \times \Theta), \|.\|_1) \ . \ \log(1/\varepsilon).$$

Due to the inequality $h^2 \leq \|.\|_1$ between Hellinger distance and total variational distance, we reach the conclusion of bracketing entropy bound.

C PARAMETER ESTIMATION UNDER THE FULL OVERLAP REGIME

In this appendix, we study the convergence rates of parameter estimation in the deviated Gaussian mixture of experts under the full overlap regime, namely when the function $g_0(Y|X)$ takes the following form:

$$g_0(Y|X) = p_{G_0}(Y|X) := \sum_{j=1}^{k_0} p_j^0 f(Y|(a_j^0)^\top X + b_j^0, \sigma_j^0), \tag{32}$$

and $\bar{k} = k_0$, where $\bar{k}$ stands for the number of overlapped components of two mixing measures G_0 and G_*.

Under this regime, it is worth noting that if $G_* = G_0$, then the conditional density function $p_{\lambda^* G_*}(Y|X)$ is reduced to $p_{\lambda^* G_*}(Y|X) = (1-\lambda^*) p_{G_0}(Y|X) + \lambda^* p_{G_0}(Y|X) = p_{G_0}(Y|X)$, which coincides with the setting $\lambda^* = 0$ that we will consider in Appendix D. For that reason, we assume that $G_* \neq G_0$ throughout this appendix.

Identiability of the deviated Gaussian mixture of experts. Similar to the partial overlap regime studied in Section 3.2, the deviated Gaussian mixture of experts under the full overlap regime is also not identifiable. Furthermore, it is even more challenging to solve the equation $p_{\lambda,G}(X,Y) = p_{\lambda^*,G_*}(X,Y)$ for almost surely (X,Y) in this regime. In particular, we have to take into account the following set of mixing proportions λ which make the term $\overline{G}_*(\lambda) = \frac{\lambda-\lambda^*}{\lambda}G_0 + \frac{\lambda^*}{\lambda}G_*$ defined in equation (10) a valid mixing measure:

$$\mathcal{T} := \{\lambda \in (0,1] : (\lambda^* - \lambda)p_i^0 \leq \lambda^* p_i^*, \forall i \in [k_0]\}.$$

Here, the set $\mathcal{T}$ contains $\lambda \in (0,1]$ such that the weights associated with components of $\overline{G}_*(\lambda)$ are non-negative, i.e. $\bar{p}_i^*(\lambda) \geq 0$ where

$$\bar{p}_i^*(\lambda) := \begin{cases} [(\lambda - \lambda^*)p_i^0 + \lambda^* p_i^*]/\lambda, & i \in [k_0], \\ \lambda^* p_i^*/\lambda, & i \in [k_*] \setminus [k_0]. \end{cases}$$

Subsequently, we solve the identifiability equation in two complement scenarios of λ with respect to the set $\mathcal{T}$.

When $\lambda \in \mathcal{T}$: Since $\overline{G}_*(\lambda)$ is valid mixing measure in this case, the identifiability equation can be rewritten as $\lambda[p_G(X,Y) - p_{\overline{G}_*(\lambda)}(X,Y)] = 0$ for almost surely (X,Y). Moreover, as $\overline{G}_*(\lambda)$ has $k_* + k_0 - \bar{k} = k_*$ components and $k > k_*$, that equation admits $(\lambda, \overline{G}_*(\lambda))$ as a solution for any $\lambda \in \mathcal{T}$. Additionally, it is worth noting that (λ^*, G_*) is a special instance of $(\lambda, \overline{G}_*(\lambda))$ when $\lambda = \lambda^* \in \mathcal{T}$.

When $\lambda \in \mathcal{T}^c$: In this case, there are some components of $\overline{G}_*(\lambda)$ having negative weights $\bar{p}_i^*(\lambda) < 0$. Thus, it is necessary to inspect such components by considering the following set:

$$I_\lambda := \{i \in [k_0] : (\lambda^* - \lambda)p_i^0 > \lambda^* p_i^*\},$$

which includes indices $i \in [k_0]$ such that $\bar{p}_i^*(\lambda) < 0$. Here, we say that I_λ is *ratio-independent* if $|I_\lambda| = 1$ or $p_i^0/p_i^* = p_j^0/p_j^*$ for all $i, j \in I_\lambda$ if $|I_\lambda| \geq 2$. An intuition behind this definition is to guarantee that all the terms $(\lambda^* - \lambda)p_i^0 - \lambda^* p_i^*$, for $i \in I_\lambda$, can be arbitrary small simultaneously. Consequently, when I_λ is a ratio-independent set, mixing measures of the following form are solutions of the identifiability equation:

$$\widetilde{G}_*(\lambda) := \frac{1}{s(\lambda)}\left(\sum_{i \in I_\lambda^c}\Big[p_i^*\lambda^* + (\lambda - \lambda^*)p_i^0\Big]\delta_{(a_i^0, b_i^0, \sigma_i^0)} + \sum_{i=k_0+1}^{k_*}\lambda^* p_i^* \delta_{(a_i^*, b_i^*, \sigma_i^*)}\right), \tag{33}$$

where $s(\lambda) := \sum_{i \in I_\lambda^c}\Big[p_i^*\lambda^* + (\lambda - \lambda^*)p_i^0\Big] + \sum_{i=k_0+1}^{k_*}\lambda^* p_i^*$ is a normalizing term. It can be seen from the above formulation that components of the mixing measure $\widetilde{G}_*(\lambda)$ are those of $\overline{G}_*(\lambda)$ with positive weights.

Voronoi loss function. Now, we are ready to define the Voronoi loss function $D_4((\lambda, G), (\lambda^*, G_*))$ for the full overlap regime as follows:

$$D_4((\lambda, G), (\lambda^*, G_*)) = \begin{cases} \mathbf{1}_{\{\lambda \in \mathcal{T}^c\}} s(\lambda) D_3(G, \widetilde{G}_*(\lambda)) + \mathbf{1}_{\{\lambda \in \mathcal{T}\}} D_3(G, \overline{G}_*(\lambda)), & \text{if } I_\lambda \text{ is ratio-independent;} \\ \mathbf{1}_{\{\lambda \in \mathcal{T}^c\}} \sum_{i \in I_\lambda}[-\lambda p_i'(\lambda)] + \mathbf{1}_{\{\lambda \in \mathcal{T}\}} D_3(G, \overline{G}_*(\lambda)), & \text{otherwise,} \end{cases} \tag{34}$$

where the loss function D_3 is given in Section 3.2.

Given the above loss of interest, we establish the convergence rate of the MLE under the full overlap regime in the following theorem.

Theorem 3. *Assume that $\lambda^* \in (0,1]$ is unknown, and let g_0 take the form in equation (32) with $\bar{k} = k_0$. Then, we achieve that $V(p_{\lambda,G}, p_{\lambda^*,G_*})$ & $D_4((\lambda, G), (\lambda^*, G_*))$ for any $(\lambda, G) \in [0,1] \times \mathcal{G}_k(\Theta)$. This bound together with Proposition 2 indicate that*

$$\mathbb{P}(D_4((\hat{\lambda}_n, \widehat{G}_n), (\lambda^*, G_*)) > C_4\sqrt{\log(n)/n}) \ . \ n^{-c_4},$$

where $C_4 > 0$ is a constant depending on $g_0, \lambda^*, G_*, \Theta$, while the constant c_4 depends only on Θ.

Proof of Theorem 3 is deferred to Appendix C.1. When $\hat{\lambda}_n \in \mathcal{T}$, the formulation of $D_4((\lambda, G), (\lambda^*, G_*))$ is simplified to that of $D_3(G, \overline{G}_*(\lambda))$. Therefore, the convergence behavior of the MLE in this case resembles the results in Theorem 2, which will not be repeated here. The difference between this theorem and its previous counterparts occurs only when $\hat{\lambda}_n \in \mathcal{T}^c$. In particular, if $I_{\hat{\lambda}_n}$ is a ratio-independent set, then the MLE $\widehat{G}_n$ converges to $\widetilde{G}_*(\hat{\lambda}_n)$ at a substantially slower rate than $\widetilde{\mathcal{O}}(n^{-1/2})$ as it depends on the convergence rate of $s(\hat{\lambda}_n)$ to zero. By contrast, if the set $I_{\hat{\lambda}_n}$ is not ratio-independent, then the discrepancy $D_4((\lambda, G), (\lambda^*, G_*))$ will not vanish as n tends to infinity. Thus, we cannot deduce any conclusions regarding the convergence rates of the MLE in this case. An underlying reason for this phenomenon is that the terms $(\lambda^* - \hat{\lambda}_n)p_i^0 - \lambda^* p_i^*$ for $i \in I_{\hat{\lambda}_n}$ cannot approach zero simultaneously as $I_{\hat{\lambda}_n}$ is not ratio-independent.

C.1 Proof of Theorem 3

Similar to previous proofs, we need to demonstrate the following inequality:

$$\inf_{\lambda \in [0,1], G \in \mathcal{G}_{k,\xi}(\Theta)} V(p_{\lambda,G}, p_{\lambda^*,G_*})/D_4((\lambda, G), (\lambda^*, G_*)) > 0. \tag{35}$$

Local inequality. Firstly, we will derive the local version of the above inequality:

$$\lim_{\varepsilon \to 0} \inf_{\substack{\lambda \in [0,1], G \in \mathcal{G}_{k,\xi}(\Theta): \\ D_4((\lambda,G),(\lambda^*,G_*)) \leq \varepsilon}} V(p_{\lambda,G}, p_{\lambda^*,G_*})/D_4((\lambda, G), (\lambda^*, G_*)) > 0. \tag{36}$$

Assume that the above claim does not hold true, then there exist a sequence of mixing measures $G_n = \sum_{i=1}^{k_n} p_i^n \delta_{(a_i^n, b_i^n, \sigma_i^n)} \in \mathcal{G}_{k,\xi}(\Theta)$ and a sequence of mixing proportions $\lambda_n \in [0, 1]$ such that

$$\begin{cases} D_{4n} := D_4((\lambda_n, G_n), (\lambda^*, G_*)) \to 0, \\ V(p_{\lambda_n G_n}, p_{\lambda^* G_*})/D_{4n} \to 0, \end{cases}$$

as $n \to \infty$. Under the setting of Theorem 3, we have $\theta_j^* = \theta_j^0$ for all $j \in [k_0]$ and

$$\begin{aligned} &p_{\lambda_n,G_n}(X, Y) - p_{\lambda^*,G_*}(X, Y) \\ &= \lambda_n \sum_{i=1}^{k_n} p_i^n f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n)\bar{f}(X) - \sum_{j=1}^{k_*} \bar{p}_j^*(\lambda_n) f(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*)\bar{f}(X), \end{aligned} \tag{37}$$

where $\bar{p}_j^*(\lambda_n) := \begin{cases} \lambda^* p_j^* + (\lambda_n - \lambda^*)p_j^0, & j \in [k_0] \\ \lambda^* p_j^*, & k_0 + 1 \leq j \leq k_* \end{cases}$.

Next, we will show that $\liminf \lambda_n$ is bounded below by some positive constant. Assume that this claim is not true, then $\lambda_n \to 0$ as $n \to \infty$. Note that

$$V(p_{\lambda,G_n}, p_{\lambda,G_*}) = \frac{V(p_{\lambda,G_n}, p_{\lambda,G_*})}{D_{4n}} \times D_{4n} \to 0.$$

Then, by the Fatou's lemma, we get that $p_{\lambda_n,G_n}(X, Y) - p_{\lambda^*,G_*}(X, Y) \to 0$ as $n \to \infty$ for almost surely (X, Y). Since $\lambda_n \to 0$ and the density $f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n)$ can be upper bounded by a function which is independent of n for almost surely (X, Y) (see the proof of part (ii) of Lemma 2 for more detail), we deduce that

$$\lambda_n \sum_{i=1}^{k_n} p_i^n f(Y|h_1(X, \theta_{1i}^n), h_2(X, \theta_{2i}^n)) \to 0.$$

It follows that $\sum_{j=1}^{k_*} \bar{p}_j^*(\lambda_n) f(Y|h_1(X, \theta_{1j}^*), h_2(X, \theta_{2j}^*)) \to 0$ as $n \to \infty$, which leads to the fact that $\bar{p}_j^*(\lambda_n) \to 0$ for all $j \in [k_*]$. This means that $p_i^* = p_i^0$ when $i \in [k_0]$ and $p_i^* = 0$ otherwise. Thus, we obtain $G_* \equiv G_0$, which is a contradiction to the assumption that $G_* \neq G_0$. Therefore, $\liminf \lambda_n$ is bounded below by some positive constant.

Subsequently, we consider two main scenarios of λ_n based on the set $\mathcal{T}$ mentioned in Section C, i.e.

$$\mathcal{T} := \{\lambda \in (0, 1] : (\lambda^* - \lambda)p_i^0 \leq \lambda^* p_i^*, \forall i \in [k_0]\}.$$

Case 1: $\lambda_n \in \mathcal{T}$ for infinitely $n \in \mathbb{N}$. WLOG, we assume that $\lambda_n \in \mathcal{T}$ for all $n \in \mathbb{N}$.

In this case, we have $D_{4n} = D_3(G_n, \overline{G}_*(\lambda_n))$, and the difference $p_{\lambda_n, G_n}(X,Y) - p_{\lambda^*, G_*}(X,Y)$ can be written as

$$\begin{aligned} p_{\lambda_n, G_n}(X,Y) - p_{\lambda^*, G_*}(X,Y) = \ & \lambda_n \Bigg\{ \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} p_i^n f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n) \\ & - \Big[\Big(1 - \frac{\lambda^*}{\lambda_n}\Big) \sum_{j=1}^{k_0} p_j^0 f(Y|(a_j^0)^\top X + b_j^0, \sigma_j^0) + \frac{\lambda^*}{\lambda_n} \sum_{j=1}^{k_*} p_j^* f(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) \Big] \Bigg\} \bar{f}(X) \\ = \ & \lambda_n \Big[p_{G_n}(X,Y) - p_{\overline{G}_*(\lambda_n)}(X,Y) \Big]. \end{aligned} \tag{38}$$

Recall that under the full overlap regime, we have $\bar{k} = k_0$, which leads to $k \geq k_* = k_* + k_0 - \bar{k}$. Moreover, since $\lambda_n \in \mathcal{T}$, we get that $\overline{G}_*(\lambda_n)$ is a valid mixing measure. Thus, by employing arguments utilized in Case 2.2 in Appendix A.2, we obtain the local inequality in equation (36) for this case.

Case 2: $\lambda_n \notin \mathcal{T}$ for infinitely $n \in \mathbb{N}$. WLOG, we assume that $\lambda_n \notin \mathcal{T}$ for all $n \in \mathbb{N}$.

For each $n \in \mathbb{N}$, since $\lambda_n \notin \mathcal{T}$, there exists an index $i \in [k_0]$ such that $\lambda^* p_i^* - (\lambda^* - \lambda_n) p_i^0 < 0$. In other words, the set $I_{\lambda_n} := \{i \in [k_0] : (\lambda^* - \lambda_n) p_i^0 > \lambda^* p_i^*\}$ is not empty and has at least one element. In addition, we also have that $\bar{p}_j^*(\lambda_n) < 0$ for any $j \in I_{\lambda_n}$ in this case.

Next, we will consider two different settings of the set I_{λ_n} as follows:

Case 2.1: I_{λ_n} is not ratio-independent.

From the formulation of metric D_3 in equation (34), we have $D_{4n} = \sum_{j \in I_{\lambda_n}} [-\bar{p}_j^*(\lambda_n)] \to 0$ in this case. Recall that we have $-\bar{p}_j^*(\lambda_n) > 0$ for all $j \in I_{\lambda_n}$, then it follows that $\bar{p}_j^*(\lambda_n) \to 0$ as $n \to \infty$ for all $j \in I_{\lambda_n}$. This leads to the fact that $p_i^*/p_i^0 = p_j^*/p_j^0$ for all $i, j \in I_{\lambda_n}$, which is a contradiction to the assumption that I_{λ_n} is not ratio-independent. Therefore, we obtain the local inequality in equation (36) for this case.

Case 2.2: I_{λ_n} is ratio-independent.

In this case, we have $D_{4n} = s(\lambda_n) D_3(G_n, \widetilde{G}_*(\lambda_n))$. Next, we will demonstrate that $s(\lambda_n) \not\to 0$ as $n \to \infty$. Assume by contrary that $s(\lambda_n) \to 0$, then $p_j^* = 0$ for all $j > k_0$ and $\bar{p}_j^*(\lambda_n) = \lambda^* p_j^* + (\lambda_n - \lambda^*) p_j^0 \to 0$ for all $j \notin I_{\lambda_n}$. Note that

$$V(p_{\lambda_n, G_n}, p_{\lambda^*, G_*}) = \frac{V(p_{\lambda_n, G_n}, p_{\lambda^*, G_*})}{D_{4n}} \times D_{4n} \to 0.$$

Therefore, by means of Fatou's lemma, we get that $p_{\lambda_n, G_n}(X,Y) - p_{\lambda^*, G_*}(X,Y) \to 0$ when $n \to \infty$. Recall that

$$\begin{aligned} & p_{\lambda_n, G_n}(X,Y) - p_{\lambda^*, G_*}(X,Y) \\ = & \sum_{j \in I_{\lambda_n}} (-\bar{p}_j^*(\lambda_n)) f(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) \bar{f}(X) + \Big[\lambda_n \sum_{i=1}^{K} p_i^n f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n) \bar{f}(X) \\ & - \sum_{j \in I_{\lambda_n}^c} \bar{p}_j^*(\lambda_n) f(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) \bar{f}(X) - \sum_{j=k_0+1}^{k_*} \bar{p}_j^*(\lambda_n) f(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) \bar{f}(X) \Big], \end{aligned}$$

we get $\bar{p}_j^* \to 0$ for all $j \in I_{\lambda_n}$ and $\lambda_n \to 0$, which is a contradiction to the result that $\liminf \lambda_n$ is bounded below by a positive constant. Thus, $s(\lambda_n) \not\to 0$ as $n \to \infty$.

From the definition of $\widetilde{G}_*(\lambda_n)$, we can rewrite it as $\widetilde{G}_*(\lambda_n) := \sum_{j \in \mathcal{J}_{\lambda_n}} \frac{\bar{p}_j^*(\lambda_n)}{s(\lambda_n)} \delta_{(a_j^*, b_j^*, \sigma_j^*)}$, where

$$\mathcal{J}_{\lambda_n} := I_{\lambda_n}^c \cup \{k_0 + 1, \ldots, k_*\}.$$

Next, we will use the following Voronoi cells to study the discrepancy $D_3(G_n, \widetilde{G}_*(\lambda_n))$:

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

for any $\forall j \in \mathcal{J}_{\lambda_n}$, where $\theta_i^n := (a_i^n, b_i^n, \sigma_i^n)$ and $\theta_j^* = (a_j^*, b_j^*, \sigma_j^*)$.

As $k_n \leq k$ for all n, there exists a subsequence of G_n such that k_n does not change with n. Thus, by replacing G_n with this subsequence, we assume that $k_n = k$ for all n. Additionally, $\mathcal{C}_j = \mathcal{C}_j^n$ does not change with n for all $j \in [k_*]$, either. Then, we rewrite the difference $p_{\lambda_n,G_n}(X,Y) - p_{\lambda^*,G_*}(X,Y)$ as follows:

$$\begin{aligned}
p_{\lambda_n,G_n}(X,Y) - p_{\lambda^*,G_*}(X,Y) &= \sum_{j\in I_{\lambda_n}} [-\bar{p}_j^*(\lambda_n)] f(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*)\bar{f}(X) \\
&+ \sum_{j:|\mathcal{C}_j|>1} \sum_{i\in\mathcal{C}_j} \lambda_n p_i^n [f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n) - f(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*)]\bar{f}(X) \\
&+ \sum_{j:|\mathcal{C}_j|=1} \sum_{i\in\mathcal{C}_j} \lambda_n p_i^n [f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n) - f(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*)]\bar{f}(X) \\
&+ \sum_{j\in\mathcal{J}_{\lambda_n}} \left(\sum_{i\in\mathcal{C}_j} \lambda_n p_i^n - \bar{p}_j^*(\lambda_n) \right) f(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*)\bar{f}(X) \\
&:= C_n + A_{n,1} + A_{n,2} + B_n.
\end{aligned}$$

For each $j \in \mathcal{J}_{\lambda_n} : |\mathcal{C}_j| > 1$, by applying the Taylor expansion up to order $\bar{r}(|\mathcal{C}_j|)$ as in Appendix A.1, we can rewrite $A_{n,1}$ as

$$A_{n,1} = \sum_{j:|\mathcal{C}_j|>1} \sum_{|\alpha_1|=0}^{\bar{r}(|\mathcal{C}_j|)} \sum_{\ell=0}^{2(\bar{r}(|\mathcal{C}_j|)-|\alpha_1|)} E_{\alpha_1,\ell}^n(j) X^{\alpha_1} \cdot \frac{\partial^{|\alpha_1|+\ell} f}{\partial h_1^{|\alpha_1|+\ell}}(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*)\bar{f}(X) + R_5(X,Y),$$

where $R_5(X,Y)$ is a Taylor remainder such that $R_5(X,Y)/D_{4n}$, and

$$E_{\alpha_1,\ell}^n(j) := \sum_{i\in\mathcal{C}_j} \sum_{\substack{\alpha_2+2\alpha_3=\ell \\ \alpha_2+\alpha_3\geq 1-|\alpha_1|}} \frac{\lambda_n p_i^n}{2^{\alpha_3}\alpha!} \cdot (\Delta a_{ij}^n)^{\alpha_1} (\Delta b_{ij}^n)^{\alpha_2} (\Delta\sigma_{ij}^n)^{\alpha_3}, \tag{39}$$

for any $j \in \mathcal{J}_{\lambda_n} : |\mathcal{C}_j| > 1$, $0 \leq |\alpha_1| \leq \bar{r}(|\mathcal{C}_j|)$ and $0 \leq \ell \leq 2(\bar{r}(|\mathcal{C}_j|) - |\alpha_1|)$.

On the other hand, by means of Taylor expansion up to the first order, we can decompose $A_{n,2}$ as

$$A_{n,2} = \sum_{j:|\mathcal{C}_j|=1} \sum_{|\alpha_1|=0}^{1} \sum_{\ell=0}^{2(1-|\alpha_1|)} E_{\alpha_1,\ell}^n(j) X^{\alpha_1} \cdot \frac{\partial^{|\alpha_1|+\ell} f}{\partial h_1^{|\alpha_1|+\ell}}(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*)\bar{f}(X) + R_6(X,Y),$$

where $R_6(X,Y)$ is a Taylor remainder term such that $R_6(X,Y)/D_{1n} \to 0$ as $n \to \infty$, and $E_{\alpha_1,\ell}(j)$ is defined similarly as in equation (39) but for $j \in \mathcal{J}_{\lambda_n} : |\mathcal{C}_j| = 1$, $0 \leq |\alpha_1| \leq 1$ and $0 \leq \ell \leq 2(\bar{r}(|\mathcal{C}_j|) - |\alpha_1|)$. Additionally, we also utilize the notation $E_{\alpha_1,\ell}^n(j)$ to denote the coefficients in C_n as $E_{\mathbf{0}_d,0}^n(j) := -\bar{p}_j^*(\lambda_n)$ for any $j \in I_{\lambda_n}$, and those in B_n as

$$E_{\mathbf{0}_d,0}^n(j) := \sum_{i\in\mathcal{C}_j} \lambda_n p_i^n - \bar{p}_j^*(\lambda_n),$$

for any $j \in \mathcal{J}_{\lambda_n}$. Therefore, $A_{n,1}$, $A_{n,2}$, B_n and C_n can be viewed as linear combinations of elements of the following set:

$$\mathcal{H}_3 := \left\{ X^{\alpha_1} \cdot \frac{\partial^{|\alpha_1|+\ell} f}{\partial h_1^{|\alpha_1|+\ell}}(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*)\bar{f}(X) : j \in \mathcal{J}_{\lambda_n},\ 0 \leq |\alpha_1| \leq \bar{r}(|\mathcal{C}_j|), 0 \leq \ell \leq 2(\bar{r}(|\mathcal{C}_j|) - |\alpha_1|), \right\}. \tag{40}$$

Assume that all the coefficients in the formulations of $A_{n,1}/D_{4n}$, $A_{n,2}/D_{4n}$, B_n/D_{4n} and C_n/D_{4n} go to zero as

$n \to \infty$. Now, we consider the following quantity:

$$\begin{aligned}
1 = \frac{D_{4n}}{D_{4n}} &= \frac{s(\lambda_n) D_3(G_n, \widetilde{G}_*(\lambda_n))}{D_{4n}} \\
&= \frac{s(\lambda_n) \sum_{j \in \mathcal{J}_{\lambda_n}} |\sum_{i \in \mathcal{C}_j} p_i^n - \bar{p}_j^*(\lambda_n)/s(\lambda_n)|}{D_{4n}} \\
&+ \frac{s(\lambda_n) \sum_{j:|\mathcal{C}_j|=1} \sum_{i \in \mathcal{C}_j} p_i^n (\|\Delta a_{ij}^n\| + |\Delta b_{ij}^n| + |\Delta \sigma_{ij}^n|)}{D_{4n}} \\
&+ \frac{s(\lambda_n) \sum_{j:|\mathcal{C}_j|>1} \sum_{i \in \mathcal{C}_j} p_i^n (\|\Delta a_{ij}^n\|^2 + |\Delta b_{ij}^n|^{\bar{r}(|\mathcal{C}_j|)} + |\Delta \sigma_{ij}^n|^{\bar{r}(|\mathcal{C}_j|)/2})}{D_{4n}}
\end{aligned} \tag{41}$$

For $j \in \mathcal{J}_{\lambda_n} : |\mathcal{C}_j| > 1$, we take summation of the limits of $E_{\alpha_1,0}(j)$, where $\alpha_1 \in \{2e_1, 2e_2, \ldots, 2e_d\}$ with $e_u := (0, \ldots, 0, \underbrace{1}_{u\text{-}th}, 0, \ldots, 0)$, and obtain that

$$\frac{1}{D_{4n}} \cdot \sum_{j \in \mathcal{J}_{\lambda_n} : |\mathcal{C}_j| > 1} \sum_{i \in \mathcal{C}_j} \lambda_n p_i^n \|\Delta a_{ij}^n\|^2 \to 0, \tag{42}$$

For $j \in \mathcal{J}_{\lambda_n}$ such that $|\mathcal{C}_j| = 1$, we combine the limits of $E_{\mathbf{0}_d,1}(j)/D_{4n}$, $E_{\mathbf{0}_d,2}(j)/D_{4n}$ and $E_{\alpha_1,0}(j)/D_{4n}$ for any $\alpha_1 \in \{e_1, e_2, \ldots, e_d\}$, then

$$\frac{1}{D_{4n}} \cdot \sum_{j \in \mathcal{J}_{\lambda_n} : |\mathcal{C}_j| = 1} \sum_{i \in \mathcal{C}_j} \lambda_n p_i^n \Big(\|\Delta a_{ij}^n\|_1 + |\Delta b_{ij}^n| + |\Delta \sigma_{ij}^n| \Big) \to 0.$$

Due to the topological equivalence between 1-norm and 2-norm, we receive

$$\frac{1}{D_{4n}} \cdot \sum_{j \in \mathcal{J}_{\lambda_n} : |\mathcal{C}_j| = 1} \sum_{i \in \mathcal{C}_j} \lambda_n p_i^n \Big(\|\Delta a_{ij}^n\| + |\Delta b_{ij}^n| + |\Delta \sigma_{ij}^n| \Big) \to 0. \tag{43}$$

Since $s(\lambda_n) \not\to 0$, it follows from equations (42) and (43) that

$$\frac{s(\lambda_n)}{D_{4n}} \cdot \sum_{j \in \mathcal{J}_{\lambda_n} : |\mathcal{C}_j| > 1} \sum_{i \in \mathcal{C}_j} \lambda_n p_i^n \|\Delta a_{ij}^n\|^2 + \frac{s(\lambda_n)}{D_{4n}} \cdot \sum_{j \in \mathcal{J}_{\lambda_n} : |\mathcal{C}_j| = 1} \sum_{i \in \mathcal{C}_j} \lambda_n p_i^n \Big(\|\Delta a_{ij}^n\|_1 + |\Delta b_{ij}^n| + |\Delta \sigma_{ij}^n| \Big) \to 0. \tag{44}$$

By taking the summation of the limits of $|E_{\mathbf{0}_d,0}(j)|/D_{4n}$ for $j \in \mathcal{J}_{\lambda_n}$, we get that

$$\frac{1}{D_{4n}} \cdot \sum_{j \in \mathcal{J}_{\lambda_n}} \left| \sum_{i \in \mathcal{C}_j} \lambda_n p_i^n - \bar{p}_j^*(\lambda_n) \right| \to 0.$$

From the above hypothesis, we take the summation of all the coefficients in the representation of C_n/D_{4n} and get that

$$\frac{1}{D_{4n}} \cdot \sum_{j \in I_{\lambda_n}} -\bar{p}_j^*(\lambda_n) \to 0.$$

Then, we have

$$\begin{aligned}
0 \leq \frac{s(\lambda_n) |\sum_{i \in \mathcal{C}_j} p_i^n - \bar{p}_j^*(\lambda_n)/s(\lambda_n)|}{D_{4n}} &= \frac{\sum_{i \in \mathcal{J}_{\lambda_n}} |s(\lambda_n) \sum_{i \in \mathcal{C}_j} p_i^n - \bar{p}_j^*(\lambda_n)|}{D_{4n}} \\
\leq \frac{\sum_{j \in \mathcal{J}_{\lambda_n}} |\lambda_n \sum_{i \in \mathcal{C}_j} p_i^n - \bar{p}_j^*(\lambda_n)|}{D_{4n}} &+ \left(\sum_{j \in \mathcal{J}_{\lambda_n}} \sum_{i \in \mathcal{C}_j} p_i^n \right) \frac{\sum_{j \in I_{\lambda_n}} -\bar{p}_j^*(\lambda_n)}{D_{4n}} \to 0,
\end{aligned}$$

which leads to

$$\frac{1}{D_{4n}} \cdot s(\lambda_n) \left| \sum_{i \in \mathcal{C}_j} p_i^n - \bar{p}_j^*(\lambda_n)/s(\lambda_n) \right| \to 0. \tag{45}$$

By plugging in the limits in equations (44) and (45) into the equation(41), we deduce that

$$\frac{1}{D_{4n}} \cdot s(\lambda_n) \sum_{j:|\mathcal{C}_j|>1} \sum_{i \in \mathcal{C}_j} p_i^n \left(|\Delta b_{ij}^n|^{\bar{r}(|\mathcal{C}_j|)} + |\Delta \sigma_{ij}^n|^{\bar{r}(|\mathcal{C}_j|)/2} \right) \to 1.$$

Therefore, we can find an index $j^* \in \mathcal{J}_{\lambda_n}$ such that $|\mathcal{C}_j| > 1$ satisfies

$$\frac{1}{D_{4n}} \cdot s(\lambda_n) \sum_{i \in \mathcal{C}_{j^*}} p_i^n \left(|(\Delta b_{ij^*}^n)^{(1)}|^{\bar{r}(|\mathcal{C}_{j^*}|)} + |\Delta \sigma_{ij^*}^n|^{\bar{r}(|\mathcal{C}_{j^*}|)/2} \right) \not\to 0.$$

WLOG, we assume that $j^* = 1$. From the hypothesis, as $E_{\mathbf{0}_d,\ell}(1)/D_{4n} \to 0$ as $n \to \infty$ for any $1 \leq \ell \leq \bar{r}(|\mathcal{C}_1|)$, we have

$$\frac{\sum_{i \in \mathcal{C}_1} p_i^n \sum_{\alpha_2 + 2\alpha_3 = \ell} \frac{(\Delta b_{i1}^n)^{\alpha_2} (\Delta \sigma_{i1}^n)^{\alpha_3}}{2^{\alpha_3} \alpha_2! \alpha_3!}}{s(\lambda_n) \sum_{i \in \mathcal{C}_1} p_i^n \left(|\Delta b_{i1}^n|^{\bar{r}(|\mathcal{C}_1|)} + |\Delta \sigma_{ij}^n|^{\bar{r}(|\mathcal{C}_1|)/2} \right)} \to 0,$$

for any $1 \leq \ell \leq \bar{r}(|\mathcal{C}_1|)$. Recall that $s(\lambda_n) \not\to 0$, then

$$\frac{\sum_{i \in \mathcal{C}_1} p_i^n \sum_{\alpha_2 + 2\alpha_3 = \ell} \frac{(\Delta b_{i1}^n)^{\alpha_2} (\Delta \sigma_{i1}^n)^{\alpha_3}}{2^{\alpha_3} \alpha_2! \alpha_3!}}{\sum_{i \in \mathcal{C}_1} p_i^n \left(|\Delta b_{i1}^n|^{\bar{r}(|\mathcal{C}_1|)} + |\Delta \sigma_{ij}^n|^{\bar{r}(|\mathcal{C}_1|)/2} \right)} \to 0. \tag{46}$$

Subsequently, we denote

$$\overline{M}_n = \max\{|\Delta b_{i1}^n|, |\Delta \sigma_{ij}^n|^{1/2} : i \in \mathcal{C}_1\}, \quad \bar{p}_n = \max_{i \in \mathcal{C}_1} p_i^n.$$

Since the sequence $p_i^n/\bar{p}_n$ is bounded, we can substitute it by its subsequence which admits a non-negative limit $s_i^2 = \lim_{n \to \infty} p_i^n/\bar{p}_n$. Furthermore, as $p_i^n \geq \xi > 0$ for all $i \in \mathcal{C}_1$, at least one among the limit s_i^2 is equal to 1. Similarly, let $(\Delta b_{i1}^n)/\overline{M}_n \to t_{1i}$ and $(\Delta \sigma_{i1}^n)/(2\overline{M}_n^2) \to t_{2i}$ as $n \to \infty$ for any $i \in \mathcal{C}_1$. Then, at least one among t_{1i} and t_{2i} for $i \in \mathcal{C}_1$ is equal to either 1 or -1.

Then, we divide both the numerator and the denominator of the ratio in equation (46) by $\bar{p}_n \overline{M}_n^{\ell}$, and obtain the following system of polynomial equations:

$$\sum_{i \in \mathcal{C}_1} \sum_{\alpha_2 + 2\alpha_3 = \ell} \frac{s_i^2 \, t_{1i}^{\alpha_2} \, t_{2i}^{\alpha_3}}{\alpha_2! \; \alpha_3!} = 0, \quad \forall \ell = 1, 2, \ldots, \bar{r}(|\mathcal{C}_1|).$$

It follows from the definition of $\bar{r}(|\mathcal{C}_j|)$ that this system of polynomial equations will not admit any non-trivial solutions $(s_i, t_{1i}, t_{2i})_{i \in \mathcal{C}_j}$, which is a contradiction to the fact that $s_i > 0$ for all $i \in \mathcal{C}_1$.

Consequently, at least one among the coefficients in the representations of $A_{n,1}/D_{4n}$, $A_{n,2}/D_{4n}$, B_n/D_{4n} and C_n/D_{4n} does not go to zero as $n \to \infty$. Let us denote by m_n the maximum of the absolute values of those aforementioned coefficients, i.e.

$$m_n = \max_{\substack{j \in [k_*], 0 \leq |\alpha_1| \leq \bar{r}(|\mathcal{C}_j|), \\ 0 \leq \ell \leq 2(\bar{r}(|\mathcal{C}_j|) - |\alpha_1|)}} \left\{ \frac{|E_{\alpha_1,\ell}^n(j)|}{D_{4n}} \right\}.$$

Additionally, we define

$$E_{\alpha_1,\ell}^n(j)/m_n \to \tau_{\alpha_1,\ell}(j)$$

as $n \to \infty$ for all $j \in [k_*]$, $0 \leq |\alpha_1| \leq \bar{r}(|\mathcal{C}_j|)$, $0 \leq \ell \leq 2(\bar{r}(|\mathcal{C}_j|) - |\alpha_1|)$. Here, at least one among $\tau_{\alpha_1,\ell}(j)$ is non-zero. By applying the Fatou's lemma, we get

$$0 = \lim_{n\to\infty} \frac{1}{m_n} \frac{2V(p_{\lambda_n,G_n}, p_{\lambda^*,G_*})}{D_{4n}} \geq \int \liminf_{n\to\infty} \frac{1}{m_n} \frac{|p_{\lambda_n,G_n}(X,Y) - p_{\lambda^*,G_*}(X,Y)|}{D_{4n}} \mathrm{d}(X,Y) \geq 0.$$

Note that

$$\frac{1}{m_n} \frac{p_{\lambda_n,G_n}(X,Y) - p_{\lambda^*,G_*}(X,Y)}{D_{4n}} \to \sum_{j,\alpha_1,\ell} \tau_{\alpha_1,\ell}(j) X^{\alpha_1} \cdot \frac{\partial^{|\alpha_1|+\ell} f}{\partial h_1^{|\alpha_1|+\ell}}(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) \bar{f}(X).$$

As a result, we get

$$\sum_{j,\alpha_1,\ell} \tau_{\alpha_1,\ell}(j) X^{\alpha_1} \cdot \frac{\partial^{|\alpha_1|+\ell} f}{\partial h_1^{|\alpha_1|+\ell}}(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) \bar{f}(X) = 0, \tag{47}$$

By employing similar arguments for showing the set $\mathcal{H}_2$ is linearly independent as in Appendix A.2, we can demonstrate that $\mathcal{H}_3$ defined in equation (40) is also a linearly independent set. Thus, equation (47) indicates that

$$\sum_{j,\alpha_1,\ell} \tau_{\alpha_1,\ell}(j) X^{\alpha_1} = 0,$$

for all $j \in [k_*]$ and $0 \leq |\alpha_1| \leq \bar{r}(|\mathcal{C}_j|)$ and $0 \leq \ell \leq 2(\bar{r}(|\mathcal{C}_j|) - |\alpha_1|)$. As the left hand side of the above equation is a polynomial of $X \in \mathcal{X}$, which is a bounded set of $\mathbb{R}^d$. Then, $\tau_{\alpha_1,\ell}(j) = 0$ for all $j \in [k_*]$, $0 \leq |\alpha_1| \leq \bar{r}(|\mathcal{C}_j|)$ and $0 \leq \ell \leq 2(\bar{r}(|\mathcal{C}_j|) - |\alpha_1|)$. This is a contradiction to the fact that at least one among $\tau_{\alpha_1,\ell}(j)$ is different from 0. Therefore, we reach the local inequality in equation (36), which means that there exists a positive constant ε_0 such that

$$\inf_{\substack{\lambda\in[0,1], G\in\mathcal{G}_{k,\xi}(\Theta):\\ D_4((\lambda,G),(\lambda^*,G_*))\leq\varepsilon_0}} V(p_{\lambda,G}, p_{\lambda^*,G_*})/D_4((\lambda,G),(\lambda^*,G_*)) > 0.$$

Global inequality. Thus, it is sufficient to demonstrate that

$$\inf_{\substack{\lambda\in[0,1], G\in\mathcal{G}_{k,\xi}(\Theta):\\ D_4((\lambda,G),(\lambda^*,G_*))>\varepsilon_0}} V(p_{\lambda,G}, p_{\lambda^*,G_*})/D_4((\lambda,G),(\lambda^*,G_*)) > 0. \tag{48}$$

Suppose that the above inequality does not hold, then there exist sequences $\lambda_n' \in [0,1]$ and $G_n' \in \mathcal{G}_{k,\xi}(\Theta)$ such that

$$\begin{cases} D_4((\lambda_n', G_n'), (\lambda^*, G_*)) > \varepsilon_0 \\ V(p_{\lambda_n',G_n'}, p_{\lambda^*,G_*})/D_4((\lambda_n', G_n'), (\lambda^*, G_*)) \to 0, \end{cases}$$

which implies that $V(p_{\lambda_n',G_n'}, p_{\lambda^*,G_*}) \to 0$ as $n \to \infty$. Note that the sets Θ and $[0,1]$ are bounded, we can find a subsequence of G_n' and a subsequence of λ_n' such that $G_n' \to G'$ and $\lambda_n' \to \lambda'$, where $G' \in \mathcal{G}_{k,\xi}(\Theta)$ and $\lambda' \in [0,1]$. By replacing G_n' and λ_n' with their subsequences, we get that $D_4((\lambda', G'), (\lambda^*, G_*)) > \varepsilon_0$. By the Fatou's lemma, we obtain that

$$\begin{aligned} 0 = \lim_{n\to\infty} 2V(p_{\lambda_n',G_n'}, p_{\lambda^*,G_*}) &\geq \int \liminf_{n\to\infty} \left|p_{\lambda_n',G_n'}(X,Y) - p_{\lambda^*,G_*}(X,Y)\right| d(X,Y) \\ &= \int \left|p_{\lambda',G'}(X,Y) - p_{\lambda^*,G_*}(X,Y)\right| d(X,Y) \geq 0, \end{aligned}$$

which indicates that $p_{\lambda',G'}(X,Y) = p_{\lambda^*,G_*}(X,Y)$ for almost surely (X,Y).

Case 1: $\lambda_n' \in \mathcal{T}$ for infinitely $n \in \mathbb{N}$. WLOG, we assume that $\lambda_n' \in \mathcal{T}$ for all $n \in \mathbb{N}$.

In this case, $D_4((\lambda', G'), (\lambda^*, G_*)) = D_3(G', \overline{G}_*(\lambda')) > \varepsilon_0$. It follows from equation (38) that

$$0 = p_{\lambda',G'}(X,Y) - p_{\lambda^*,G_*}(X,Y) = \lambda'[p_{G'}(X,Y) - p_{\overline{G}_*(\lambda')}(X,Y)],$$

Since $\liminf \lambda'_n$ is lower bounded by a positive constant, then $\lambda' > 0$. Combining this with the above result, we get that $p_G(X,Y) = p_{\overline{G}_*(\lambda')}(X,Y)$ for almost surely (X,Y). Due to the identifiability of the Gaussian mixture of experts Ho et al. (2022), we obtain that $G' = \overline{G}_*(\lambda')$. This means that $D_3(G', \overline{G}_*(\lambda')) = 0$, which is a contradiction to the fact that $D_3(G', \overline{G}_*(\lambda')) > \varepsilon_0 > 0$.

Case 2: $\lambda'_n \in \mathcal{T}^c$ for infinitely $n \in \mathbb{N}$. WLOG, we assume that $\lambda'_n \in \mathcal{T}^c$ for all $n \in \mathbb{N}$.

Case 2.1: $I_{\lambda'_n}$ is not ratio-independent

In this case, $D_4((\lambda', G'), (\lambda^*, G_*)) = \sum_{j \in I_{\lambda'}} -\bar{p}^*_j(\lambda') > \varepsilon_0$. It follows from equation (37) that

$$\begin{aligned}
0 = p_{\lambda', G'}(X,Y) - p_{\lambda^*, G_*}(X,Y) &= \sum_{j \in I_{\lambda'}} -\bar{p}^*_j(\lambda') f(Y|(a^*_j)^\top X + b^*_j, \sigma^*_j)\bar{f}(X) \\
&\quad + \Big[\sum_{i=1}^{k'} \lambda' p'_i f(Y|(a'_i)^\top X + b'_i, \sigma'_i)\bar{f}(X) - \sum_{j \in \mathcal{J}_{\lambda'}} \bar{p}^*_j(\lambda') f(Y|(a^*_j)^\top X + b^*_j, \sigma^*_j)\bar{f}(X)\Big] \\
&= \sum_{j \in I_{\lambda'}} -\bar{p}^*_j(\lambda') f(Y|(a^*_j)^\top X + b^*_j, \sigma^*_j)\bar{f}(X) + [p_{G'}(X,Y) - p_{\overline{G}_*(\lambda')}(X,Y)] \\
&= \sum_{j \in I_{\lambda'}} -\bar{p}^*_j(\lambda') f(Y|(a^*_j)^\top X + b^*_j, \sigma^*_j)\bar{f}(X).
\end{aligned}$$

Recall that $-\bar{p}^*_j(\lambda') > 0$ for all $j \in I_{\lambda'}$. Thus, $\bar{p}^*_j(\lambda') = 0$ as $n \to \infty$ for all $j \in I_{\lambda'}$. This leads to the fact that $p^*_i/p^0_i = p^*_j/p^0_j = (\lambda^* - \lambda')/\lambda^*$ for all $i, j \in I_{\lambda'}$, which is a contradiction to the fact that $I_{\lambda'}$ is not ratio-independent, which follows from the ratio-independece of $I_{\lambda'_n}$.

Case 2.2: $I_{\lambda'_n}$ is ratio-independent.

In this case, $D_4((\lambda', G'), (\lambda^*, G_*)) = D_3(G', \widetilde{G}_*(\lambda'))$. The result $p_{\lambda', G'}(X,Y) = p_{\lambda^*, G_*}(X,Y)$ for almost surely (X,Y) indicates that $G' = \widetilde{G}_*(\lambda')$. Then, we have $D_3(G', \widetilde{G}_*(\lambda'_n)) = 0$, which contradicts to the fact that $D_3(G', \widetilde{G}_*(\lambda'_n)) > \varepsilon_0 > 0$.

Hence, the proof is completed.

D PARAMETER ESTIMATION WITH VANISHING MIXING PROPORTION

In this appendix, we resume the discussion about parameter estimation rates under the deviated Gaussian mixture of experts when the mixing proportion vanishes, that is, $\lambda^* = 0$. For that purpose, we consider the distinguishable and non-distinguishable settings in Appendix D.1 and Appendix D.2, respectively.

D.1 Distinguishable Settings

First of all, we explore the convergence behavior of parameter estimation under the distinguishable settings.

Theorem 4. *Assume that the distinguishability condition in Definition 1 holds and $\lambda^* = 0$. Then, the Total Variation lower bound $V(p_{\lambda,G}, p_{\lambda^*,G_*})$ & λ holds for any $(\lambda, G) \in [0,1] \times \mathcal{G}_k(\Theta)$. This bound together with Proposition 2 suggest that we can find a positive constant C_5 that depends only on g_0, λ^*, Θ such that*

$$\mathbb{P}(\hat{\lambda}_n > C_5\sqrt{\log(n)/n}) \ . \ n^{-c_5},$$

where c_5 is a positive constant depending only on Θ.

When $\lambda^* = 0$, the mixture part p_{G_*} is no longer involved in the formulation of the true conditional density function $p_{\lambda^*,G_*}(Y|X)$. Moreover, since p_{G_*} is distinguishable from the known function g_0, then we are not able to access the convergence behavior of the MLE $\widehat{G}_n$. Nevertheless, Theorem 4 indicates that the mixing proportion estimation $\hat{\lambda}_n$ converges to $\lambda^* = 0$ at a parametric rate of order $\mathcal{O}(n^{-1/2})$.

Proof of Theorem 4. From the result of Proposition 2, it is sufficient to show that $V(p_{\hat{\lambda}_n, \widehat{G}_n}, p_{\lambda^*,G_*})$ & $\hat{\lambda}_n$. When $\hat{\lambda}_n = 0$, this problem becomes trivial. Therefore, we will consider only the case when $\hat{\lambda}_n > 0$, in which the

problem turns into proving that

$$\inf_{\lambda\in(0,1],G\in\mathcal{G}_{k,\xi}(\Theta)} \frac{V(p_{\lambda,G},p_{\lambda^*,G_*})}{\lambda} > 0.$$

Assume that the above inequality does not hold, which implies that there exist sequences $\lambda_n \in (0,1]$ and $G_n = \sum_{i=1}^{k_n} p_i^n \delta_{(a_i^n,b_i^n,\sigma_i^n)} \in \mathcal{G}_{k,\xi}(\Theta)$ such that $V(p_{\lambda_n,G_n},p_{\lambda^*,G_*})/\lambda_n \to 0$ as $n\to\infty$. Since Θ is a compact set, we can find a subsequence of G_n such that $G_n \to \widetilde{G}$, where $\widetilde{G} := \sum_{i=1}^{\widetilde{k}} \widetilde{p}_i \delta_{(\widetilde{a}_i,\widetilde{b}_i,\widetilde{\sigma}_i)} \in \mathcal{G}_{k,\xi}(\Theta)$. By replacing G_n with this subsequence and applying the Fatou's lemma with a note that $\lambda^* = 0$, we get

$$\lim_{n\to\infty} \frac{2V(p_{\lambda_n,G_n},p_{\lambda^*,G_*})}{\lambda_n} \geq \int \liminf_{n\to\infty} \left|\sum_{i=1}^{k_n} p_i^n f(Y|(a_i^n)^\top X + b_i^n,\sigma_i^n) - g_0(Y|X)\right| \bar{f}(X) d(X,Y).$$

It follows from the hypothesis $V(p_{\lambda_n,G_n},p_{\lambda^*,G_*})/\lambda_n \to 0$ that $\sum_{i=1}^{\widetilde{k}} \widetilde{p}_i f(Y|(\widetilde{a}_i)^\top X + \widetilde{b}_i,\widetilde{\sigma}_i) - g_0(Y|X) = 0$, for almost surely (X,Y). This contradicts the assumption that $p_{\widetilde{G}}$ is distinguishable from g_0. Hence, we reach the conclusion of this part. □

D.2 Non-distinguishable Settings

We now draw our attention to parameter estimation rates under the non-distinguishable settings when the mixing proportion vanishes, namely when the function g_0 takes the following form:

$$g_0(Y|X) = p_{G_0}(Y|X) := \sum_{j=1}^{k_0} p_j^0 f(Y|(a_j^0)^\top X + b_j^0,\sigma_j^0), \tag{49}$$

where $k_0 \in [k_*]$.

Since $\lambda^* = 0$, the mixture part p_{G_*} is not involved in the formulation of the true conditional density $p_{\lambda^*,G_*}(Y|X)$. As a result, we do not have any interaction between two functions $g_0(Y|X)$ and $p_{G_*}(Y|X)$. Therefore, it is unnecessary to divide the non-distinguishable setting into partial overlap regime and full overlap regime. Instead, we establish the parameter estimations under the general non-distinguishable settings in the following theorem:

Theorem 5. *Suppose that the function g_0 takes the form in equation (49) and $\lambda^* = 0$. Then, the Total Variation lower bound $V(p_{\lambda,G},p_{\lambda^*,G_*})$ & $\lambda D_3(G,G_0)$ holds for any $(\lambda,G) \in [0,1]\times\mathcal{G}_k(\Theta)$. This bound together with Proposition 2 indicates that there exists a positive constants C_6 depending on g_0,λ^*,Θ such that*

$$\mathbb{P}(\hat{\lambda}_n D_3(\widehat{G}_n,G_0) > C_6\sqrt{\log(n)/n}) \ . \ n^{-c_6},$$

where c_6 is a constant that depends only on Θ.

Different from the results of all previous theorems, the MLE $\widehat{G}_n$ converges to the mixing measure G_0 rather than G_* under the loss function D_3 due to the disappearance of the mixture part $p_{G_*}(Y|X)$ in the conditional density $p_{\lambda^*,G_*}(Y|X)$. Moreover, the rate of that convergence depends on the vanishing rate of $\hat{\lambda}_n$, therefore, it is no better than the parametric rate of order $\mathcal{O}(n^{-1/2})$.

Proof of Theorem 5. Note that the problem is trivial when $\hat{\lambda}_n = 0$, therefore, we consider only the case when $\hat{\lambda}_n > 0$. From Proposition 2, it is sufficient to show that

$$\inf_{\lambda\in(0,1],G\in\mathcal{G}_{k,\xi}(\Theta)} \frac{V(p_{\lambda,G},p_{\lambda^*,G_*})}{\lambda D_3(G,G_0)} > 0. \tag{50}$$

Since $\lambda^* = 0$, we get that

$$\begin{aligned} p_{\lambda,G}(X,Y) - p_{\lambda^*,G_*}(X,Y) &= (1-\lambda)g_0(Y|X)\bar{f}(X) + \lambda p_G(X,Y) - g_0(Y|X)\bar{f}(X) \\ &= \lambda\,[p_G(X,Y) - g_0(Y|X)\bar{f}(X)] \\ &= \lambda\,[p_G(X,Y) - p_{G_0}(X,Y)]. \end{aligned}$$

As a result, equation (50) becomes $\inf_{G\in\mathcal{G}_{k,\xi}(\Theta)} V(p_G, p_{G_0})/D_3(G, G_0) > 0$.

Local inequality: We first prove that

$$\lim_{\varepsilon\to 0}\inf_{G\in\mathcal{G}_{k,\xi}(\Theta):D_3(G,G_0)\leq\varepsilon}\frac{V(p_G, p_{G_0})}{D_3(G, G_0)} > 0.$$

Assume by contrary that the above claim is not true. Then, there exists a sequence $G_n = \sum_{i=1}^{k_n} p_i^n \delta_{(a_i^n, b_i^n, \sigma_i^n)} \in \mathcal{G}_{k,\xi}(\Theta)$ such that as $n \to \infty$, we have

$$\begin{cases} D_3(G_n, G_0) \to 0, \\ V(p_{G_n}, p_{G_0})/D_3(G_n, G_0) \to 0. \end{cases}$$

By employing arguments (with adapted notations) used in Case 2.2 in Appendix A.2 for showing contradiction to the fact that $V(p_{G_n}, p_{\overline{G}_*(\lambda_n)}) \to 0$ as $n \to \infty$, we also get a contradiction here. Consequently, there exists a positive constant ε_0 such that

$$\inf_{G\in\mathcal{G}_{k,\xi}(\Theta):D_3(G,G_0)\leq\varepsilon_0}\frac{V(p_G, p_{G_0})}{D_3(G, G_0)} > 0.$$

Global inequality: From the above result, we only need to show that

$$\inf_{G\in\mathcal{G}_{k,\xi}(\Theta):D_3(G,G_0)>\varepsilon_0}\frac{V(p_G, p_{G_0})}{D_3(G, G_0)} > 0.$$

Assume that the above inequality does not hold. Then, there exists a sequence $G'_n \in \mathcal{G}_{k,\xi}(\Theta)$ satisfying $V(p_{G'_n}, p_{G_0})/D_3(G'_n, G_0) \to 0$ as $n \to \infty$, whereas $D_3(G'_n, G_0) > \varepsilon_0$ for all $n \in \mathbb{N}$. Therefore, $V(p_{G'_n}, p_{G_0}) \to 0$ as $n \to \infty$. Note that Θ is a compact set, then we can find a subsequence of G'_n such that $G'_n \to G'$ for some $G' \in \mathcal{G}_{k,\xi}(\Theta)$. By replacing the sequence G'_n by that subsequence, we obtain that $D_3(G', G_0) > \varepsilon_0$ as a result of $D_3(G'_n, G_0) > \varepsilon_0$ for all $n \in \mathbb{N}$. By Fatou's lemma, we get

$$\begin{aligned} 0 = \lim_{n\to\infty} V(p_{G'_n}, p_{G_0}) &\geq \frac{1}{2}\int \liminf_{n\to\infty} |p_{G'_n}(X, Y) - p_{G_0}(X, Y)| d(X, Y) \\ &= \frac{1}{2}\int |p_{G'}(X, Y) - p_{G_0}(X, Y)| d(X, Y) \geq 0, \end{aligned}$$

which implies that $p_{G'}(X, Y) = p_{G_0}(X, Y)$ for almost surely $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. Since the Gaussian mixture of experts is identifiable (cf. Proposition 3 in Ho et al. (2022)), the previous equation indicates that $G' \equiv G_0$. This contradicts to the fact that $D_3(G', G_0) \geq \varepsilon > 0$.

Hence, the proof is completed. □