Towards Convergence Rates for Parameter Estimation in Gaussian-gated Mixture of Experts

Huy Nguyen[◇,⋆] **TrungTin Nguyen**[○,†,⋆] **Khai Nguyen**[◇] **Nhat Ho**[◇]

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

School of Mathematics and Physics, The University of Queensland[○]

Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France[†]

Abstract

Originally introduced as a neural network for ensemble learning, mixture of experts (MoE) has recently become a fundamental building block of highly successful modern deep neural networks for heterogeneous data analysis in several applications of machine learning and statistics. Despite its popularity in practice, a satisfactory level of theoretical understanding of the MoE model is far from complete. To shed new light on this problem, we provide a convergence analysis for maximum likelihood estimation (MLE) in the Gaussian-gated MoE model. The main challenge of that analysis comes from the inclusion of covariates in the Gaussian gating functions and expert networks, which leads to their intrinsic interaction via some partial differential equations with respect to their parameters. We tackle these issues by designing novel Voronoi loss functions among parameters to accurately capture the heterogeneity of parameter estimation rates. Our findings reveal that the MLE has distinct behaviors under two complement settings of location parameters of the Gaussian gating functions, namely when all these parameters are non-zero versus when at least one among them vanishes. Notably, these behaviors can be characterized by the solvability of two different systems of polynomial equations. Finally, we conduct a simulation study to empirically verify our theoretical results.

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

1 INTRODUCTION

Mixture of experts (MoE) (Jacobs et al., 1991; Jordan and Jacobs, 1994) is a popular statistical machine learning model where experts are either regression functions or classifiers, while the input-dependent weights (also called gating functions) softly partition the input space into different regions and define which regions each expert is responsible for (see (Yuksel et al., 2012; Masoudnia and Ebrahimpour, 2014; Fedus et al., 2022a) for further details). In regression analysis with heterogeneous data, softmax-gated MoE (Jacobs et al., 1991; Jordan and Jacobs, 1994) and *Gaussian-gated MoE* (GMoE) (Xu et al., 1995) models are the most popular choices. One of the main drawbacks of the softmax-gated MoE models is the difficulty of applying an expectation-maximization (EM) algorithm (Dempster et al., 1977), which requires an internal iterative numerical optimization procedure, e.g., Newton-Raphson algorithm, to update the softmax parameters in the maximization step. On the other hand, parameters of the GMoE models can be updated analytically, which helps reduce the computational complexity of the estimation routine. For those reasons, GMoE has become a fundamental component of modern deep neural networks in various fields, including speech recognition (Fritsch et al., 1996; You et al., 2022), computer vision (Lathuilière et al., 2017; Puigcerver et al., 2021), natural language processing (Shazeer et al., 2017; Fedus et al., 2022b; Mustafa et al., 2022; Do et al., 2023; Pham et al., 2024), medical images (Han et al., 2024), robot dynamics (Sato and Ishii, 2000; Moody and Darken, 1989), remote sensing (Deleforge et al., 2015; Kugler et al., 2022; Forbes et al., 2022a,b), and econometrics (Norets and Pelenis, 2021; Norets and Pati, 2017; Diani et al., 2022). However, there is a paucity of work aiming at theoretically understanding the density estimation and parameter estimation in the GMoE models, which has remained poorly understood in the literature to

the best of our knowledge.

Related literature. In the GMoE setting, early classical research focused on identifiability issues (Jiang and Tanner, 1999c) and parameter estimation in the exact-fitted setting, assuming the true number of components k_0 is known (Jiang and Tanner, 1999a). For most applications, it is a too strong presumption as the true number of components is seldom known. To deal with this problem, there are three common practical approaches. The first approach is based on model selection, most importantly the Bayesian information criterion from asymptotic theory (Forbes et al., 2022b; Chamroukhi and Huynh, 2019; Khalili, 2010) and the slope heuristic (Baudry et al., 2012; Birgé and Massart, 2007) in a non-asymptotic framework (Nguyen et al., 2022b, 2023d,e, 2022a). In particular, the bias term can be substantially reduced with a sufficiently large model collection w.r.t. the number of mixture components k by well-studied universal approximations theorems (Nguyen et al., 2021; Mendes and Jiang, 2012; Jiang and Tanner, 1999b). However, since we have to search for the optimal k over all possible values, this approach is computationally expensive. The second approach is to design a tractable Bayesian nonparametric GMoE model. For example, Nguyen et al. (2023c) avoided any commitment to an arbitrary k with posterior consistency guarantee thanks to the merge-truncate-merge post-processing in Guha et al. (2021). However, this approach still depends on a tuning parameter, which prevents the direct application of this approach to real data sets. The last approach is to use prior knowledge to over-specify the true model, i.e. specifying more mixture components than necessary, where most existing work is limited to its particular case, including mixture models (Ho and Nguyen, 2016a,b, 2019; Guha et al., 2021; Manole and Ho, 2022) and mixture of experts (Ho et al., 2022; Nguyen et al., 2023b,a, 2024b,a). It is worth noting that the convergence behavior of parameter estimations in the GMoE model has remained an open question, which we aim to answer in this paper. Before going into further details, we first formally introduce an affine instance of the GMoE model. This is a simplified but standard setting where we use linear functions for Gaussian mean experts.

GMoE setting. GMoE models are used to capture the non-linear and heterogeneous relationship between the response $Y \in \mathcal{Y} \subset \mathbb{R}$ and the set of covariates $X \in \mathcal{X} \subset \mathbb{R}^d$, $d \in \mathbb{N}$. In the affine GMoE model, the response Y is approximated by a k_0 local affine:

$$Y = \sum_{j=1}^{k_0} \mathbb{I}\left(Z = j\right)\left[(a_j^0)^\top X + b_j^0 + e_j^0\right]. \tag{1}$$

Here $\mathbb{I}$ is an indicator function and Z is a latent variable that captures a cluster relationship, such that $Z = j$ if Y comes from cluster $j \in [k_0] := \{1, 2, \ldots, k_0\}$. Vectors $a_j^0 \in \mathbb{R}^d$ and scalars $b_j^0 \in \mathbb{R}$ define cluster-specific affine transformations. In addition, e_j^0 are error terms that capture both the reconstruction error (due to the local affine approximations) and the observation noise in $\mathbb{R}$. Let $\mathcal{F}_d := \left\{f(\cdot|\psi, \Sigma) : \psi \in \mathbb{R}^d, \Sigma \in \mathcal{S}_d^+\right\}$ be the family of d-dimensional Gaussian density functions with mean ψ and positive-definite covariance matrix Σ, where $\mathcal{S}_d^+$ indicates the set of all symmetric positive-definite matrices on $\mathbb{R}^{d\times d}$. Following the usual assumption that e_j^0 is a zero-mean Gaussian variable with variance $\nu_j^0 \in \mathbb{R}_+$, it follows that

$$p\left(Y|X, Z = j\right) = f_{\mathcal{D}}\left(Y|(a_j^0)^\top X + b_j^0, \nu_j^0\right), \quad f_{\mathcal{D}} \in \mathcal{F}_1.$$

To enforce the affine transformations to be local, X is defined as a mixture of k_0 Gaussian components:

$$p\left(X|Z = j\right) = f_{\mathcal{L}}\left(X|c_j^0, \Gamma_j^0\right), \; p\left(Z = j\right) = \pi_j^0, \tag{2}$$

where $f_{\mathcal{L}} \in \mathcal{F}_d$. Here, we refer to $f_{\mathcal{D}}$ and $f_{\mathcal{L}}$ as the data density and the local density, respectively. Additionally, $\pi_j^0 > 0$ are called mixing proportions (or weights), satisfying $\sum_{j=1}^{k_0} \pi_j^0 = 1$. Via the law of total probability, we obtain the GMoE model of order k_0 whose joint density function $p_{G_0}(X, Y)$ is given by:

$$\sum_{j=1}^{k_0} \pi_j^0 f_{\mathcal{L}}(X|c_j^0, \Gamma_j^0) \cdot f_{\mathcal{D}}(Y|(a_j^0)^\top X + b_j^0, \nu_j^0). \tag{3}$$

Here, $G_0 := \sum_{j=1}^{k_0} \pi_j^0 \delta_{(c_j^0, \Gamma_j^0, a_j^0, b_j^0, \nu_j^0)}$ denotes a true but unknown probability mixing measure, where δ is the Dirac measure and for $j \in [k_0]$, $(c_j^0, \Gamma_j^0, a_j^0, b_j^0, \nu_j^0) \in \Theta \subset \mathbb{R}^d \times \mathcal{S}_d^+ \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+$ are called components of G_0. We assume that $\{(X_i, Y_i)\}_{i\in[n]}$ are i.i.d. samples of random variable (X, Y), coming from the GMoE model of order k_0. To facilitate our theoretical guarantee, we assume that Θ is compact and $\mathcal{X}$ is bounded.

Maximum likelihood estimation. We propose a general theoretical framework for analyzing the statistical performance of *maximum likelihood estimation* (MLE) for parameters under the setting of the GMoE model. Since the true order k_0 is generally unknown in practice, it is necessary to over-specify the number of components of mixing measures to at most k, where $k > k_0$. In particular, we consider

$$\widehat{G}_n \in \underset{G\in\mathcal{O}_k(\Theta)}{\arg\max} \sum_{i=1}^{n} \log(p_G(X_i, Y_i)), \tag{4}$$

where $\mathcal{O}_k(\Theta) := \{G = \sum_{i=1}^{k'} \pi_i \delta_{(c_i, \Gamma_i, a_i, b_i, \nu_i)} : 1 \leq k' \leq k, \; \sum_{i=1}^{k} \pi_i = 1, \; (c_i, \Gamma_i, a_i, b_i, \nu_i) \in \Theta\}$ denotes the set of all mixing measures with at most k components.

Theoretical challenges. For the purpose of deriving parameter estimation rates in the GMoE model,

we first use the Taylor expansion to decompose the term $p_{\widehat{G}_n}(X,Y) - p_{G_0}(X,Y)$ into a linear combination of elements which belong to a linearly independent set and associate with coefficients involving the discrepancies between parameter estimations and true parameters. By doing so, when the density estimation $p_{\widehat{G}_n}$ converges to the true density p_{G_0}, those parameter discrepancies also go to zero and we then obtain our desired parameter estimation rates. Nevertheless, the density decomposition is challenging due to a number of linearly dependent derivative terms in the Taylor expansion. In particular, we find out two interactions among the parameters of either function $f_{\mathcal{D}}$ or $f_{\mathcal{L}}$ via the following partial differential equations (PDEs):

$$\frac{\partial^2 f_{\mathcal{D}}}{\partial b^2} = 2\frac{\partial f_{\mathcal{D}}}{\partial \nu}, \quad \frac{\partial^2 f_{\mathcal{L}}}{\partial c\, \partial c^\top} = 2\frac{\partial f_{\mathcal{L}}}{\partial \Gamma}. \tag{5}$$

We refer to those interactions as *interior interactions* since each of them involves either parameters b, ν of function $f_{\mathcal{D}}$ or parameters c, Γ of function $f_{\mathcal{L}}$. Furthermore, we also figure out an interaction between the parameters of functions $f_{\mathcal{D}}$ and $f_{\mathcal{L}}$. More specifically, let us denote $F(X,Y|\theta) := f_{\mathcal{L}}(X|c,\Gamma) f_{\mathcal{D}}(Y|a^\top X + b, \nu)$ where $\theta := (c, \Gamma, a, b, \nu)$. Then, by taking the derivatives of F with respect to its parameters as follows:

$$\frac{\partial^2 F}{\partial c\, \partial b}(X,Y|\theta_j^0) = \Gamma^{-1}(X - c_j^0) \cdot f_{\mathcal{L}} \cdot \frac{\partial f_{\mathcal{D}}}{\partial b};$$
$$\frac{\partial F}{\partial a}(X,Y|\theta_j^0) = X \cdot f_{\mathcal{L}} \cdot \frac{\partial f_{\mathcal{D}}}{\partial b},$$

it can be seen that the following PDE holds true when the location parameter of $f_{\mathcal{L}}$ vanishes, i.e. $c_j^0 = 0$:

$$\frac{\partial^2 F}{\partial c\, \partial b}(X,Y|\theta_j^0) = \Gamma^{-1} \cdot \frac{\partial F}{\partial a}(X,Y|\theta_j^0). \tag{6}$$

We refer to the interaction among parameters c, b, a in equation (6) as the *exterior interaction*. Back to the density decomposition, it is necessary to aggregate linearly dependent derivative terms in equations (5) and (6) by taking the summation of their associated coefficients. As a result, we achieve our desired linear combination of linearly independent terms. However, the structure of associated coefficients in that combination becomes complex owing to the previous aggregation. Thus, when those coefficients converge to zero, we have to cope with two complex systems of polynomial equations given in equations (9) and (12).

Overall contributions. In this paper, we characterize the convergence behavior of maximum likelihood estimation in the GMoE model. Firstly, we demonstrate that the density estimation $p_{\widehat{G}_n}$ converges to the true density p_{G_0} under the Total Variation distance V at the parametric rate $V(p_{\widehat{G}_n}, p_{G_0}) = \mathcal{O}(n^{-1/2})$. Regarding the parameter estimation problem, given the above challenge discussion, we consider two complement settings of the location parameters $c_1^0, c_2^0, \ldots, c_{k_0}^0$ based on the validity of the PDE in equation (6) as follows (see also Table 1):

1. Type I setting: *all the values of $c_1^0, c_2^0, \ldots, c_{k_0}^0$ are different from zero.* Since the PDE (6) does not hold under this setting, we have to deal with only the interior interactions in equation (5). Thus, we propose a novel Voronoi loss function $\overline{D}(G, G_0)$ defined in equation (10) to capture those interactions, and then establish the Total Variation lower bound $\overline{D}(\widehat{G}_n, G_0) \lesssim V(p_{\widehat{G}_n}, p_{G_0}) = \mathcal{O}(n^{-1/2})$. This result together with the formulation of $\overline{D}(\widehat{G}_n, G_0)$ indicate that exact-fitted parameters $c_j^0, \Gamma_j^0, a_j^0, b_j^0, \nu_j^0$, which are approximated by exactly one component, share the same estimation rate of order $\mathcal{O}(n^{-1/2})$. By contrast, the rates for estimating over-fitted parameters $c_j^0, \Gamma_j^0, b_j^0, \nu_j^0$, which are fitted by at least two components, depend on the solvability of the system of polynomial equations (9) and become no faster than $\mathcal{O}(n^{-1/4})$. These slow rates are due to the interior interactions among those parameters in equation (5). As over-fitted parameters a_j^0 are not involved in those interactions, their estimation rates keep unchanged of order $\mathcal{O}(n^{-1/4})$.

2. Type II setting: *at least one among the values of $c_1^0, c_2^0, \ldots, c_{k_0}^0$ is equal to zero.* Without loss of generality, we assume that $c_1^0, c_2^0, \ldots, c_{\widetilde{k}}^0$ equal zero, where $1 \leq \widetilde{k} \leq k_0$, while other c_j's are non-zero. Since the PDE (6) holds true under this setting, we have to confront both interior and exterior interactions among parameters. For that purpose, we construct another novel Voronoi loss function $\widetilde{D}(G, G_0)$ in equation (13) to handle those interactions, and then derive the Total Variation lower bound $\widetilde{D}(\widehat{G}_n, G_0) \lesssim V(p_{\widehat{G}_n}, p_{G_0}) = \mathcal{O}(n^{-1/2})$. Due to the occurrence of both interior and exterior interactions, the rates for estimating over-fitted parameters $c_j^0, \Gamma_j^0, a_j^0, b_j^0, \nu_j^0$ are now determined by the solvability of both systems of polynomial equations (9) and (12). Meanwhile, the estimation rates for their exact-fitted counterparts remain the same of order $\mathcal{O}(n^{-1/2})$.

Practical implication. In practice, the parameters specific to each mixing component may carry useful information about the heterogeneity of the underlying (latent) subpopulations. Since in reality there is a tendency to "over-fit" the mixture generously by adding many more mixing components, our theory warns against this because, as we have shown, the convergence rate via standard methods such as MLE for subpopulation-specific parameters deteriorates rapidly with the number of redundant components. Hopefully, the theoretical results will suggest practical ways to identify benign scenarios and impose helpful constraints when GMoE models have favourable convergence rates,

Setting	Exact-fitted $c_j^0, \Gamma_j^0, a_j^0, b_j^0, \nu_j^0$	Over-fitted a_j^0		Over-fitted c_j^0, b_j^0		Over-fitted Γ_j^0, ν_j^0	
		$j \in [\widetilde{k}]$	$j \in [k_0] \setminus [\widetilde{k}]$	$j \in [\widetilde{k}]$	$j \in [k_0] \setminus [\widetilde{k}]$	$j \in [\widetilde{k}]$	$j \in [k_0] \setminus [\widetilde{k}]$
Type I	$\mathcal{O}(n^{-1/2})$	$\mathcal{O}(n^{-1/4})$		$\mathcal{O}(n^{-1/2\bar{r}_j})$		$\mathcal{O}(n^{-1/\bar{r}_j})$	
Type II	$\mathcal{O}(n^{-1/2})$	$\mathcal{O}(n^{-1/\widetilde{r}_j})$	$\mathcal{O}(n^{-1/4})$	$\mathcal{O}(n^{-1/2\widetilde{r}_j})$	$\mathcal{O}(n^{-1/2\bar{r}_j})$	$\mathcal{O}(n^{-1/\widetilde{r}_j})$	$\mathcal{O}(n^{-1/\bar{r}_j})$

Table 1: Summary of parameter estimation rates in the GMoE model under the Type I and Type II settings. Recall that the cardinality of Voronoi cells $\mathcal{A}_j$ (see Section 2) generated by true components $(c_j^0, \Gamma_j^0, a_j^0, b_j^0, \nu_j^0)$ indicates the number of components fitting them. When $|\mathcal{A}_j| = 1$, we call them exact-fitted parameters, but when $|\mathcal{A}_j| > 1$, they are referred to as over-fitted parameters. Additionally, the notations $\bar{r}_j := \bar{r}(|\mathcal{A}_j|)$ and $\widetilde{r}_j := \widetilde{r}(|\mathcal{A}_j|)$ stand for the solvability of two polynomial equation systems (9) and (12), respectively. For example, if $|\mathcal{A}_j| = 2$, then we have $\bar{r}_j = \widetilde{r}_j = 4$. Meanwhile, we get $\bar{r}_j = \widetilde{r}_j = 6$ if $|\mathcal{A}_j| = 3$.

and detect pathological scenarios that practitioners would do well to avoid. In particular, practitioners can consistently estimate the true number of components based on our important threshold on the convergence rates of the MLE using the merge-truncate-merge procedure (Guha et al., 2021) or Group-Sort-Fuse (Manole and Khalili, 2021).

Paper organization. The rest of this paper proceeds as follows. In Section 2, we begin with providing some background on the identifiability of the GMoE model and the rate for estimating the joint density function under that model. Next, in Section 3, we establish the convergence rates of parameter estimation under both Type I and Type II settings, which are then empirically verified by simulation studies in Section 4. Finally, we conclude the paper in Section 5 and defer proofs of all theoretical results to the supplementary material.

Notation. Throughout the paper, $\{1, 2, \ldots, n\}$ is abbreviated as $[n]$ for any $n \in \mathbb{N}$. Given any two sequences of positive real numbers $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n = \mathcal{O}(b_n)$ or $a_n \lesssim b_n$ to indicate that there exists a constant $C > 0$ such that $a_n \leq C b_n$ for all $n \geq 1$. Next, for any vector $v \in \mathbb{R}^d$, we denote $|v| := v_1 + v_2 + \ldots + v_d$, whereas $\|v\|_p$ stands for its p-norm with a note that $\|v\|$ implicitly indicates the 2-norm unless stating otherwise. By abuse of notation, we also denote by $\|A\|$ the Frobenius norm of any matrix $A \in \mathbb{R}^{d \times d}$. Additionally, the notation $|S|$ represents for the cardinality of any set S. Finally, given two probability density functions p, q with respect to the Lebesgue measure μ, we define $V(p, q) := \frac{1}{2} \int |p - q| d\mu$ as their Total Variation distance, while $h^2(p, q) := \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2 d\mu$ denotes the squared Hellinger distance between them.

2 PRELIMINARIES

In this section, we first verify the identifiability of the GMoE model, and then establish the density estimation rate under that model. Lastly, we introduce a notion of a Voronoi cells, which will be used to build Voronoi loss functions in Section 3.

Firstly, we demonstrate in the following proposition that the GMoE model is identifiable:

Proposition 1 (Identifiability of the GMoE model)**.** *Let G and G' be two mixing measures in $\mathcal{O}_k(\Theta)$. If the equation $p_G(X, Y) = p_{G'}(X, Y)$ holds true for almost surely $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, then we obtain that $G \equiv G'$.*

The proof of Proposition 1 is deferred to Appendix C.1. Given the above result, we know that two mixing measures G and G' are equivalent if and only if they share the same joint density function.

Next, we characterize the convergence rate of the joint density estimation $p_{\widehat{G}_n}$ to its true counterpart p_{G_0} under the Total Variation distance.

Proposition 2 (Joint Density Estimation Rate)**.** *With the MLE $\widehat{G}_n$ defined in equation (4), the following bound indicates that the density estimation $p_{\widehat{G}_n}$ converges to the true density p_{G_0} under the Total Variation distance at the parametric rate of order $\mathcal{O}(n^{-1/2})$ (up to a logarithmic term):*

$$\mathbb{P}(V(p_{\widehat{G}_n}, p_{G_0}) > C_1 \sqrt{\log(n)/n}) \lesssim n^{-C_2},$$

where C_1 and C_2 are universal constants.

The proof of Proposition 2 can be found in Appendix C.2. This result is a key ingredient to study the parameter estimation problem in the GMoE model in subsequent sections. In particular, if we are able to establish the lower bound of the Total Variation distance in terms of some loss function D between two mixing measures, i.e., $V(p_G, p_{G_0}) \gtrsim D(G, G_0)$ for any mixing measure $G \in \mathcal{O}_k(\Theta)$, then the MLE $\widehat{G}_n$ also converges to the true mixing measure G_0 at the parametric rate of $\mathcal{O}(n^{-1/2})$. Based on this result, we then achieve the parameter estimation rates through the formulation of the loss function $D(\widehat{G}_n, G_0)$. For that purpose, let us

introduce a notion of Voronoi cells which are essential to construct Voronoi loss functions later in Section 3.

Voronoi cells. In general, true parameters which are fitted by exactly one component should enjoy faster estimation rates than those approximated by more than one component. Therefore, in order to capture the convergence behavior of parameter estimations accurately, we define k_0 different index sets called Voronoi cells to control the number of fitted components approaching each of the k_0 true components. More formally, for any $G \in \mathcal{O}_k(\Theta)$, the Voronoi cell $\mathcal{A}_j := \mathcal{A}_j(G)$ generated by $\theta_j^0 := (c_j^0, \Gamma_j^0, a_j^0, b_j^0, \nu_j^0)$ is defined as

$$\mathcal{A}_j := \{i \in [k] : \|\theta_i - \theta_j^0\| \leq \|\theta_i - \theta_\ell^0\|, \ \forall \ell \neq j\}, \quad (7)$$

for any $j \in [k_0]$, where $\theta_i := (c_i, \Gamma_i, a_i, b_i, \nu_i)$. An illustration of Voronoi cells is given in Appendix A. Notably, the cardinality of each Voronoi cell $\mathcal{A}_j$ is exactly the number of fitted components approximating the true component θ_j^0.

3 PARAMETER ESTIMATION RATES

In this section, we conduct the convergence analysis for parameter estimation in the GMoE model under the Type I and Type II settings in Section 3.1 and Section 3.2, respectively. Then, we sketch the proof for main results in both settings in Section 3.3.

3.1 Type I Setting

Let us recall that under this setting, all the values of $c_1^0, c_2^0, \ldots, c_{k_0}^0$ are non-zero. Although the exterior interaction between the parameters of two functions $f_{\mathcal{L}}$ and $f_{\mathcal{D}}$ mentioned in equation (6) does not hold in this scenario, we encounter two interior interactions among parameters b, ν and c, Γ via the following PDEs:

$$\frac{\partial^2 f_{\mathcal{D}}}{\partial b^2} = 2\frac{\partial f_{\mathcal{D}}}{\partial \nu}, \quad \frac{\partial^2 f_{\mathcal{L}}}{\partial c \, \partial c^\top} = 2\frac{\partial f_{\mathcal{L}}}{\partial \Gamma}. \quad (8)$$

System of polynomial equations. To precisely characterize the estimation rates for those parameters, we need to consider the solvability of a system of polynomial equations which was previously studied in Ho and Nguyen (2016a). In particular, for each $m \geq 2$, let $\bar{r}(m)$ be the smallest positive integer r such that the system:

$$\sum_{l=1}^{m} \sum_{\substack{n_1, n_2 \in \mathbb{N}: \\ n_1 + 2n_2 = s}} \frac{p_l^2 \, q_{1l}^{n_1} \, q_{2l}^{n_2}}{n_1! \, n_2!} = 0, \quad s = 1, 2, \ldots, r, \quad (9)$$

does not have any non-trivial solutions for the unknown variables $\{p_l, q_{1l}, q_{2l}\}_{l=1}^m$. Here, a solution is called non-trivial if all the values of p_l are different from zero, whereas at least one among q_{1l} is non-zero. The following lemma gives us the values of $\bar{r}(m)$ at some specific points m.

Lemma 1 (Proposition 2.1, (Ho and Nguyen, 2016a))**.** *When $m = 2$, we have that $\bar{r}(m) = 4$, while for $m = 3$, we get $\bar{r}(m) = 6$. If $m \geq 4$, then $\bar{r}(m) \geq 7$.*

Proof of Lemma 1 is in Ho and Nguyen (2016a). Now, we are ready to introduce a Voronoi loss function used for this setting.

Voronoi loss function. For simplicity, we denote $\Delta c_{ij} := c_i - c_j^0$, $\Delta \Gamma_{ij} := \Gamma_i - \Gamma_j^0$, $\Delta a_{ij} := a_i - a_j^0$, $\Delta b_{ij} := b_i - b_j^0$, $\Delta \nu_{ij} := \nu_i - \nu_j^0$ and $\bar{r}_j := \bar{r}(|\mathcal{A}_j|)$. Additionally, we also define mappings $K_{ij} : \mathbb{N}^5 \to \mathbb{R}$ such that $K_{ij}(\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5) := \|\Delta c_{ij}\|^{\kappa_1} + \|\Delta \Gamma_{ij}\|^{\kappa_2} + \|\Delta a_{ij}\|^{\kappa_3} + |\Delta b_{ij}|^{\kappa_4} + |\Delta \nu_{ij}|^{\kappa_5}$, for any $j \in [k_0]$ and $i \in \mathcal{A}_j$. Then, the Voronoi loss function $\overline{D}(G, G_0)$ of interest in this setting is given by:

$$\overline{D}(G, G_0) := \sum_{\substack{j : |\mathcal{A}_j| > 1, \\ i \in \mathcal{A}_j}} \pi_i K_{ij}\Big(\bar{r}_j, \frac{\bar{r}_j}{2}, 2, \bar{r}_j, \frac{\bar{r}_j}{2}\Big) + \sum_{\substack{j : |\mathcal{A}_j| = 1, \\ i \in \mathcal{A}_j}} \pi_i K_{ij}(1, 1, 1, 1, 1) + \sum_{j=1}^{k_0} \Big| \sum_{i \in \mathcal{A}_j} \pi_i - \pi_j^0 \Big|. \quad (10)$$

Given this loss function, we capture parameter estimation rates in the GMoE model in the following theorem.

Theorem 1. *Under the Type I setting, the Total Variation lower bound $V(p_G, p_{G_0}) \gtrsim \overline{D}(G, G_0)$ holds for any $G \in \mathcal{O}_k(\Theta)$, which implies that there exists a universal constant $C_3 > 0$ depending on G_0 and Θ satisfying*

$$\mathbb{P}(\overline{D}(\widehat{G}_n, G_0) > C_3 \sqrt{\log(n)/n}) \lesssim n^{-C_4},$$

where $C_4 > 0$ is a constant that depends only on Θ.

Proof of Theorem 1 is in Appendix B.1. It follows from Theorem 1 that the discrepancy $\overline{D}(\widehat{G}_n, G_0)$ vanishes at a rate of order $\mathcal{O}(n^{-1/2})$ up to a logarithmic constant, which leads to following observations: **(i)** True parameters $c_j^0, \Gamma_j^0, a_j^0, b_j^0, \nu_j^0$, which are fitted by exactly one component, share the same estimation rate of order $\mathcal{O}(n^{-1/2})$; **(ii)** By contrast, the rates for estimating parameters fitted by more than one element are significantly slower. In particular, the estimation rates for c_j^0, b_j^0 are of order $\mathcal{O}(n^{-1/2\bar{r}(|\mathcal{A}_j^n|)})$, whereas those for Γ_j^0, ν_j^0 are of order $\mathcal{O}(n^{-1/\bar{r}(|\mathcal{A}_j^n|)})$ in which $\mathcal{A}_j^n := \mathcal{A}_j(\widehat{G}_n)$. For instance, if we have $|\mathcal{A}_j^n| = 3$, then Lemma 1 indicates that the previous two rates become $\mathcal{O}(n^{-1/12})$ and $\mathcal{O}(n^{-1/6})$, respectively. These slow rates are owing to the interior interactions among those parameters in equation (8). Meanwhile, a_j^0 admits a much faster rate of order $\mathcal{O}(n^{-1/4})$ as it does not interact with other parameters.

3.2 Type II Setting

Next, we consider the Type II setting, namely when at least one among $c_1^0, c_2^0, \ldots, c_{k_0}^0$ is equal to vector $\mathbf{0}_d$. Without loss of generality, we assume that $c_1^0 = c_2^0 = \ldots = c_{\widetilde{k}}^0 = \mathbf{0}_d$, while $c_{\widetilde{k}+1}^0, c_{\widetilde{k}+2}^0, \ldots, c_{k_0}^0$ are different from $\mathbf{0}_d$. Under this setting, we encounter not only the two interior interactions in equation (8) but also the exterior interaction expressed by the following PDE:

$$\frac{\partial^2 F}{\partial c\, \partial b}(X, Y|\theta_j^0) = \Gamma^{-1} \cdot \frac{\partial F}{\partial a}(X, Y|\theta_j^0), \tag{11}$$

where $F(X, Y|\theta) := f_{\mathcal{L}}(X|c, \Gamma) f_{\mathcal{D}}(Y|a^\top X + b, \nu)$ and $\theta := (c, \Gamma, a, b, \nu)$. This phenomenon poses a lot of challenges in the parameter estimation problem. Therefore, we will only present the results when $d = 1$ for simplicity, while those for the setting $d > 1$ can be argued in a similar fashion but with more complex notations.

System of polynomial equations. Due to the emergence of the exterior interaction, we need to control the solvability of a totally new system of polynomial equations, which is given by

$$\sum_{l=1}^{m} \sum_{\alpha \in \mathcal{J}_{\ell_1, \ell_2}} \frac{p_l^2\, q_{1l}^{\alpha_1}\, q_{2l}^{\alpha_2}\, q_{3l}^{\alpha_3}\, q_{4l}^{\alpha_4}\, q_{5l}^{\alpha_5}}{\alpha_1!\, \alpha_2!\, \alpha_3!\, \alpha_4!\, \alpha_5!} = 0, \tag{12}$$

for all $\ell_1, \ell_2 \geq 0$ satisfying $1 \leq \ell_1 + \ell_2 \leq r$, where $\mathcal{J}_{\ell_1, \ell_2} := \{\alpha = (\alpha_l)_{l=1}^5 \in \mathbb{N}^5 : \alpha_1 + 2\alpha_2 + \alpha_3 = \ell_1,\ \alpha_3 + \alpha_4 + 2\alpha_5 = \ell_2\}$. Now, we define $\widetilde{r}(m)$ as the smallest natural number r such that the system in equation (12) does not have any non-trivial solutions for the unknown variables $\{p_l, q_{1l}, q_{2l}, q_{3l}, q_{4l}, q_{5l}\}_{l=1}^m$, namely, all of p_l are non-zero, whereas at least one among q_{4l} is different from zero. The following lemma establishes a connection between $\widetilde{r}(m)$ and $\bar{r}(m)$ as well as provides the values of $\widetilde{r}(m)$ given some specific choices of m.

Lemma 2. *In general, we have $\widetilde{r}(m) \leq \bar{r}(m)$ for all $m \in \mathbb{N}$. Furthermore, the equality occurs when $m = 2$ and $m = 3$, meaning that $\widetilde{r}(2) = 4$ and $\widetilde{r}(3) = 6$.*

Proof of Lemma 2 is in Appendix C.3. Next, we introduce a Voronoi loss function tailored to this setting.

Voronoi loss function. Firstly, let us reformulate the mappings K_{ij} defined in Section 3.1 for $d = 1$ as $K_{ij}(\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5) := |\Delta c_{ij}|^{\kappa_1} + |\Delta \Gamma_{ij}|^{\kappa_2} + |\Delta a_{ij}|^{\kappa_3} + |\Delta b_{ij}|^{\kappa_4} + |\Delta \nu_{ij}|^{\kappa_5}$. In addition, we denote $\widetilde{r}_j := \widetilde{r}(|\mathcal{A}_j|)$ and $\bar{r}_j := \bar{r}(|\mathcal{A}_j|)$, for any $j \in [k_0]$. Then, the Voronoi loss of interest $\widetilde{D}(G, G_0)$ is defined as follows:

$$\begin{aligned}
\widetilde{D}(G, G_0) :=& \sum_{\substack{j \in [\widetilde{k}]: |\mathcal{A}_j| > 1, \\ i \in \mathcal{A}_j}} \pi_i K_{ij}\Big(\widetilde{r}_j, \frac{\widetilde{r}_j}{2}, \frac{\widetilde{r}_j}{2}, \widetilde{r}_j, \frac{\widetilde{r}_j}{2}\Big) \\
&+ \sum_{\substack{j \in [k_0] \setminus [\widetilde{k}]: |\mathcal{A}_j| > 1, \\ i \in \mathcal{A}_j}} \pi_i K_{ij}\Big(\bar{r}_j, \frac{\bar{r}_j}{2}, 2, \bar{r}_j, \frac{\bar{r}_j}{2}\Big) \\
&+ \sum_{\substack{j \in [k_0]: |\mathcal{A}_j| = 1, \\ i \in \mathcal{A}_j}} \pi_i K_{ij}(1, 1, 1, 1, 1) + \sum_{j=1}^{k_0} \Big| \sum_{i \in \mathcal{A}_j} \pi_i - \pi_j^0 \Big|.
\end{aligned} \tag{13}$$

Given the above loss function, we derive the rates for estimating parameters under the Type II setting in the following theorem.

Theorem 2. *Under the Type II setting, the Total Variation lower bound $V(p_G, p_{G_0}) \gtrsim \widetilde{D}(G, G_0)$ holds for any $G \in \mathcal{O}_k(\Theta)$, which indicates that we can find a constant $C_5 > 0$ depending on G_0, Θ such that*

$$\mathbb{P}(\widetilde{D}(\widehat{G}_n, G_0) > C_5 \sqrt{\log(n)/n}) \lesssim n^{-C_6},$$

where $C_6 > 0$ is a constant that depends only on Θ.

Proof of Theorem 2 is in Appendix B.2. Similar to Theorem 1, the Voronoi loss $\widetilde{D}(\widehat{G}_n, G_0)$ also converges to zero at a rate of order $\mathcal{O}(n^{-1/2})$ (up to a logarithmic term) under the Type II setting. Moreover, true parameters $c_j^0, \Gamma_j^0, a_j^0, b_j^0, \nu_j^0$ enjoy the same estimation rates as their counterparts in Section 3.1 for any $j \in [k_0] : |\mathcal{A}_j^n| = 1$ and $j \in [k_0] \setminus [\widetilde{k}] : |\mathcal{A}_j^n| > 1$. However, the difference in the convergence behavior occurs when $j \in [\widetilde{k}] : |\mathcal{A}_j^n| > 1$. In particular, the rates for estimating parameters a_j^0 now drop substantially to $\mathcal{O}(n^{-1/\widetilde{r}(|\mathcal{A}_j|)})$ in comparison with $\mathcal{O}(n^{-1/4})$ under the Type I settings. This phenomenon happens due to the interaction of a_j^0 with parameters c_j^0, b_j^0 via the PDE in equation (11).

3.3 Proof Sketch

Since arguments used for the proof of Theorem 1 are included in that of Theorem 2, we will present the former proof sketch implicitly inside the latter. In particular, we focus on establishing the bound $\inf_{G \in \mathcal{O}_k(\Theta)} V(p_G, p_{G_0})/\widetilde{D}(G, G_0) > 0$ under the Type II setting when $d = 1$. For that purpose, we will respectively demonstrate its local and global versions by contradiction as follows:

Local bound: We wish to prove that

$$\lim_{\varepsilon > 0} \inf_{G \in \mathcal{O}_k(\Theta), \widetilde{D}(G, G_0) \leq \varepsilon} V(p_G, p_{G_0})/\widetilde{D}(G, G_0) > 0.$$

Assume that this bound does not hold, then we can find a sequence $G_n = \sum_{i=1}^{k_n} \pi_i^n \delta_{\theta_i^n} \in \mathcal{O}_k(\Theta)$, where $\theta_i^n := (c_i^n, \Gamma_i^n, a_i^n, b_i^n, \nu_i^n)$, such that $V(p_{G_n}, p_{G_0})/\widetilde{D}(G_n, G_0)$ and $\widetilde{D}(G_n, G_0)$ both vanish as $n \to \infty$. Now, we decompose $\Xi_n := p_{G_n}(X,Y) - p_{G_0}(X,Y)$ as

$$\Xi_n = \sum_{j=1}^{k_0} \sum_{i \in \mathcal{A}_j} \pi_i^n [F(X,Y|\theta_i^n) - F(X,Y|\theta_j^0)] + \sum_{j=1}^{k_0} \left(\sum_{i \in \mathcal{A}_j} \pi_i^n - \pi_j^0 \right) F(X,Y|\theta_j^0),$$

where $\theta_j^0 := (c_j^0, \Gamma_j^0, a_j^0, b_j^0, \nu_j^0)$. Let us denote $h_1(X,a,b) = a^\top X + b$ for any $a \in \mathbb{R}^d, b \in \mathbb{R}$. Then, for $i \in \mathcal{A}_j$ and $i' \in \mathcal{A}_{j'}$ where $j \in [\widetilde{k}]$ and $j' \in [k_0] \setminus [\widetilde{k}]$, we invoke the Taylor expansion up to some orders r_{1j} and $r_{2j'}$ (we will choose later) for $F(X,Y|\theta_i^n)$ and $F(X,Y|\theta_{i'}^n)$, respectively, as follows:

$$\begin{aligned}
&F(X,Y|\theta_i^n) - F(X,Y|\theta_j^0) \\
&= \sum_{\ell_1+\ell_2=1}^{2r_{1j}} Q^n_{\ell_1,\ell_2}(j) \cdot X^{\ell_1} f_{\mathcal{L}}(X|c_j^0, \Gamma_j^0) \\
&\quad \times \frac{\partial^{\ell_2} f_{\mathcal{D}}}{\partial h_1^{\ell_2}}(Y|a_j^0 X + b_j^0, \nu_j^0) + R_{1ij}(X,Y), \\
&F(X,Y|\theta_{i'}^n) - F(X,Y|\theta_{j'}^0) \\
&= R_{2i'j'}(X,Y) + \sum_{\alpha_3=0}^{r_{2j'}} \sum_{\tau_1+\tau_2=0}^{2(r_{2j'}-\alpha_3)} T^n_{\alpha_3,\tau_1,\tau_2}(j') \cdot X^{\alpha_3} \\
&\quad \times \frac{\partial^{\tau_1} f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_{j'}^0, \Gamma_{j'}^0) \frac{\partial^{\alpha_3+\tau_2} f_{\mathcal{D}}}{\partial h_1^{\alpha_3+\tau_2}}(Y|a_{j'}^0 X + b_{j'}^0, \nu_{j'}^0).
\end{aligned}$$

Here $R_{1ij}(X,Y)$ and $R_{2i'j'}(X,Y)$ are Taylor remainders such that their ratios to $\widetilde{D}(G_n, G_0)$ vanishes as $n \to \infty$. Thus, we can treat $\Xi_n/\widetilde{D}(G_n, G_0)$ as a linear combination of linearly independent terms

$$X^{\ell_1} \cdot f_{\mathcal{L}}(X|c_j^0, \Gamma_j^0) \cdot \frac{\partial^{\ell_2} f_{\mathcal{D}}}{\partial h_1^{\ell_2}}(Y|a_j^0 X + b_j^0, \nu_j^0),$$

$$X^{\alpha_3} \cdot \frac{\partial^{\tau_1} f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_{j'}^0, \Gamma_{j'}^0) \cdot \frac{\partial^{\alpha_3+\tau_2} f_{\mathcal{D}}}{\partial h_1^{\alpha_3+\tau_2}}(Y|a_{j'}^0 X + b_{j'}^0, \nu_{j'}^0)$$

associated with coefficients $Q^n_{\ell_1,\ell_2}(j)$ and $T^n_{\alpha_3,\tau_1,\tau_2}(j')$, respectively. Moreover, it follows from Fatou's lemma that $\Xi_n/\widetilde{D}(G_n, G_0)$ approaches zero when $n \to \infty$. Consequently, all the coefficients in the representation of $\Xi_n/\widetilde{D}(G_n, G_0)$, i.e. $Q^n_{\ell_1,\ell_2}(j)/\widetilde{D}(G_n, G_0)$ and $T^n_{\alpha_3,\tau_1,\tau_2}(j')/\widetilde{D}(G_n, G_0)$, go to zero as $n \to \infty$. Therefore, in order to point out a contradiction, we need to choose the values of r_{1j} and $r_{2j'}$ such that at least one among these coefficients does not vanish. As a result, we achieve the aforementioned local bound. Now, we will show how to determine such values of r_{1j} and $r_{2j'}$. It is worth noting that if we set $\widetilde{k} = 0$, then Type II settings reduces to Type I settings and we only need to deal with $r_{2j'}$ as follows:

Type I setting: We will specify an appropriate of $r_{2j'}$ during proving by contradiction that not all the coefficients $T^n_{\alpha_3,\tau_1,\tau_2}(j')/\widetilde{D}(G_n, G_0)$ tend to zero. Assume that these coefficients all vanish, then we extract some useful limits among them for our arguments and end up with the following system of polynomial equations:

$$\sum_{i' \in \mathcal{A}_{j'}} \sum_{n_1+2n_2=s} \frac{p_l^2 \, q_{1l}^{n_1} \, q_{2l}^{n_2}}{n_1! \, n_2!} = 0, \quad s = 1, 2, \ldots, r_{2j'}.$$

By construction, this system must have at least one non-trivial solution. Thus, to contradict this condition, we set $r_{2j'} = \bar{r}(|\mathcal{A}_{j'}|)$, which makes the above system has no non-trivial solutions.

Type II setting: When $\widetilde{k} > 0$, i.e. there exist some zero-valued parameter c_j, we will keep $r_{2j'} = \bar{r}(|\mathcal{A}_{j'}|)$ for all $j' \in [k_0] \setminus [\widetilde{k}]$ and find the desired values of r_{1j} for $j \in [\widetilde{k}]$ by showing by contradiction that not all the coefficients $Q^n_{\ell_1,\ell_2}(j)/\widetilde{D}(G_n, G_0)$ go to zero. En route to pointing out a contradiction to the hypothesis, we come across a more complex system of polynomial equations than its counterpart in the previous setting, specifically

$$\sum_{i \in \mathcal{A}_j} \sum_{\alpha \in \mathcal{J}_{\ell_1,\ell_2}} \frac{p_i^2 \, q_{1i}^{\alpha_1} \, q_{2i}^{\alpha_2} \, q_{3i}^{\alpha_3} \, q_{4i}^{\alpha_4} \, q_{5i}^{\alpha_5}}{\alpha_1! \, \alpha_2! \, \alpha_3! \, \alpha_4! \, \alpha_5!} = 0,$$

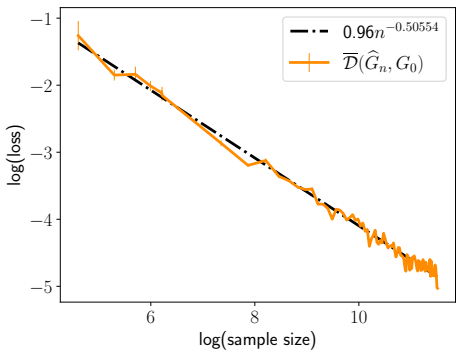
for all $\ell_1, \ell_2 \geq 0$ such that $1 \leq \ell_1 + \ell_2 \leq r_{1j}$, where $\mathcal{J}_{\ell_1,\ell_2} := \{\alpha = (\alpha_i)_{i=1}^5 \in \mathbb{N}^5 : \alpha_1 + 2\alpha_2 + \alpha_3 = \ell_1, \ \alpha_3 + \alpha_4 + 2\alpha_5 = \ell_2\}$. Since this system necessarily has a non-trivial solution, we choose $r_{1j} = \widetilde{r}(|\mathcal{A}_j|)$ so that it admits only trivial solutions, which contradicts the previous claim. Consequently, we can find a constant $\varepsilon' > 0$ such that

$$\inf_{G \in \mathcal{O}_k(\Theta), \widetilde{D}(G,G_0) \leq \varepsilon'} V(p_G, p_{G_0})/\widetilde{D}(G, G_0) > 0.$$

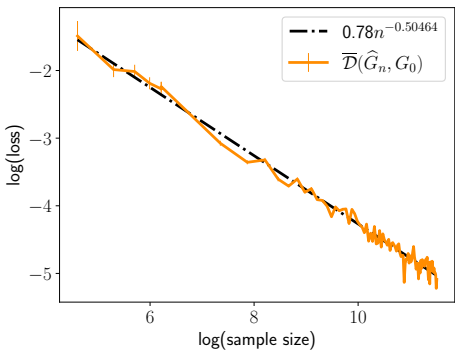
Global bound: Thus, to complete the proof, it is sufficient to demonstrate the global bound

$$\inf_{G \in \mathcal{O}_k(\Theta), \widetilde{D}(G,G_0) > \varepsilon'} V(p_G, p_{G_0})/\widetilde{D}(G, G_0) > 0.$$

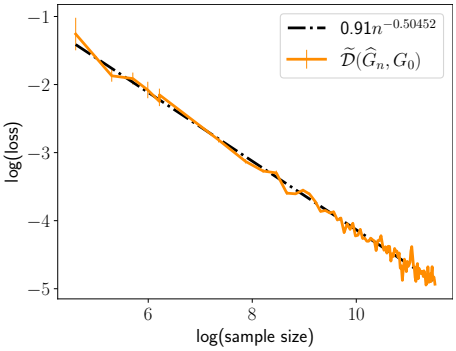
If this bound did not hold, there would be a mixing measure $G' \in \mathcal{O}_k(\Theta)$ that satisfies $p_{G'}(X,Y) = p_{G_0}(X,Y)$ for almost surely (X,Y), which leads to $G' \equiv G_0$ by Proposition 1. As a result, we obtain $\widetilde{D}(G', G_0) = 0$, which contradicts the constraint that $\widetilde{D}(G', G_0) > \varepsilon'$. Hence, the proof sketch is completed.



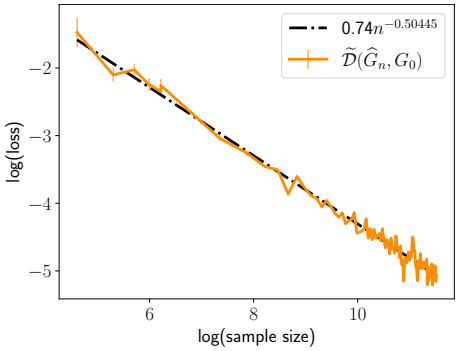
(a) Model I, $k = 4$



(b) Model I, $k = 5$



(c) Model II, $k = 4$



(d) Model II, $k = 5$

Figure 1: Log-log scaled plots of the empirical mean of discrepancies $\overline{D}(\widehat{G}_n, G_0)$ and $\widetilde{D}(\widehat{G}_n, G_0)$ and G_0 (orange lines with error bars) and least-squares fitted linear regression (black dash-dotted lines) when $d = 1$ and $k_0 = 3$.

4 EXPERIMENTS

In this section, we empirically validate the convergence rates of parameter estimation in four GMoE models which satisfy the assumptions of Type I and Type II settings, respectively, when $k_0 = 3$. Note that for simplicity, we only perform a simulation study to illustrate the convergence rates of Theorems 1 and 2 for the GMoE model when X lies in one- and two-dimensional space with unknown location and scale parameters. All code to reproduce our simulation study is publicly available[1] and all simulations below were performed in Python 3.9.13 on a standard Unix machine.

Numerical schemes. In Model I, we set G_0 as follows:

$$\sum_{j=1}^{3} \pi_j^0 \delta_{(c_j^0, \Gamma_j^0, a_j^0, b_j^0, \nu_j^0)} = 0.3\delta_{(-0.1,\, 0.04,\, 0.40,\, 0.34,\, 0.01)} + 0.4\delta_{(0.1,\, 0.02,\, -0.71,\, -0.33,\, 0.03)} + 0.3\delta_{(0.5,\, 0.01,\, 0,\, 0.2,\, 0.02)}.$$

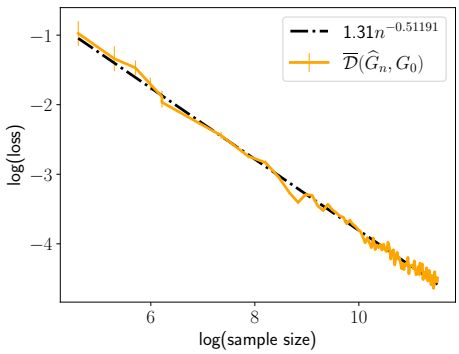
For Model II, we consider the same setting as in Model I but with $c_1^0 = 0$ and $b_1^0 = 0.3$. To demonstrate the claim that the empirical convergence rates of parameter estimation under the Type I (Model III) and Type II (Model IV) settings also hold in higher dimensions, we conduct a numerical simulation for $d = 2$ and $k_0 = 3$. In Model III, we set G_0 as

$$\sum_{j=1}^{3} \pi_j^0 \delta_{(c_j^0, \Gamma_j^0, a_j^0, b_j^0, \nu_j^0)} = 0.3\delta_{(-0.1\cdot \mathbf{1}_d,\, 0.04\cdot \mathbf{I}_d,\, 0.4\cdot \mathbf{1}_d,\, 0.34,\, 0.01)} + 0.4\delta_{(0.1\cdot \mathbf{1}_d,\, 0.02\cdot \mathbf{I}_d,\, -0.71\cdot \mathbf{1}_d,\, -0.33,\, 0.03)} + 0.3\delta_{(0.5\cdot \mathbf{1}_d,\, 0.01\cdot \mathbf{I}_d,\, \mathbf{0}_d,\, 0.2,\, 0.02)},$$

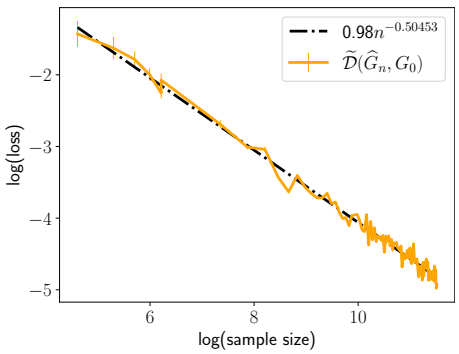
where $\mathbf{1}_d = (1, 1)$, $\mathbf{0}_d = (0, 0)$ and $\mathbf{I}_d$ is the identity matrix of size d. In Model IV, we consider the same setting of G_0 as in Model III but with $c_1^0 = \mathbf{0}_d$ and $b_1^0 = 0.3$.

Numerical details. In accordance with the hierarchical GMoE setting of (2), we generate 20 samples $(X_i, Y_i)_{i \in [n]}$ of size n for each setting, given 100 different choices of sample size n between 10^2 and 10^5. Then, we compute the MLE $\widehat{G}_n$ w.r.t. a number of components k for each sample. For both of these settings, we choose $k \in \{k_0 + 1, k_0 + 2\}$ with corresponding $\overline{r}, \widetilde{r} \in \{4, 6\}$ using Lemmas 1 and 2. Here we implement the MLE using the EM algorithm for GMoE. This is a simplification of a general hybrid GMoE-EM from (Deleforge et al., 2015, Section 5). We choose the convergence criteria $\epsilon = 10^{-5}$ and 2000 maximum EM iterations. Our goal is to illustrate the theoretical properties of the estimator $\widehat{G}_n$. Therefore, we have initialized the EM al-

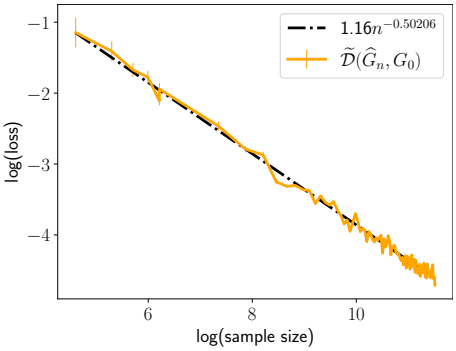
[1] https://github.com/Trung-TinNGUYEN/CRPE-GMoE



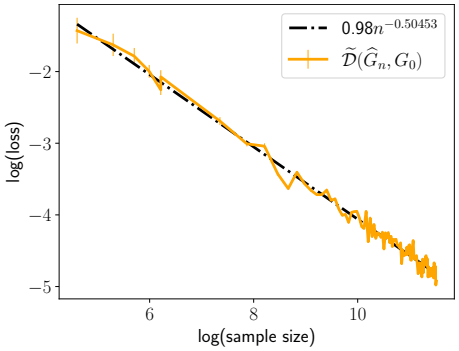
(a) Model III, $k = 4$



(b) Model III, $k = 5$



(c) Model IV, $k = 4$



(d) Model IV, $k = 5$

Figure 2: Log-log scaled plots of the empirical mean of discrepancies $\overline{D}(\widehat{G}_n, G_0)$ and $\widetilde{D}(\widehat{G}_n, G_0)$ (orange lines with error bars) and least-squares fitted linear regression (black dash-dotted lines) when $d = 2$ and $k_0 = 3$.

gorithm in a favourable way. More specifically, we first randomly partitioned the set $\{1, \ldots, k\}$ into k_0 index sets $J_1, \ldots, J_{k_0}$, each containing at least one point, for any given k and k_0 and for each replication. Finally, we sampled c_j^0 (resp. $\Gamma_j^0, a_j^0, b_j^0, \nu_j^0$) from a unique Gaussian distribution centered on c_t^0 (resp. $\Gamma_t^0, a_t^0, b_t^0, \nu_t^0$), with vanishing covariance so that $j \in J_t$.

Empirical convergence rates. The empirical mean of discrepancies $\overline{D}$ and $\widetilde{D}$ between $\widehat{G}_n$ and G_0, and the choice of k for Models I-II are reported in Figure 1. It can be observed from Figure 1 that those average discrepancies vanish at a rate of order $\mathcal{O}(n^{-1/2})$, which matches the results of Theorems 1 and 2, where the only theoretical assumption that can be violated is the global convergence of the MLE. Note that the use of the joint density function allows the GMoE to be linked to a hierarchical mixture model, which guarantees global convergence for parameter estimation for arbitrary dimensions, see recent advances, e.g., (Kwon et al., 2021; Kwon and Caramanis, 2020; Kwon et al., 2019). We can therefore guarantee that the rates in Theorems 1 and 2 also hold in higher dimensions. Indeed, it can be observed from Figure 2 that the average discrepancies $\overline{D}(\widehat{G}_n, G_0)$ and $\widetilde{D}(\widehat{G}_n, G_0)$ also approach zero at the rate of order $\mathcal{O}(n^{-1/2})$ for $d = 2$, confirming the empirical behaviour of Theorems 1 and 2 under the high dimensional settings.

5 CONCLUSION

In this paper, we conduct a convergence analysis for density estimation and parameter estimation in the Gaussian-gated mixture of experts (GMoE) under two complement settings of location parameters of the gating function. We demonstrate that the density estimation rate remains parametric on the sample size under both settings. On the other hand, due to several challenges induced by the interior and exterior interactions among parameters arising in those settings, we have to solve two complex systems of polynomial equations and then propose two corresponding novel Voronoi loss functions among parameters. We show that these Voronoi losses are able to capture the dependence of parameter estimation rates on the number of fitted components, which are more accurate than those characterized by the generalized Wasserstein loss used in previous works. We believe that our current techniques can be extended to the GMoE model with general experts in Ho et al. (2022) and to the hierarchical MoE for exponential family models in Jiang and Tanner (1999a). In addition, understanding the convergence behavior of least squares estimation under the deterministic MoE model (Nguyen et al., 2024c) with Gaussian gate is also a potential direction. However, we leave such non-trivial developments for future work.

Acknowledgements

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

References

J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012. ISSN 1573-1375. doi: 10.1007/s11222-011-9236-1. (Cited on page 2.)

L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1):33–73, 2007. doi: 10.1007/s00440-006-0011-8. Publisher: Springer. (Cited on page 2.)

F. Chamroukhi and B.-T. Huynh. Regularized Maximum Likelihood Estimation and Feature Selection in Mixtures-of-Experts Models. *Journal de la Société Française de Statistique*, 160(1):57–85, 2019. (Cited on page 2.)

A. Deleforge, F. Forbes, and R. Horaud. High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25(5):893–911, 2015. ISSN 1573-1375. doi: 10.1007/s11222-014-9461-5. (Cited on pages 1 and 8.)

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. doi: 10.1111/j.2517-6161.1977.tb01600.x. Publisher: John Wiley & Sons, Ltd. (Cited on page 1.)

C. Diani, G. Galimberti, and G. Soffritti. Multivariate cluster-weighted models based on seemingly unrelated linear regression. *Computational Statistics & Data Analysis*, page 107451, Feb. 2022. ISSN 0167-9473. doi: 10.1016/j.csda.2022.107451. (Cited on page 1.)

T. G. Do, H. K. Le, T. Nguyen, Q. Pham, B. T. Nguyen, T.-N. Doan, C. Liu, S. Ramasamy, X. Li, and S. HOI. HyperRouter: Towards Efficient Training and Inference of Sparse Mixture of Experts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, Dec. 2023. Association for Computational Linguistics. (Cited on page 1.)

W. Fedus, J. Dean, and B. Zoph. A review of sparse expert models in deep learning. *arXiv preprint arXiv:2209.01667*, 2022a. (Cited on page 1.)

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(120):1–39, 2022b. (Cited on page 1.)

F. Forbes, H. D. Nguyen, T. Nguyen, and J. Arbel. Mixture of expert posterior surrogates for approximate Bayesian computation. In *JDS 2022 - 53èmes Journées de Statistique de la Société Française de Statistique (SFdS)*, Lyon, France, June 2022a. (Cited on page 1.)

F. Forbes, H. D. Nguyen, T. Nguyen, and J. Arbel. Summary statistics and discrepancy measures for approximate Bayesian computation via surrogate posteriors. *Statistics and Computing*, 32(5):85, Oct. 2022b. ISSN 1573-1375. doi: 10.1007/s11222-022-10155-6. (Cited on pages 1 and 2.)

J. Fritsch, M. Finke, and A. Waibel. Adaptively growing hierarchical mixtures of experts. In *Advances in Neural Information Processing Systems*, volume 9, 1996. (Cited on page 1.)

A. Guha, N. Ho, and X. Nguyen. On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli*, 27(4):2159 – 2188, 2021. doi: 10.3150/20-BEJ1275. Publisher: Bernoulli Society for Mathematical Statistics and Probability. (Cited on pages 2 and 4.)

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 1.)

N. Ho and X. Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6):2726 – 2755, 2016a. doi: 10.1214/16-AOS1444. Publisher: Institute of Mathematical Statistics and Bernoulli Society. (Cited on pages 2 and 5.)

N. Ho and X. Nguyen. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10(1):271–307, 2016b. Publisher: The Institute of Mathematical Statistics and the Bernoulli Society. (Cited on page 2.)

N. Ho and X. Nguyen. Singularity Structures and Impacts on Parameter Estimation in Finite Mixtures of Distributions. *SIAM Journal on Mathematics of Data Science*, 1(4):730–758, Jan. 2019. doi: 10.1137/18M122947X. Publisher: Society for Industrial and Applied Mathematics. (Cited on page 2.)

N. Ho, C.-Y. Yang, and M. I. Jordan. Convergence Rates for Gaussian Mixtures of Experts. *Journal of Machine Learning Research*, 23(323):1–81, 2022. (Cited on pages 2 and 9.)

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural*

computation, 3(1):79–87, 1991. Publisher: MIT Press. (Cited on page 1.)

W. Jiang and M. A. Tanner. Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *Annals of Statistics*, pages 987–1011, 1999a. (Cited on pages 2 and 9.)

W. Jiang and M. A. Tanner. Hierarchical Mixtures-of-Experts for Generalized Linear Models: Some Results on Denseness and Consistency. In D. Heckerman and J. Whittaker, editors, *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*, volume R2 of *Proceedings of Machine Learning Research*. PMLR, Jan. 1999b. (Cited on page 2.)

W. Jiang and M. A. Tanner. On the identifiability of mixtures-of-experts. *Neural Networks*, 12(9):1253–1258, 1999c. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(99)00066-0. (Cited on page 2.)

M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2):181–214, 1994. Publisher: MIT Press. (Cited on page 1.)

A. Khalili. New estimation and feature selection methods in mixture-of-experts models. *Canadian Journal of Statistics*, 38(4):519–539, 2010. Publisher: Wiley Online Library. (Cited on page 2.)

B. Kugler, F. Forbes, and S. Douté. Fast Bayesian inversion for high dimensional inverse problems. *Statistics and Computing*, 32(2):31, Mar. 2022. ISSN 1573-1375. doi: 10.1007/s11222-021-10019-5. (Cited on page 1.)

J. Kwon and C. Caramanis. EM Converges for a Mixture of Many Linear Regressions. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1727–1736. PMLR, Aug. 2020. (Cited on page 9.)

J. Kwon, W. Qian, C. Caramanis, Y. Chen, and D. Davis. Global Convergence of the EM Algorithm for Mixtures of Two Component Linear Regression. In A. Beygelzimer and D. Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2055–2110. PMLR, June 2019. (Cited on page 9.)

J. Kwon, N. Ho, and C. Caramanis. On the Minimax Optimality of the EM Algorithm for Learning Two-Component Mixed Linear Regression. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1405–1413. PMLR, Apr. 2021. (Cited on page 9.)

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 1.)

T. Manole and N. Ho. Refined Convergence Rates for Maximum Likelihood Estimation under Finite Mixture Models. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14979–15006. PMLR, July 2022. (Cited on page 2.)

T. Manole and A. Khalili. Estimating the number of components in finite mixture models via the Group-Sort-Fuse procedure. *The Annals of Statistics*, 49(6): 3043 – 3069, 2021. doi: 10.1214/21-AOS2072. Publisher: Institute of Mathematical Statistics. (Cited on page 4.)

S. Masoudnia and R. Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42 (2):275–293, 2014. ISSN 1573-7462. doi: 10.1007/s10462-012-9338-y. (Cited on page 1.)

E. F. Mendes and W. Jiang. On convergence rates of mixtures of polynomial experts. *Neural computation*, 24(11):3025–3051, 2012. Publisher: MIT Press. (Cited on page 2.)

J. Moody and C. J. Darken. Fast Learning in Networks of Locally-Tuned Processing Units. *Neural Computation*, 1(2):281–294, 1989. doi: 10.1162/neco.1989.1.2.281. (Cited on page 1.)

B. Mustafa, C. R. Ruiz, J. Puigcerver, R. Jenatton, and N. Houlsby. Multimodal Contrastive Learning with LIMoE: the Language-Image Mixture of Experts. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. (Cited on page 1.)

H. Nguyen, P. Akbarian, T. Nguyen, and N. Ho. A general theory for softmax gating multinomial logistic mixture of experts. *arXiv preprint arXiv:2310.14188*, 2023a. (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*, 2023b. (Cited on page 2.)

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 2.)

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 2.)

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. D. Nguyen, T. Nguyen, F. Chamroukhi, and G. J. McLachlan. Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models. *Journal of Statistical Distributions and Applications*, 8(1):13, Aug. 2021. ISSN 2195-5832. doi: 10.1186/s40488-021-00125-0. (Cited on page 2.)

T. Nguyen, F. Chamroukhi, H. D. Nguyen, and F. Forbes. Model selection by penalization in mixture of experts models with a non-asymptotic approach. In *JDS 2022 - 53èmes Journées de Statistique de la Société Française de Statistique (SFdS)*, Lyon, France, June 2022a. (Cited on page 2.)

T. Nguyen, H. D. Nguyen, F. Chamroukhi, and F. Forbes. A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models. *Electronic Journal of Statistics*, 16(2):4742 – 4822, 2022b. doi: 10.1214/22-EJS2057. Publisher: Institute of Mathematical Statistics and Bernoulli Society. (Cited on page 2.)

T. Nguyen, F. Forbes, J. Arbel, and H. D. Nguyen. Bayesian nonparametric mixture of experts for high-dimensional inverse problems. *hal-04015203*, Mar. 2023c. (Cited on page 2.)

T. Nguyen, D. N. Nguyen, H. D. Nguyen, and F. Chamroukhi. A non-asymptotic theory for model selection in high-dimensional mixture of experts via joint rank and variable selection. *Preprint hal-03984011*, Feb. 2023d. (Cited on page 2.)

T. Nguyen, H. D. Nguyen, F. Chamroukhi, and G. J. McLachlan. Non-asymptotic oracle inequalities for the Lasso in high-dimensional mixture of experts. *arXiv:2009.10622*, Feb. 2023e. (Cited on page 2.)

A. Norets and D. Pati. Adaptive Bayesian estimation of conditional densities. *Econometric Theory*, 33(4): 980–1012, 2017. doi: 10.1017/S0266466616000220. Publisher: Cambridge University Press. (Cited on page 1.)

A. Norets and J. Pelenis. Adaptive Bayesian estimation of conditional discrete-continuous distributions with an application to stock market trading activity. *Journal of Econometrics*, 2021. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2021.11.004. (Cited on page 1.)

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 1.)

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 1.)

M. Sato and S. Ishii. On-line EM algorithm for the normalized gaussian network. *Neural computation*, 12(2):407–432, feb 2000. ISSN 0899-7667 (Print). doi: 10.1162/089976600300015853. (Cited on page 1.)

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*, 2017. (Cited on page 1.)

S. van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000. (Cited on page 29.)

L. Xu, M. Jordan, and G. E. Hinton. An Alternative Model for Mixtures of Experts. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1995. (Cited on page 1.)

Z. You, S. Feng, D. Su, and D. Yu. Speechmoe2: Mixture-of-Experts Model with Improved Routing. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7217–7221, 2022. doi: 10.1109/ICASSP43922.2022.9747065. (Cited on page 1.)

S. E. Yuksel, J. N. Wilson, and P. D. Gader. Twenty Years of Mixture of Experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8): 1177–1193, 2012. ISSN 2162-2388 VO - 23. doi: 10.1109/TNNLS.2012.2200299. (Cited on page 1.)

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]
   (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. [Yes]

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes]

(b) Complete proofs of all theoretical results. [Yes]

(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). [Yes]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(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]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

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 "Towards Convergence Rates for Parameter Estimation in Gaussian-gated Mixture of Experts"

In this supplementary material, we first include an illustration of Voronoi cells in Appendix A to help the readers understand this concept better. Then, we provide the proof of Theorem 1 and Theorem 2 in Appendix B. Finally, proofs for the remaining results are presented in Appendix C.

A ILLUSTRATION OF VORONOI CELLS

In this appendix, we aim to illustrate the Voronoi cells defined in Section 2. For that purpose, let us recall the definition of that concept here. In particular, for any mixing measure $G \in \mathcal{O}_k(\Theta)$, the Voronoi cell $\mathcal{A}_j := \mathcal{A}_j(G)$ generated by a true component $\theta_j^0 := (c_j^0, \Gamma_j^0, a_j^0, b_j^0, \nu_j^0)$ of G_0 is given by

$$\mathcal{A}_j := \{i \in [k] : \|\theta_i - \theta_j^0\| \leq \|\theta_i - \theta_\ell^0\|, \ \forall \ell \neq j\}, \tag{14}$$

for any $j \in [k_0]$, where $\theta_i := (c_i, \Gamma_i, a_i, b_i, \nu_i)$ is a component of G. Now, we provide an illustration of the above Voronoi cells under the setting when $k_0 = 6$ and $k = 10$ in Figure 3.

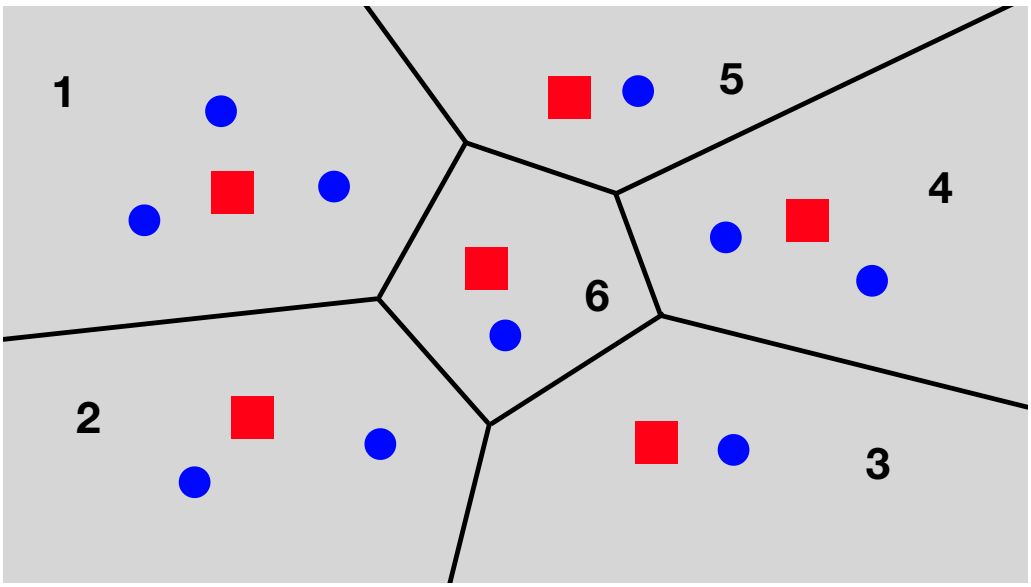


Figure 3: Illustration of Voronoi cells defined in equation (14) when $k_0 = 6$ and $k = 10$. In this figure, red squares represent for true components (i.e. components of G_0), while blue circles indicate fitted components (i.e. components of G). By definition, each Voronoi cell is generated by one true component, and its cardinality is exactly the number of corresponding fitted components. For example, the square in cell 4 is approximated by two rounds, which means that the cardinality of cell 4 is two.

Connection to Theorem 1. Under the Type I setting, parameters of the true components $(c_j^0, \Gamma_j^0, a_j^0, b_j^0, \nu_j^0)$ in cells 3, 5 and 6, which are fitted by one component, enjoy a parametric estimation rate of order $\mathcal{O}(n^{-1/2})$. Next, the rates for estimating parameters c_1^0, b_1^0 of the true component in cell 1, which are approximated by three components, stand at order $\mathcal{O}(n^{-1/2\bar{r}(3)}) = \mathcal{O}(n^{-1/12})$, while those for Γ_1^0, ν_1^0 are of order $\mathcal{O}(n^{-1/\bar{r}(3)}) = \mathcal{O}(n^{-1/6})$. Meanwhile, the estimation rate for a_1^0 is independent of the cardinality of its corresponding Voronoi cell and remains stable at order $\mathcal{O}(n^{-1/4})$.

Connection to Theorem 2. Parameter estimation rates under the Type II setting share the same behavior as those in Theorem 1 except for the rates of estimating a_j^0. More specifically, if $c_1^0 = 0$, the estimation rate for a_1^0 now depends on the cardinality of cell 1, and experiences a drop to order $\mathcal{O}(n^{-1/\bar{r}(3)}) = \mathcal{O}(n^{-1/6})$.

B PROOF OF MAIN RESULTS

Before going to the proofs for Theorems 1 and 2 in Appendices B.1 and B.2, respectively, let us define some necessary notations used throughout this appendix. Firstly, for any vector $v \in \mathbb{R}^d$, either v_i or $v^{(i)}$ represents the i-th entry of v, while the sum of its entries is abbreviated as $|v| := v_1 + v_2 + \ldots + v_d$. Next, for any vector $p \in \mathbb{N}^d$, we denote $v^p := v_1^{p_1} v_2^{p_2} \ldots v_d^{p_d}$ and $p! := p_1! p_2! \ldots p_d!$. Additionally, we sometimes use the notation h_1 and h_2 to denote the expert functions considered in this work. In particular, we define $h_1(X, a, b) = a^\top X + b$ as the mean expert function for any $X \in \mathcal{X} \subset \mathbb{R}^d$, $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$, whereas $h_2(X, \nu) = \nu$ stands for the variance expert function for any $\nu \in \mathbb{R}_+$. Finally, since parameters in the proofs for Theorem 1 and Theorem 2 belong to various high-dimensional spaces, we summarize their domains in Table 2 and Table 3, respectively, to help readers keep track of them.

	c	Γ	a	b	ν	τ_1	τ_2	α_1	α_2	α_3	α_4	α_5	ℓ_1	ℓ_2
Thm 1	$\mathbb{R}^d$	$\mathcal{S}_d^+$	$\mathbb{R}^d$	$\mathbb{R}$	$\mathbb{R}_+$	$\mathbb{N}^d$	$\mathbb{N}$	$\mathbb{N}^d$	$\mathbb{N}^{d\times d}$	$\mathbb{N}^d$	$\mathbb{N}$	$\mathbb{N}^d$	N/A	N/A

Table 2: Domains for parameters used in the proof of Theorem 1

	c	Γ	a	b	ν	τ_1	τ_2	α_1	α_2	α_3	α_4	α_5	ℓ_1	ℓ_2
Thm 2	$\mathbb{R}$	$\mathbb{R}_+$	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}_+$	$\mathbb{N}$	$\mathbb{N}$	$\mathbb{N}$	$\mathbb{N}$	$\mathbb{N}$	$\mathbb{N}$	$\mathbb{N}$	$\mathbb{N}$	$\mathbb{N}$

Table 3: Domains for parameters used in the proof of Theorem 2

B.1 Proof of Theorem 1

Our goal is to show the following inequality:

$$\inf_{G \in \mathcal{O}_{k,\beta}(\Theta)} V(p_G, p_{G_0}) / \overline{D}(G, G_0) > 0, \tag{15}$$

which implies the desired Total Variation lower bound $V(p_{\widehat{G}_n}, p_{G_0}) \gtrsim \overline{D}(\widehat{G}_n, G_0)$. Given this bound, the joint density estimation rate in Proposition 2 then leads to the convergence rate of the MLE $\widehat{G}_n$ to G_0 under the loss $\overline{D}$ as follows:

$$\mathbb{P}(\overline{D}(\widehat{G}_n, G_0) > C_3 \sqrt{\log(n)/n}) \lesssim n^{-C_4},$$

for some universal constants C_3 and C_4. Note that the infimum in equation (15) is subject to all the mixing measures in the set $\mathcal{O}_{k,\beta}(\Theta) := \{G = \sum_{i=1}^{k'} \pi_i \delta_{(c_i,\Gamma_i,a_i,b_i,\nu_i)} : 1 \leq k' \leq k,\ \sum_{i=1}^{k} \pi_i = 1,\ \pi_i \geq \beta,\ (c_i, \Gamma_i, a_i, b_i, \nu_i) \in \Theta\}$, for some positive constant β. Now, we divide the proof of inequality (15) into two parts which we refer to as local bound and global bound.

Local bound: Firstly, we will prove the local version of inequality (15):

$$\lim_{\varepsilon \to 0} \inf_{\substack{G \in \mathcal{O}_{k,\beta}(\Theta) \\ \overline{D}(G, G_0) \leq \varepsilon}} V(p_G, p_{G_0}) / \overline{D}(G, G_0) > 0. \tag{16}$$

Assume by contrary that the claim in equation (16) does not hold. Then, there exists a sequence of mixing measures $G_n = \sum_{i=1}^{k_n} \pi_i^n \delta_{(c_i^n, \Gamma_i^n, a_i^n, b_i^n, \nu_i^n)} \in \mathcal{O}_{k,\beta}(\Theta)$ such that $\overline{D}(G_n, G_0) \to 0$ and $V(p_{G_n}, p_{G_0}) / \overline{D}(G_n, G_0) \to 0$ as $n \to \infty$. Moreover, since $k_n \leq k$ for all $n \in \mathbb{N}$, we can replace (G_n) by its subsequence that admits a fixed number of atoms $k_n = k' \leq k$. Additionally, $\mathcal{A}_j = \mathcal{A}_j^n$ does not change with n for all $j \in [k_0]$.

Step 1 - Taylor expansion for density decomposition: Now, we consider the quantity

$$\begin{aligned}
&p_{G_n}(X,Y) - p_{G_0}(X,Y)\\
&= \sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\pi_i^n[f_{\mathcal{L}}(X|c_i^n,\Gamma_i^n)f_{\mathcal{D}}(Y|(a_i^n)^\top X+b_i^n,\nu_i^n) - f_{\mathcal{L}}(X|c_j^0,\Gamma_j^0)f_{\mathcal{D}}(Y|(a_j^0)^\top X+b_j^0,\nu_j^0)]\\
&+ \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\pi_i^n[f_{\mathcal{L}}(X|c_i^n,\Gamma_i^n)f_{\mathcal{D}}(Y|(a_i^n)^\top X+b_i^n,\nu_i^n) - f_{\mathcal{L}}(X|c_j^0,\Gamma_j^0)f_{\mathcal{D}}(Y|(a_j^0)^\top X+b_j^0,\nu_j^0)]\\
&+ \sum_{j=1}^{k_0}\left(\sum_{i\in\mathcal{A}_j}\pi_i^n - \pi_j^0\right) f_{\mathcal{L}}(X|c_j^0,\Gamma_j^0)f_{\mathcal{D}}(Y|(a_j^0)^\top X+b_j^0,\nu_j^0)\\
&:= A_n + B_n + E_n.
\end{aligned}$$

For each $j\in[k_0]:|\mathcal{A}_j|>1$, we perform a Taylor expansion up to the $\bar{r}(|\mathcal{A}_j|)$-th order, and then rewrite A_n with a note that $\alpha=(\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5)\in\mathbb{N}^d\times\mathbb{N}^{d\times d}\times\mathbb{N}^d\times\mathbb{N}\times\mathbb{N}$ as follows:

$$\begin{aligned}
A_n = \sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\pi_i^n\sum_{|\alpha|=1}^{\bar{r}(|\mathcal{A}_j|)}\frac{1}{\alpha!}(\Delta c_{ij}^n)^{\alpha_1}(\Delta\Gamma_{ij}^n)^{\alpha_2}(\Delta a_{ij}^n)^{\alpha_3}(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5}\\
\times\frac{\partial^{|\alpha_1|+|\alpha_2|}f_{\mathcal{L}}}{\partial c^{\alpha_1}\partial\Gamma^{\alpha_2}}(X|c_j^0,\Gamma_j^0)\cdot\frac{\partial^{|\alpha_3|+\alpha_4+\alpha_5}f_{\mathcal{D}}}{\partial a^{\alpha_3}\partial b^{\alpha_4}\partial\nu^{\alpha_5}}(Y|(a_j^0)^\top X+b_j^0,\nu_j^0)+R_1(X,Y)
\end{aligned}$$

where $R_1(X,Y)$ is a remainder term such that $R_1(X,Y)/\overline{D}(G_n,G_0)\to 0$ as $n\to\infty$, which is due to the uniform Holder continuity of a location-scale Gaussian family. Since $f_{\mathcal{L}}$ d-dimensional Gaussian density functions, we have the following partial differential equation (PDE):

$$\frac{\partial^{|\alpha_1|+|\alpha_2|}f_{\mathcal{L}}}{\partial c^{\alpha_1}\partial\Gamma^{\alpha_2}}(X|c_j^0,\Gamma_j^0) = \frac{1}{2^{|\alpha_2|}}\cdot\frac{\partial^{|\alpha_1|+2|\alpha_2|}f_{\mathcal{L}}}{\partial c^{\tau(\alpha_1,\alpha_2)}}(X|c_j^0,\Gamma_j^0),$$

where $\tau(\alpha_1,\alpha_2):=\left(\alpha_1^{(v)}+\sum_{u=1}^d(\alpha_2^{(uv)}+\alpha_2^{(vu)})\right)_{v=1}^d=\left(\alpha_1^{(v)}+2\sum_{u=1}^d\alpha_2^{(uv)}\right)_{v=1}^d\in\mathbb{N}^d$. Similarly, as $f_{\mathcal{L}}$ is an univariate Gaussian density function, then

$$\frac{\partial^{|\alpha_3|+\alpha_4+\alpha_5}f_{\mathcal{D}}}{\partial a^{\alpha_3}\partial b^{\alpha_4}\partial\nu^{\alpha_5}}(Y|(a_j^0)^\top X+b_j^0,\nu_j^0) = \frac{X^{\alpha_3}}{2^{\alpha_5}}\cdot\frac{\partial^{|\alpha_3|+\alpha_4+2\alpha_5}f_{\mathcal{D}}}{\partial h_1^{|\alpha_3|+\alpha_4+2\alpha_5}}(Y|(a_j^0)^\top X+b_j^0,\nu_j^0),$$

where $h_1(X,a,b)=a^\top X+b$ is the mean expert function. Combine these results together, A_n can be represented as follows:

$$\begin{aligned}
A_n = \sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\pi_i^n\sum_{|\alpha|=1}^{\bar{r}(|\mathcal{A}_j|)}\frac{1}{\alpha!}(\Delta c_{ij}^n)^{\alpha_1}(\Delta\Gamma_{ij}^n)^{\alpha_2}(\Delta a_{ij}^n)^{\alpha_3}(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5}\\
\times\frac{1}{2^{|\alpha_2|}}\frac{\partial^{|\alpha_1|+2|\alpha_2|}f_{\mathcal{L}}}{\partial c^{\tau(\alpha_1,\alpha_2)}}(X|c_j^0,\Gamma_j^0)\cdot\frac{X^{\alpha_3}}{2^{\alpha_5}}\frac{\partial^{|\alpha_3|+\alpha_4+2\alpha_5}f_{\mathcal{D}}}{\partial h_1^{|\alpha_3|+\alpha_4+2\alpha_5}}(Y|(a_j^0)^\top X+b_j^0,\nu_j^0)+R_1(X,Y),
\end{aligned}$$

Let $\tau_1=\tau(\alpha_1,\alpha_2)\in\mathbb{N}^d$ and $\tau_2=\alpha_4+2\alpha_5\in\mathbb{N}$, we can rewrite A_n as

$$\begin{aligned}
A_n = \sum_{j:|\mathcal{A}_j|>1}\sum_{|\alpha_3|=0}^{\bar{r}(|\mathcal{A}_j|)}\sum_{|\tau_1|+\tau_2=0}^{2(\bar{r}(|\mathcal{A}_j|)-|\alpha_3|)}\sum_{\substack{\tau(\alpha_1,\alpha_2)=\tau_1\\ \alpha_4+2\alpha_5=\tau_2}}\sum_{i\in\mathcal{A}_j}\frac{\pi_i^n}{2^{|\alpha_2|+\alpha_5}\alpha!}(\Delta c_{ij}^n)^{\alpha_1}(\Delta\Gamma_{ij}^n)^{\alpha_2}(\Delta a_{ij}^n)^{\alpha_3}\\
\times(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5}\cdot X^{\alpha_3}\frac{\partial^{|\tau_1|}f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_j^0,\Gamma_j^0)\frac{\partial^{|\alpha_3|+\tau_2}f_{\mathcal{D}}}{\partial h_1^{|\alpha_3|+\tau_2}}(Y|(a_j^0)^\top X+b_j^0,\nu_j^0)+R_1(X,Y),
\end{aligned}$$

Analogously, for each $j \in [k_0] : |\mathcal{A}_j| = 1$, by means of Taylor expansion up to the first order, B_n is rewritten as follows:

$$B_n = \sum_{j:|\mathcal{A}_j|=1} \sum_{|\alpha_3|=0}^{1} \sum_{|\tau_1|+\tau_2=0}^{2(1-|\alpha_3|)} \sum_{\substack{\tau(\alpha_1,\alpha_2)=\tau_1 \\ \alpha_4+2\alpha_5=\tau_2}} \sum_{i\in\mathcal{A}_j} \frac{\pi_i^n}{2^{|\alpha_2|+\alpha_5}\alpha!} (\Delta c_{ij}^n)^{\alpha_1} (\Delta \Gamma_{ij}^n)^{\alpha_2} (\Delta a_{ij}^n)^{\alpha_3}$$
$$\times (\Delta b_{ij}^n)^{\alpha_4} (\Delta \nu_{ij}^n)^{\alpha_5} \cdot X^{\alpha_3} \frac{\partial^{|\tau_1|} f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_j^0, \Gamma_j^0) \frac{\partial^{|\alpha_3|+\tau_2} f_{\mathcal{D}}}{\partial h_1^{|\alpha_3|+\tau_2}}(Y|(a_j^0)^\top X + b_j^0, \nu_j^0) + R_2(X,Y), \tag{17}$$

where $R_2(X,Y)$ is a remainder such that $R_2(X,Y)/\overline{D}(G_n, G_0) \to 0$ as $n \to \infty$.

It is worth noting that A_n, B_n and E_n can be treated as linear combinations of elements of the following set:

$$\mathcal{F} := \Big\{ X^{\alpha_3} \frac{\partial^{|\tau_1|} f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_j^0, \Gamma_j^0) \frac{\partial^{|\alpha_3|+\tau_2} f_{\mathcal{D}}}{\partial h_1^{|\alpha_3|+\tau_2}}(Y|(a_j^0)^\top X + b_j^0, \nu_j^0) : \ j \in [k_0], 0 \le |\alpha_3| \le \bar{r}(|\mathcal{A}_j|),$$
$$0 \le |\tau_1| + \tau_2 \le 2(\bar{r}(|\mathcal{A}_j|) - |\alpha_3|) \Big\}. \tag{18}$$

Let $T^n_{\alpha_3,\tau_1,\tau_2}(j)$ be the coefficients of

$$X^{\alpha_3} \frac{\partial^{|\tau_1|} f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_j^0, \Gamma_j^0) \frac{\partial^{|\alpha_3|+\tau_2} f_{\mathcal{D}}}{\partial h_1^{|\alpha_3|+\tau_2}}(Y|(a_j^0)^\top X + b_j^0, \nu_j^0)$$

in the representations of A_n, B_n and E_n.

Step 2 - Proof of non-vanishing coefficients by contradiction: Assume that all the coefficients in the representations of $A_n/\overline{D}(G_n, G_0)$, $B_n/\overline{D}(G_n, G_0)$ and $E_n/\overline{D}(G_n, G_0)$ go to 0 as $n \to \infty$. Then, by taking the summation of the absolute values of coefficients in $E_n/\overline{D}(G_n, G_0)$, which are $|T_{\mathbf{0}_d,\mathbf{0}_d,0}(j)|/\overline{D}(G_n, G_0)$ for all $j \in [k_0]$, we get that

$$\frac{1}{\overline{D}(G_n, G_0)} \cdot \sum_{j=1}^{k_0} \Big| \sum_{i\in\mathcal{A}_j} \pi_i^n - \pi_j^0 \Big| \to 0. \tag{19}$$

Subsequently, from the formulation of B_n in equation (17), we have

$$\frac{1}{\overline{D}(G_n, G_0)} \cdot \sum_{j:|\mathcal{A}_j|=1} \sum_{i\in\mathcal{A}_j} \pi_i^n (\|\Delta c_{ij}^n\|_1 + \|\Delta \Gamma_{ij}^n\|_1 + \|\Delta a_{ij}^n\|_1 + |\Delta b_{ij}^n| + |\Delta \nu_{ij}^n|) \to 0.$$

It follows from the topological equivalence of 1-norm and 2-norm that

$$\frac{1}{\overline{D}(G_n, G_0)} \cdot \sum_{j:|\mathcal{A}_j|=1} \sum_{i\in\mathcal{A}_j} \pi_i^n (\|\Delta c_{ij}^n\| + \|\Delta \Gamma_{ij}^n\| + \|\Delta a_{ij}^n\| + |\Delta b_{ij}^n| + |\Delta \nu_{ij}^n|) \to 0. \tag{20}$$

Next, from the formulation of A_n, by combining all terms of the form $|T_{\alpha_3,\mathbf{0}_d,0}(j)|/\overline{D}(G_n, G_0)$ where $j \in [k_0] : |\mathcal{A}_j| > 1$ and $\alpha_3 \in \{2e_1, 2e_2, \ldots, 2e_d\}$ with $e_u := (0, \ldots, 0, \underbrace{1}_{u\text{-}th}, 0, \ldots, 0)$ being a one-hot vector in $\mathbb{R}^d$ for all $u \in [d]$, we obtain that

$$\frac{1}{\overline{D}(G_n, G_0)} \cdot \sum_{j:|\mathcal{A}_j|>1} \sum_{i\in\mathcal{A}_j} \pi_i^n \|\Delta a_{ij}^n\|^2 \to 0. \tag{21}$$

Putting the results in equations (19), (20) and (21) together with the formulation of $\overline{D}(G_n, G_0)$ in equation (10), we deduce that

$$\frac{\sum_{j:|\mathcal{A}_j|>1} \sum_{i\in\mathcal{A}_j} \pi_i^n (\|\Delta c_{ij}^n\|^{\bar{r}(|\mathcal{A}_j|)} + \|\Delta \Gamma_{ij}^n\|^{\bar{r}(|\mathcal{A}_j|)/2} + |\Delta b_{ij}^n|^{\bar{r}(|\mathcal{A}_j|)} + |\Delta \nu_{ij}^n|^{\bar{r}(|\mathcal{A}_j|)/2})}{\overline{D}(G_n, G_0)} \to 1.$$

As a result, we can find an index $j^* \in [k_0]$ such that $|\mathcal{A}_j| > 1$ and

$$\frac{\sum_{i\in\mathcal{A}_{j^*}} \pi_i^n(\|\Delta c_{ij^*}^n\|^{\bar{r}(\mathcal{A}_{j^*})} + \|\Delta\Gamma_{ij^*}^n\|^{\bar{r}(\mathcal{A}_{j^*})/2} + |\Delta b_{ij^*}^n|^{\bar{r}(\mathcal{A}_{j^*})} + |\Delta\nu_{ij^*}^n|^{\bar{r}(\mathcal{A}_{j^*})/2})}{\overline{D}(G_n, G_0)} \not\to 0. \tag{22}$$

Without loss of generality (WLOG), we may assume that $j^* = 1$. Now, we divide our arguments into two main cases as follows:

Case 1: $\frac{1}{\overline{D}(G_n, G_0)} \cdot \sum_{i\in\mathcal{A}_1} \pi_i^n(\|\Delta c_{i1}^n\|^{\bar{r}(|\mathcal{A}_1|)} + \|\Delta\Gamma_{i1}^n\|^{\bar{r}(|\mathcal{A}_1|)/2}) \not\to 0.$

Here, we continue to split this case into two possibilities:

Case 1.1: $\frac{1}{\overline{D}(G_n, G_0)} \cdot \sum_{i\in\mathcal{A}_1} \pi_i^n\Big(\|\Delta c_{i1}^n\|^{\bar{r}(|\mathcal{A}_1|)} + \|((\Delta\Gamma_{i1}^n)^{(uu)})_{u=1}^d\|^{\bar{r}(|\mathcal{A}_1|)/2}\Big) \not\to 0.$

In this case, it must hold for some index $u^* \in [d]$ that

$$\frac{1}{\overline{D}(G_n, G_0)} \cdot \sum_{i\in\mathcal{A}_1} \pi_i^n\Big(|(\Delta c_{i1}^n)^{(u^*)}|^{\bar{r}(|\mathcal{A}_1|)} + |(\Delta\Gamma_{i1}^n)^{(u^*u^*)}|^{\bar{r}(|\mathcal{A}_1|)/2}\Big) \not\to 0. \tag{23}$$

WLOG, we assume that $u^* = 1$ throughout case 1.1. In the representation of A_n, we consider the following coefficient:

$$T_{\mathbf{0}_d,\tau_1,0}(1) = \sum_{i\in\mathcal{A}_1} \sum_{\substack{\alpha_1,\alpha_2:\\ \tau(\alpha_1,\alpha_2)=\tau_1}} \frac{\pi_i^n}{2^{|\alpha_2|}\alpha_1!\alpha_2!}(\Delta c_{i1}^n)^{\alpha_1}(\Delta\Gamma_{i1}^n)^{\alpha_2}, \tag{24}$$

where $\tau_1 \in \mathbb{N}^d$ such that $\tau_1^{(u)} = 0$ for all $u = 2, \ldots, d$. Thus, the constraint $\tau(\alpha_1, \alpha_2) = \tau_1$ holds if and only if $\alpha_1^{(u)} = \alpha_2^{(u1)} = \alpha_2^{(1v)} = \alpha_2^{(uv)} = 0$ for all $u, v = 2, \ldots, d$. Therefore, by assumption, we have

$$\frac{T_{\mathbf{0}_d,\tau_1,0}(1)}{\overline{D}(G_n, G_0)} = \frac{1}{\overline{D}(G_n, G_0)} \cdot \sum_{i\in\mathcal{A}_1} \sum_{\alpha_1^{(1)}+2\alpha_2^{(11)}=\tau_1^{(1)}} \frac{\pi_i^n}{2^{\alpha_2^{(11)}}\alpha_1^{(1)}!\ \alpha_2^{(11)}!}(\Delta c_{i1}^n)^{\alpha_1^{(1)}}(\Delta\Gamma_{i1}^n)^{\alpha_2^{(11)}} \to 0. \tag{25}$$

Collect results in equations (23) and (25), we obtain that

$$\frac{\sum_{i\in\mathcal{A}_1}\sum_{\alpha_1^{(1)}+2\alpha_2^{(11)}=\tau_1^{(1)}} \frac{\pi_i^n}{2^{\alpha_2^{(11)}}\alpha_1^{(1)}!\ \alpha_2^{(11)}!}(\Delta c_{i1}^n)^{\alpha_1^{(1)}}(\Delta\Gamma_{i1}^n)^{\alpha_2^{(11)}}}{\sum_{i\in\mathcal{A}_1}\pi_i^n\Big(|(\Delta c_{i1}^n)^{(1)}|^{\bar{r}(|\mathcal{A}_1|)} + |(\Delta\Gamma_{i1}^n)^{(11)}|^{\bar{r}(|\mathcal{A}_1|)/2}\Big)} \to 0. \tag{26}$$

Next, we define $\overline{M}_n = \max\{|(\Delta c_{i1}^n)^{(1)}|, |(\Delta\Gamma_{i1}^n)^{(11)}|^{1/2} : i \in \mathcal{A}_1\}$ and $\overline{\pi}_n = \max_{i\in\mathcal{A}_1}\pi_i^n$. For any $i \in \mathcal{A}_1$, it is clear that the sequence of positive real numbers $(\pi_i^n/\overline{\pi}_n)$ is bounded, therefore, we can replace it by its subsequence that admits a non-negative limit denoted by $p_i^2 = \lim_{n\to\infty}\pi_i^n/\overline{\pi}_n$. In addition, let us denote $(\Delta c_{i1}^n)^{(1)}/\overline{M}_n \to \eta_i$ and $(\Delta\Gamma_{i1}^n)^{(11)}/2\overline{M}_n^2 \to \gamma_i$. From the formulation of $\mathcal{O}_{k,\beta}(\Theta)$, since $\pi_i^n \geq \beta$, the real numbers p_i will not vanish, and at least one of them is equal to 1. Analogously, at least one of the η_i and γ_i is equal to either 1 or -1.

Note that $\sum_{i\in\mathcal{A}_1}\pi_i^n\Big(|(\Delta c_{i1}^n)^{(1)}|^{\bar{r}(|\mathcal{A}_1|)} + |(\Delta\Gamma_{i1}^n)^{(11)}|^{\bar{r}(|\mathcal{A}_1|)/2}\Big)/(\overline{\pi}_n\overline{M}_n^{\tau_1^{(1)}}) \not\to 0$ for all $\tau_1^{(1)} \in [\bar{r}(|\mathcal{A}_1|)]$. Thus, we are able to divide both the numerator and the denominator in equation (26) by $\overline{\pi}_n\overline{M}_n^{\tau_1^{(1)}}$ and let $n \to \infty$ in order to achieve the following system of polynomial equations:

$$\sum_{i\in\mathcal{A}_1}\sum_{\alpha_1^{(1)}+2\alpha_2^{(11)}=\tau_1^{(1)}} \frac{p_i^2\eta_i^{\alpha_1^{(1)}}\gamma_i^{\alpha_2^{(11)}}}{\alpha_1^{(1)}!\ \alpha_2^{(11)}!} = 0, \quad \tau_1^{(1)} \in [\bar{r}(|\mathcal{A}_1|)].$$

However, by the definition of $\bar{r}(|\mathcal{A}_1|)$, the above system cannot admit any non-trivial solutions, which is a contradiction. Thus, case 1.1 cannot happen.

Case 1.2: $\frac{1}{\overline{D}(G_n, G_0)} \cdot \sum_{i\in\mathcal{A}_1} \pi_i^n\Big(\|((\Delta\Gamma_{i1}^n)^{(uv)})_{1\leq u\neq v\leq d}\|^{\bar{r}(|\mathcal{A}_1|)/2}\Big) \not\to 0.$

In this case, it must hold for some indices $u^* \neq v^*$ that

$$\frac{1}{\overline{D}(G_n, G_0)} \cdot \sum_{i \in \mathcal{A}_1} \pi_i^n |(\Delta\Gamma_{i1}^n)^{(u^* v^*)}|^{\bar{r}(|\mathcal{A}_1|)/2} \not\to 0.$$

Recall that $|\mathcal{A}_1| > 1$, or equivalently, $|\mathcal{A}_1| \geq 2$, we have that $\bar{r}(|\mathcal{A}_1|) \geq 4$. Therefore, the above equation leads to

$$\frac{1}{\overline{D}(G_n, G_0)} \cdot \sum_{i \in \mathcal{A}_1} \pi_i^n |(\Delta\Gamma_{i1}^n)^{(u^* v^*)}|^2 \not\to 0. \tag{27}$$

WLOG, we assume that $u^* = 1$ and $v^* = 2$ throughout case 1.2. We continue to consider the coefficient $T_{\mathbf{0}_d, \tau_1, 0}$ in equation (24) with $\tau_1 = (2, 2, 0, \ldots, 0) \in \mathbb{N}^d$. By assumption, we have $T_{\mathbf{0}_d, \tau_1, 0}/\overline{D}(G_n, G_0) \to 0$, which together with equation (27) imply that

$$\frac{\sum_{i \in \mathcal{A}_1} \sum_{\substack{\alpha_1, \alpha_2: \\ \tau(\alpha_1, \alpha_2) = \tau_1}} \frac{\pi_i^n}{2^{|\alpha_2|} \alpha_1! \alpha_2!} (\Delta c_{i1}^n)^{\alpha_1} (\Delta\Gamma_{i1}^n)^{\alpha_2}}{\sum_{i \in \mathcal{A}_1} \pi_i^n |(\Delta\Gamma_{i1}^n)^{(12)}|^2} \to 0. \tag{28}$$

Similarly, by combining the fact that case 1.1 does not hold and the result in equation (27), we get

$$\frac{\sum_{i \in \mathcal{A}_1} \pi_i^n \left(\|\Delta c_{i1}^n\|^{\bar{r}(|\mathcal{A}_1|)} + \|((\Delta\Gamma_{i1}^n)^{(uu)})_{u=1}^d\|^{\bar{r}(|\mathcal{A}_1|)/2} \right)}{\sum_{i \in \mathcal{A}_1} \pi_i^n |(\Delta\Gamma_{i1}^n)^{(12)}|^2} \to 0.$$

Since $\bar{r}(|\mathcal{A}_1|) \geq 4$, the above limit indicates that any terms in equation (28) with $\alpha_1^{(u)} > 0$ and $\alpha_2^{(uu)} > 0$ for $u \in \{1, 2\}$ will vanish. Consequently, we deduce from equation (28) that

$$1 = \frac{\sum_{i \in \mathcal{A}_1} \pi_i^n |(\Delta\Gamma_{i1}^n)^{(12)}|^2}{\sum_{i \in \mathcal{A}_1} \pi_i^n |(\Delta\Gamma_{i1}^n)^{(12)}|^2} \to 0,$$

which is a contradiction. Thus, case 1.2 cannot happen.

Case 2: $\frac{1}{\overline{D}(G_n, G_0)} \cdot \sum_{i \in \mathcal{A}_1} \pi_i^n (|\Delta b_{i1}^n|^{\bar{r}(|\mathcal{A}_1|)} + |\Delta \nu_{i1}^n|^{\bar{r}(|\mathcal{A}_1|)/2}) \not\to 0.$

In this case, we consider the coefficient $T_{\mathbf{0}_d, \mathbf{0}_d, 0}(1)$ in the formulation of A_n. By assumption,

$$\frac{T_{\mathbf{0}_d, \mathbf{0}_d, 0}(1)}{\overline{D}(G_n, G_0)} = \frac{1}{\overline{D}(G_n, G_0)} \cdot \sum_{i \in \mathcal{A}_1} \pi_i^n \sum_{\substack{\alpha_4, \alpha_5: \\ \alpha_4 + 2\alpha_5 = \tau_2}} \frac{(\Delta b_{i1}^n)^{\alpha_4} (\Delta \nu_{i1}^n)^{\alpha_5}}{2^{\alpha_5} \alpha_4! \alpha_5!} \to 0.$$

Consequently, we obtain that

$$\frac{\sum_{i \in \mathcal{A}_1} \pi_i^n \sum_{\substack{\alpha_4, \alpha_5: \\ \alpha_4 + 2\alpha_5 = \tau_2}} \frac{(\Delta b_{i1}^n)^{\alpha_4} (\Delta \nu_{i1}^n)^{\alpha_5}}{2^{\alpha_5} \alpha_4! \alpha_5!}}{\sum_{i \in \mathcal{A}_1} \pi_i^n (|\Delta b_{i1}^n|^{\bar{r}(|\mathcal{A}_1|)} + |\Delta \nu_{i1}^n|^{\bar{r}(|\mathcal{A}_1|)/2})} \to 0.$$

By employing the same arguments for showing that the equation (26) does not hold in case 1.1, we obtain that the above limit does not hold, either. Thus, case 2 cannot happen.

From the above results of the two main cases, we conclude that not all the coefficients in the representations of $A_n/\overline{D}(G_n, G_0)$, $B_n/\overline{D}(G_n, G_0)$ and $E_n/\overline{D}(G_n, G_0)$ vanish as $n \to \infty$.

Step 3 - Application of Fatou's lemma: Subsequently, we denote by m_n the maximum of the absolute values of the coefficients in the representations of $A_n/\overline{D}(G_n, G_0)$, $B_n/\overline{D}(G_n, G_0)$ and $E_n/\overline{D}(G_n, G_0)$, that is,

$$m_n := \max_{(\alpha_3, \tau_1, \tau_2, j) \in \mathcal{S}} |T_{\alpha_3, \tau_1, \tau_2}^n(j)|/\overline{D}(G_n, G_0),$$

where the constraint set $\mathcal{S}$ is defined as

$$\mathcal{S} := \Big\{ (\alpha_3, \tau_1, \tau_2, j) \in \mathbb{N}^d \times \mathbb{N}^d \times \mathbb{N} \times [k_0] : 0 \leq |\alpha_3| \leq \bar{r}(|\mathcal{A}_j|), 0 \leq |\tau_1|, \tau_2 \leq 2(\bar{r}(|\mathcal{A}_j|) - |\alpha_3|) \Big\}.$$

Additionally, we define $T^n_{\alpha_3,\tau_1,\tau_2}(j)/m_n \to \xi_{\alpha_3,\tau_1,\tau_2}(j)$ as $n \to \infty$ for all $(\alpha_3, \tau_1, \tau_2, j) \in \mathcal{S}$. Since not all the coefficients in the representations of $A_n/\overline{D}(G_n, G_0)$, $B_n/\overline{D}(G_n, G_0)$ and $E_n/\overline{D}(G_n, G_0)$ vanish as $n \to \infty$, at least one among $\xi_{\alpha_3,\tau_1,\tau_2}(j)$ is different from zero and $m_n \not\to 0$. Then, by applying the Fatou's lemma, we get that

$$0 = \lim_{n\to\infty} \frac{1}{m_n} \cdot \frac{2V(p_{G_n}, p_G)}{\overline{D}(G_n, G_0)} \geq \int \liminf_{n\to\infty} \frac{1}{m_n} \cdot \frac{|p_{G_n}(X,Y) - p_G(X,Y)|}{\overline{D}(G_n, G_0)} \, \mathrm{d}(X,Y) \geq 0.$$

Moreover, by definition, we have

$$\begin{aligned}
&\frac{1}{m_n} \cdot \frac{p_{G_n}(X,Y) - p_G(X,Y)}{\overline{D}(G_n, G_0)} \\
&\quad \to \sum_{(\alpha_3,\tau_1,\tau_2,j)\in\mathcal{S}} \xi_{\alpha_3,\tau_1,\tau_2}(j) X^{\alpha_3} \frac{\partial^{|\tau_1|} f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c^0_j, \Gamma^0_j) \frac{\partial^{|\alpha_3|+\tau_2} f_{\mathcal{D}}}{\partial h_1^{|\alpha_3|+\tau_2}}(Y|(a^0_j)^\top X + b^0_j, \nu^0_j).
\end{aligned}$$

As a consequence, we achieve that

$$\sum_{(\alpha_3,\tau_1,\tau_2,j)\in\mathcal{S}} \xi_{\alpha_3,\tau_1,\tau_2}(j) X^{\alpha_3} \frac{\partial^{|\tau_1|} f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c^0_j, \Gamma^0_j) \frac{\partial^{|\alpha_3|+\tau_2} f_{\mathcal{D}}}{\partial h_1^{|\alpha_3|+\tau_2}}(Y|(a^0_j)^\top X + b^0_j, \nu^0_j) = 0,$$

for almost surely (X, Y). Since elements of the set $\mathcal{F}$ defined in equation (18) are linearly independent (proof of this claim is deferred to the end of this proof), the above equation implies that $\xi_{\alpha_3,\tau_1,\tau_2}(j) = 0$ for all $(\alpha_3, \tau_1, \tau_2, j) \in \mathcal{S}$, which contradicts the fact that at least one among $\xi_{\alpha_3,\tau_1,\tau_2}(j)$ is different from zero. Hence, we reach the conclusion in equation (16), which indicates that there exists some $\varepsilon_0 > 0$ such that

$$\inf_{\substack{G\in\mathcal{O}_{k,\beta}(\Theta) \\ \overline{D}(G,G_0)\leq\varepsilon_0}} V(p_G, p_{G_0})/\overline{D}(G, G_0) > 0.$$

Global bound: Given the above result, in order to achieve the inequality in equation (15), we only need to prove its following global version:

$$\inf_{\substack{G\in\mathcal{O}_{k,\beta}(\Theta) \\ \overline{D}(G,G_0)>\varepsilon_0}} V(p_G, p_{G_0})/\overline{D}(G, G_0) > 0.$$

Assume by contrary that the above claim is not true. Then, there exists a sequence $G'_n \in \mathcal{O}_{k,\beta}(\Theta)$ such that $V(p_{G'_n}, p_{G_0})/\overline{D}(G'_n, G_0) \to 0$ and $\overline{D}(G'_n, G_0) > \varepsilon_0$ for all $n \in \mathbb{N}$. Since the set Θ is compact, we can replace G'_n by its subsequence that converges to some mixing measure $G' \in \mathcal{O}_{k,\beta}(\Theta)$. Consequently, we deduce that $\overline{D}(G', G_0) = \lim_{n\to\infty} \overline{D}(G'_n, G_0) \geq \varepsilon_0$. This result together with the fact that $V(p_{G'_n}, p_{G_0})/\overline{D}(G'_n, G_0) \to 0$ lead to the limit $V(p_{G'_n}, p_{G_0}) \to 0$ as $n \to \infty$. Again, by applying the Fatou's lemma, we obtain that

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

As a consequence, we have that $p_{G'}(X,Y) = p_{G_0}(X,Y)$ for almost surely (X,Y). Due to the identifiability of the model, this equality leads to $G' \equiv G_0$, which contradicts the bound $\overline{D}(G', G_0) \geq \varepsilon_0 > 0$. Hence, we achieve the conclusion in equation (15).

Linear independence of elements in $\mathcal{F}$: For completion, we will demonstrate elements of the set $\mathcal{F}$ defined in equation (18) are linearly independent by definition. In particular, assume that there exist real numbers $\xi_{\alpha_3,\tau_1,\tau_2}(j)$, where $(\alpha_3, \tau_1, \tau_2, j) \in \mathcal{S}$, such that the following equation holds for almost surely (X, Y):

$$\sum_{(\alpha_3,\tau_1,\tau_2,j)\in\mathcal{S}} \xi_{\alpha_3,\tau_1,\tau_2}(j) X^{\alpha_3} \frac{\partial^{|\tau_1|} f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c^0_j, \Gamma^0_j) \frac{\partial^{|\alpha_3|+\tau_2} f_{\mathcal{D}}}{\partial h_1^{|\alpha_3|+\tau_2}}(Y|(a^0_j)^\top X + b^0_j, \nu^0_j) = 0.$$

Now, we rewrite the above equation as follows:

$$\sum_{j=1}^{k_0}\sum_{\omega=0}^{2\bar{r}(|\mathcal{A}_j|)}\Big(\sum_{|\alpha_3|+\tau_2=\omega}\sum_{|\tau_1|=0}^{2(\bar{r}(|\mathcal{A}_j|)-\alpha_3)-\tau_2}\xi_{\alpha_3,\tau_1,\tau_2}(j)X^{\alpha_3}\cdot\frac{\partial^{|\tau_1|}f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_j^0,\Gamma_j^0)\Big)\times\frac{\partial^{\omega}f_{\mathcal{D}}}{\partial h_1^{\omega}}(Y|(a_j^0)^\top X+b_j^0,\nu_j^0)=0, \tag{29}$$

for almost surely (X,Y). As (a_j^0,b_j^0,ν_j^0) for $j\in[k_0]$ are k_0 distinct tuples, we deduce that $((a_j^0)^\top X+b_j^0,\nu_j^0)$ for $j\in[k_0]$ are also k_0 distinct tuples for almost surely X. Thus, for almost surely X, one has $\frac{\partial^{\omega}f_{\mathcal{D}}}{\partial h_1^{\omega}}(Y|(a_j^0)^\top X+b_j^0,\nu_j^0)$ for $j\in[k_0]$ and $0\leq\omega\leq 2\bar{r}(|\mathcal{A}_j|)$ are linearly independent with respect to Y. Given that result, the equation (29) indicates that for almost surely X,

$$\sum_{|\alpha_3|+\tau_2=\omega}\sum_{|\tau_1|=0}^{2(\bar{r}(|\mathcal{A}_j|)-\omega)}\xi_{\alpha_3,\tau_1,\tau_2}(j)X^{\alpha_3}\cdot\frac{\partial^{|\tau_1|}f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_j^0,\Gamma_j^0)=0,$$

for all $j\in[k_0]$ and $0\leq\omega\leq 2\bar{r}(|\mathcal{A}_j|)$. Note that for each $j\in[k_0]$ and $0\leq\omega\leq\bar{r}(|\mathcal{A}_j|)$, the left hand side of the above equation can be viewed as a high-dimensional polynomial of two random vectors X and $X-c_j^0$ ($c_j^0\neq\mathbf{0}_d$) in $\mathcal{X}$, which is a compact set in $\mathbb{R}^d$. As a result, the above equation holds when $\xi_{\alpha_3,\tau_1,\tau_2}(j)=0$ for all $j\in[k_0]$, $0\leq\omega\leq 2\bar{r}(|\mathcal{A}_j|)$, $|\alpha_3|+\tau_2=\omega$ and $|\tau_1|\leq 2(\bar{r}(|\mathcal{A}_j|)-\alpha_3)-\tau_2$. This is equivalent to $\xi_{\alpha_3,\tau_1,\tau_2}(j)=0$ for all $(\alpha_3,\tau_1,\tau_2,j)\in\mathcal{S}$.

Hence, we conclude that the elements of $\mathcal{F}$ are linearly independent.

B.2 Proof of Theorem 2

In order to reach the conclusion in Theorem 2, we only need to demonstrate the following inequality:

$$\inf_{G\in\mathcal{O}_{k,\beta}(\Theta)}V(p_G,p_{G_0})/\widetilde{D}(G,G_0)>0. \tag{30}$$

In this proof, we will only prove the following local version of inequality (30) while the global version can be argued in the same fashion as in Appendix (B.1):

$$\lim_{\varepsilon\to 0}\inf_{\substack{G\in\mathcal{O}_{k,\beta}(\Theta)\\ \widetilde{D}(G,G_0)\leq\varepsilon}}V(p_G,p_{G_0})/\widetilde{D}(G,G_0)>0. \tag{31}$$

Assume that the claim in equation (31) is not true. This indicates that we can find a sequence of mixing measures $G_n=\sum_{i=1}^{k_n}\pi_i^n\delta_{(c_i,\Gamma_i^n,a_i^n,b_i^n,\nu_i^n)}\in\mathcal{O}_{k,\beta}(\Theta)$ that satisfies: $\widetilde{D}(G_n,G_0)\to 0$ and $V(p_{G_n},p_{G_0})/\widetilde{D}(G_n,G_0)\to 0$ as $n\to\infty$. Additionally, since $k_n\leq k$ for all $n\in\mathbb{N}$, we are able to replace (G_n) by its subsequence which admits a fixed number of atoms $k_n\leq k'\leq k$ and $\mathcal{A}_j=\mathcal{A}_j^n$ is independent of n for all $j\in[k_0]$.

Step 1 - Taylor expansion for density decomposition: Next, we take into account the quantity

$$\begin{aligned}
&p_{G_n}(X,Y)-p_{G_0}(X,Y)\\
&=\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\pi_i^n[f_{\mathcal{L}}(X|c_i^n,\Gamma_i^n)f_{\mathcal{D}}(Y|a_i^nX+b_i^n,\nu_i^n)-f_{\mathcal{L}}(X|c_j^0,\Gamma_j^0)f_{\mathcal{D}}(Y|a_j^0X+b_j^0,\nu_j^0)]\\
&+\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\pi_i^n[f_{\mathcal{L}}(X|c_i^n,\Gamma_i^n)f_{\mathcal{D}}(Y|a_i^nX+b_i^n,\nu_i^n)-f_{\mathcal{L}}(X|c_j^0,\Gamma_j^0)f_{\mathcal{D}}(Y|a_j^0X+b_j^0,\nu_j^0)]\\
&+\sum_{j=1}^{k_0}\left(\sum_{i\in\mathcal{A}_j}\pi_i^n-\pi_j^0\right)f_{\mathcal{L}}(X|c_j^0,\Gamma_j^0)f_{\mathcal{D}}(Y|a_j^0X+b_j^0,\nu_j^0)\\
&:=A_n+B_n+E_n.
\end{aligned}$$

For each $j \in [k_0] : |\mathcal{A}_j| > 1$, by means of Taylor expansion up to the $\widetilde{r}(|\mathcal{A}_j|)$-th order, A_n can be rewritten as follows with a note that $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) \in \mathbb{N}^5$:

$$\begin{aligned}
A_n &= \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \pi_i^n \sum_{|\alpha|=1}^{\widetilde{r}(|\mathcal{A}_j|)} \frac{1}{\alpha!} (\Delta c_{ij}^n)^{\alpha_1} (\Delta \Gamma_{ij}^n)^{\alpha_2} (\Delta a_{ij}^n)^{\alpha_3} (\Delta b_{ij}^n)^{\alpha_4} (\Delta \nu_{ij}^n)^{\alpha_5} \\
&\quad \times \frac{\partial^{\alpha_1+\alpha_2} f_{\mathcal{L}}}{\partial c^{\alpha_1} \partial \Gamma^{\alpha_2}}(X|c_j^0, \Gamma_j^0) \cdot \frac{\partial^{\alpha_3+\alpha_4+\alpha_5} f}{\partial a^{\alpha_3} \partial b^{\alpha_4} \partial \nu^{\alpha_5}}(Y|a_j^0 X + b_j^0, \nu_j^0) + R_1(X, Y) \\
&= \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \pi_i^n \sum_{|\alpha|=1}^{\widetilde{r}(|\mathcal{A}_j|)} \frac{1}{\alpha!} (\Delta c_{ij}^n)^{\alpha_1} (\Delta \Gamma_{ij}^n)^{\alpha_2} (\Delta a_{ij}^n)^{\alpha_3} (\Delta b_{ij}^n)^{\alpha_4} (\Delta \nu_{ij}^n)^{\alpha_5} \\
&\quad \times \frac{1}{2^{\alpha_2}} \frac{\partial^{\alpha_1+2\alpha_2} f_{\mathcal{L}}}{\partial c^{\alpha_1+2\alpha_2}}(X|c_j^0, \Gamma_j^0) \cdot \frac{X^{\alpha_3}}{2^{\alpha_5}} \frac{\partial^{\alpha_3+\alpha_4+2\alpha_5} f_{\mathcal{D}}}{\partial h_1^{\alpha_3+\alpha_4+2\alpha_5}}(Y|a_j^0 X + b_j^0, \nu_j^0) + R_3(X, Y),
\end{aligned}$$

where $R_3(X, Y)$ is Taylor remainder such that $R_3(X, Y)/\widetilde{D}(G_n, G_0) \to 0$. Since c_j^0 is equal to zero when $j \in [\widetilde{k}]$ and different from zero otherwise, the formulation of $\dfrac{\partial^{\alpha_1+2\alpha_2} f_{\mathcal{L}}}{\partial c^{\alpha_1+2\alpha_2}}(X|c_j^0, \Gamma_j^0)$ will vary when $j \in [\widetilde{k}]$ compared to $\widetilde{k} + 1 \leq j \leq k_0$. Thus, we will consider these two cases of j separately.

For $j \in [\widetilde{k}]$, when α_1 is an even integer, we have

$$\frac{\partial^{\alpha_1+2\alpha_2} f_{\mathcal{L}}}{\partial c^{\alpha_1+2\alpha_2}}(X|c_j^0, \Gamma_j^0) = \begin{cases} \sum_{w=0}^{\alpha_1/2+\alpha_2} t_{2w,\alpha_1+2\alpha_2} X^{2w}, & j \in [\widetilde{k}] \\ \sum_{w=0}^{\alpha_1/2+\alpha_2} s_{2w,\alpha_1+2\alpha_2} (X - c_j^0)^{2w}, & \widetilde{k} + 1 \leq j \leq k_0. \end{cases}$$

On the other hand, when α_1 is an odd integer, we get

$$\frac{\partial^{\alpha_1+2\alpha_2} f_{\mathcal{L}}}{\partial c^{\alpha_1+2\alpha_2}}(X|c_j^0, \Gamma_j^0) = \begin{cases} \sum_{w=0}^{(\alpha_1-1)/2+\alpha_2} t_{2w+1,\alpha_1+2\alpha_2} X^{2w+1}, & j \in [\widetilde{k}] \\ \sum_{w=0}^{(\alpha_1-1)/2+\alpha_2} s_{2w+1,\alpha_1+2\alpha_2} (X - c_j^0)^{2w+1}, & \widetilde{k} + 1 \leq j \leq k_0. \end{cases}$$

By combining both cases, we rewrite A_n as follows:

$$\begin{aligned}
A_n &= \sum_{\substack{j:|\mathcal{A}_j|>1, \\ j \in [\widetilde{k}]}} \sum_{i \in \mathcal{A}_j} \sum_{\ell_1+\ell_2=1}^{2\widetilde{r}(|\mathcal{A}_j|)} \sum_{\alpha \in \mathcal{I}_{\ell_1,\ell_2}} \frac{\pi_i^n}{2^{\alpha_2+\alpha_5} \alpha!} (\Delta c_{ij}^n)^{\alpha_1} (\Delta \Gamma_{ij}^n)^{\alpha_2} (\Delta a_{ij}^n)^{\alpha_3} \\
&\quad \times (\Delta b_{ij}^n)^{\alpha_4} (\Delta \nu_{ij}^n)^{\alpha_5} \; t_{\ell_1-\alpha_3,\alpha_1+2\alpha_2} X^{\ell_1} \frac{\partial^{\ell_2} f_{\mathcal{D}}}{\partial h_1^{\ell_2}}(Y|a_j^0 X + b_j^0, \nu_j^0) f_{\mathcal{L}}(X|c_j^0, \Gamma_j^0) \\
&+ \sum_{\substack{j:|\mathcal{A}_j|>1 \\ \widetilde{k}+1 \leq j \leq k_0}} \sum_{i \in \mathcal{A}_j} \sum_{\alpha_3=0}^{\widetilde{r}(|\mathcal{A}_j|)} \sum_{\tau_1+\tau_2=0}^{2(\widetilde{r}(|\mathcal{A}_j|)-\alpha_3)} \sum_{\substack{\alpha_1+2\alpha_2=\tau_1 \\ \alpha_4+2\alpha_5=\tau_2}} \frac{\pi_i^n}{2^{\alpha_2+\alpha_5} \alpha!} (\Delta c_{ij}^n)^{\alpha_1} (\Delta \Gamma_{ij}^n)^{\alpha_2} (\Delta a_{ij}^n)^{\alpha_3} \\
&\quad \times (\Delta b_{ij}^n)^{\alpha_4} (\Delta \nu_{ij}^n)^{\alpha_5} \times X^{\alpha_3} \frac{\partial^{\tau_1} f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_j^0, \Gamma_j^0) \frac{\partial^{\alpha_3+\tau_2} f_{\mathcal{D}}}{\partial h_1^{\alpha_3+\tau_2}}(Y|a_j^0 X + b_j^0, \nu_j^0) + R_3(X, Y),
\end{aligned} \tag{32}$$

where for any $0 \leq \ell_1 \leq 2\bar{r}(|\mathcal{A}_j|)$ and $0 \leq \ell_2 \leq 2\bar{r}(|\mathcal{A}_j|) - \ell_1$, we define

$$\begin{aligned}
\mathcal{I}_{\ell_1,\ell_2} := \Big\{ \alpha = (\alpha_i)_{i=1}^5 \in \mathbb{N}^5 : \alpha_1 + 2\alpha_2 + \alpha_3 \geq \ell_1, \; \alpha_3 + \alpha_4 + 2\alpha_5 = \ell_2, \\
1 \leq \alpha_1 + \alpha_2 + \ldots + \alpha_5 \leq \widetilde{r}(|\mathcal{A}_j|) \Big\}.
\end{aligned}$$

Regarding the formulation of B_n, for each $j \in [k_0] : |\mathcal{A}_j| = 1$, we perform a Taylor expansion up to the first order

and obtain that

$$B_n = \sum_{\substack{j:|\mathcal{A}_j|=1,\\ j\in[\widetilde{k}]}} \sum_{i\in\mathcal{A}_j} \sum_{\ell_1+\ell_2=1}^{2} \sum_{\alpha\in\mathcal{I}_{\ell_1,\ell_2}} \frac{\pi_i^n}{2^{\alpha_2+\alpha_5}\alpha!}(\Delta c_{ij}^n)^{\alpha_1}(\Delta\Gamma_{ij}^n)^{\alpha_2}(\Delta a_{ij}^n)^{\alpha_3}$$
$$\times (\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5}\, t_{\ell_1-\alpha_3,\alpha_1+2\alpha_2} X^{\ell_1}\frac{\partial^{\ell_2} f_{\mathcal{D}}}{\partial h_1^{\ell_2}}(Y|a_j^0 X + b_j^0,\nu_j^0) f_{\mathcal{L}}(X|c_j^0,\Gamma_j^0)$$
$$+ \sum_{\substack{j:|\mathcal{A}_j|=1\\ \widetilde{k}+1\le j\le k_0}} \sum_{i\in\mathcal{A}_j} \sum_{\alpha_3=0}^{2} \sum_{\tau_1+\tau_2=0}^{2(1-\alpha_3)} \sum_{\substack{\alpha_1,\alpha_2:\\ \alpha_1+2\alpha_2=\tau_1}} \sum_{\substack{\alpha_4,\alpha_5:\\ \alpha_4+2\alpha_5=\tau_2}} \frac{\pi_i^n}{2^{\alpha_2+\alpha_5}\alpha!}(\Delta c_{ij}^n)^{\alpha_1}(\Delta\Gamma_{ij}^n)^{\alpha_2}(\Delta a_{ij}^n)^{\alpha_3}$$
$$\times (\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5} \times X^{\alpha_3}\frac{\partial^{\tau_1} f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_j^0,\Gamma_j^0)\frac{\partial^{\alpha_3+\tau_2} f_{\mathcal{D}}}{\partial h_1^{\alpha_3+\tau_2}}(Y|a_j^0 X + b_j^0,\nu_j^0) + R_4(X,Y), \tag{33}$$

where $R_4(X,Y)/\widetilde{D}(G_n,G_0) \to 0$ as $n\to\infty$.

From equations (32) and (33), we can treat $A_n/\widetilde{D}(G_n,G_0)$, $B_n/\widetilde{D}(G_n,G_0)$ and $E_n/\widetilde{D}(G_n,G_0)$ as linear combinations of elements of the following set:

$$\mathcal{H} := \Big\{X^{\ell_1}\frac{\partial^{\ell_2} f_{\mathcal{D}}}{\partial h_1^{\ell_2}}(Y|a_j^0 X + b_j^0,\nu_j^0) f_{\mathcal{L}}(X|c_j^0,\Gamma_j^0) : j\in[\widetilde{k}],\ 0\le \ell_1+\ell_2\le 2\widetilde{r}(|\mathcal{A}_j|)\Big\}$$
$$\cup \Big\{X^{\alpha_3}\frac{\partial^{\tau_1} f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_j^0,\Gamma_j^0)\frac{\partial^{\alpha_3+\tau_2} f_{\mathcal{D}}}{\partial h_1^{\alpha_3+\tau_2}}(Y|a_j^0 X + b_j^0,\nu_j^0) : \widetilde{k}+1\le j\le k_0,\ 0\le\alpha_3\le\widetilde{r}(|\mathcal{A}_j|),$$
$$0\le\tau_1+\tau_2\le 2(\widetilde{r}(|\mathcal{A}_j|)-\alpha_3)\Big\}. \tag{34}$$

For any $(j,\ell_1,\ell_2)\in\mathcal{Q} := \{(j,\ell_1,\ell_2)\in\mathbb{N}^3 : j\in[\widetilde{k}],\ 0\le\ell_1+\ell_2\le 2\widetilde{r}(|\mathcal{A}_j|)\}$, let $Q_{\ell_1,\ell_2}^n(j)$ be the coefficient of

$$X^{\ell_1}\frac{\partial^{\ell_2} f_{\mathcal{D}}}{\partial h_1^{\ell_2}}(Y|a_j^0 X + b_j^0,\nu_j^0) f_{\mathcal{L}}(X|c_j^0,\Gamma_j^0)$$

in the representations of A_n, B_n and E_n. It follows from equations (32) and (33) that $Q_{\ell_1,\ell_2}^n(j)$ is given by

$$Q_{\ell_1,\ell_2}^n(j) = \begin{cases} \sum_{\alpha\in\mathcal{I}_{\ell_1,\ell_2}}\sum_{i\in\mathcal{A}_j} \frac{\pi_i^n\cdot t_{\ell_1-\alpha_3,\alpha_1+2\alpha_2}}{2^{\alpha_2+\alpha_5}\alpha!}(\Delta c_{ij}^n)^{\alpha_1}(\Delta\Gamma_{ij}^n)^{\alpha_2}(\Delta a_{ij}^n)^{\alpha_3}(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5} & \\ \qquad\qquad (\ell_1,\ell_2)\neq(0,0), & \\ \sum_{i\in\mathcal{A}_j}\pi_i^n - \pi_j^0, \qquad\qquad (\ell_1,\ell_2)=(0,0). & \end{cases}$$

Meanwhile, we denote by $T_{\alpha_3,\tau_1,\tau_2}^n(j)$ the coefficient of

$$X^{\alpha_3}\frac{\partial^{\tau_1} f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_j^0,\Gamma_j^0)\frac{\partial^{\alpha_3+\tau_2} f_{\mathcal{D}}}{\partial h_1^{\alpha_3+\tau_2}}(Y|a_j^0 X + b_j^0,\nu_j^0),$$

for all $(j,\alpha_3,\tau_1,\tau_2)\in\mathcal{T} := \{(j,\alpha_3,\tau_1,\tau_2)\in\mathbb{N}^3 : \widetilde{k}+1\le j\le k_0,\ 0\le\alpha_3\le\widetilde{r}(|\mathcal{A}_j|),\ 0\le\tau_1+\tau_2\le 2(\widetilde{r}(|\mathcal{A}_j|)-\alpha_3)\}$. Thus, $T_{\alpha_3,\tau_1,\tau_2}^n(j)$ is represented as

$$T_{\alpha_3,\tau_1,\tau_2}^n(j) = \begin{cases} \sum_{\substack{\alpha_1+2\alpha_2=\tau_1,\\ \alpha_4+2\alpha_5=\tau_2}}\sum_{i\in\mathcal{A}_j} \frac{\pi_i^n}{2^{\alpha_2+\alpha_5}\alpha!}(\Delta c_{ij}^n)^{\alpha_1}(\Delta\Gamma_{ij}^n)^{\alpha_2}(\Delta a_{ij}^n)^{\alpha_3}(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5} & \\ \qquad\qquad (\alpha_3,\tau_1,\tau_2)\neq(0,0,0), & \\ \sum_{i\in\mathcal{A}_j}\pi_i^n - \pi_j^0, \qquad\qquad (\alpha_3,\tau_1,\tau_2)=(0,0,0). & \end{cases}$$

Step 2 - Proof of non-vanishing coefficients by contradiction: Assume by contrary that all the coefficients of elements in the set $\mathcal{H}$ in the representations of $A_n/\widetilde{D}(G_n,G_0)$, $B_n/\widetilde{D}(G_n,G_0)$ and $E_n/\widetilde{D}(G_n,G_0)$ vanish when

n tends to infinity. It is worth noting that for $(\ell_1, \ell_2) \neq (0,0)$, we have $\mathcal{I}_{\ell_1+1,\ell_2} \subseteq \mathcal{I}_{\ell_1,\ell_2}$ and

$$\mathcal{J}_{\ell_1,\ell_2} := \mathcal{I}_{\ell_1,\ell_2} \setminus \mathcal{I}_{\ell_1+1,\ell_2} = \Big\{(\alpha_1,\ldots,\alpha_5) \in \mathbb{N}^5 : \alpha_1 + 2\alpha_2 + \alpha_3 = \ell_1,\ \alpha_3 + \alpha_4 + 2\alpha_5 = \ell_2, \\ 1 \leq \alpha_1 + \alpha_2 + \ldots + \alpha_5 \leq \widetilde{r}(|\mathcal{A}_j|)\Big\}.$$

Since $Q^n_{\ell_1,\ell_2}(j)/\widetilde{D}(G_n, G_0) \to 0$ for all tuples $(j, \ell_1, \ell_2) \in \mathcal{Q}$, we achieve that

$$\begin{aligned}
\frac{S^n_{\ell_1,\ell_2}(j)}{\widetilde{D}(G_n,G_0)} &:= \frac{Q^n_{\ell_1,\ell_2}(j) - Q_{\ell_1+1,\ell_2}(j)}{\widetilde{D}(G_n,G_0)} \\
&= \frac{\sum_{\alpha\in\mathcal{J}_{\ell_1,\ell_2}}\sum_{i\in\mathcal{A}_j} \frac{\pi^n_i \cdot t_{\ell_1-\alpha_3,\alpha_1+2\alpha_2}}{2^{\alpha_2+\alpha_5}\alpha!}(\Delta c^n_{ij})^{\alpha_1}(\Delta\Gamma^n_{ij})^{\alpha_2}(\Delta a^n_{ij})^{\alpha_3}(\Delta b^n_{ij})^{\alpha_4}(\Delta\nu^n_{ij})^{\alpha_5}}{\widetilde{D}(G_n,G_0)} \\
&= \frac{\sum_{\alpha\in\mathcal{J}_{\ell_1,\ell_2}}\sum_{i\in\mathcal{A}_j} \frac{\pi^n_i}{(\Gamma^0_j)^{\alpha_1+2\alpha_2}\ 2^{\alpha_2+\alpha_5}\alpha!}(\Delta c^n_{ij})^{\alpha_1}(\Delta\Gamma^n_{ij})^{\alpha_2}(\Delta a^n_{ij})^{\alpha_3}(\Delta b^n_{ij})^{\alpha_4}(\Delta\nu^n_{ij})^{\alpha_5}}{\widetilde{D}(G_n,G_0)} \\
&\to 0,
\end{aligned}$$

where the third inequality follows from the fact that $t_{\ell_1-\alpha_3,\alpha_1+2\alpha_2} = t_{\alpha_1+2\alpha_2,\alpha_1+2\alpha_2} = (\Gamma^0_j)^{-(\alpha_1+2\alpha_2)}$. Additionally, we also let $S_{0,0}(j) := Q_{0,0}(j)$ for all $j \in [\widetilde{k}]$.

By assumption, $|S_{0,0}(j)|/\widetilde{D}(G_n,G_0) \to 0$ for all $j \in [\widetilde{k}]$ and $|T_{0,0,0}(j)|/\widetilde{D}(G_n,G_0) \to 0$ for all $\widetilde{k}+1 \leq j \leq k_0$ as $n \to \infty$. By taking the summation of all such terms, we get that

$$\frac{1}{\widetilde{D}(G_n,G_0)} \cdot \sum_{j=1}^{k_0}\Big|\sum_{i\in\mathcal{A}_j}\pi^n_i - \pi^0_j\Big| \to 0. \tag{35}$$

Next, we consider indices $j \in [k_0] : |\mathcal{A}_j| = 1$, i.e. those in the formulation of B_n. For $j \in [\widetilde{k}]$, since $|S^n_{\ell_1,\ell_2}(j)|/\widetilde{D}(G_n,G_0) \to 0$ for all $(\ell_1,\ell_2) \in \{(1,0),(0,1),(1,1),(2,0),(0,2)\}$, we get that

$$\frac{1}{\widetilde{D}(G_n,G_0)} \cdot \sum_{\substack{j:|\mathcal{A}_j|=1 \\ j\in[\widetilde{k}]}}\sum_{i\in\mathcal{A}_j}\pi^n_i\Big(|\Delta c^n_{ij}| + |\Delta\Gamma^n_{ij}| + |\Delta a^n_{ij}| + |\Delta b^n_{ij}| + |\Delta\nu^n_{ij}|\Big) \to 0. \tag{36}$$

Moreover, for $\widetilde{k}+1 \leq j \leq k_0$, as $|T^n_{\alpha_3,\tau_1,\tau_2}(j)|/\widetilde{D}(G_n,G_0) \to 0$ for all $(\alpha_3,\tau_1,\tau_2) \in \{(0,1,0),(0,2,0),(1,0,0),(0,0,1),(0,0,2)\}$, we deduce that

$$\frac{1}{\widetilde{D}(G_n,G_0)} \cdot \sum_{\substack{j:|\mathcal{A}_j|=1 \\ \widetilde{k}+1\leq j\leq k_0}}\sum_{i\in\mathcal{A}_j}\pi^n_i\Big(|\Delta c^n_{ij}| + |\Delta\Gamma^n_{ij}| + |\Delta a^n_{ij}| + |\Delta b^n_{ij}| + |\Delta\nu^n_{ij}|\Big) \to 0. \tag{37}$$

Let us denote

$$K^n_{ij}(\kappa_1,\kappa_2,\kappa_3,\kappa_4,\kappa_5) := |\Delta c^n_{ij}|^{\kappa_1} + |\Delta\Gamma^n_{ij}|^{\kappa_2} + |\Delta a^n_{ij}|^{\kappa_3} + |\Delta b^n_{ij}|^{\kappa_4} + |\Delta\nu^n_{ij}|^{\kappa_5}.$$

Then, equations (36) and (37) indicates that

$$\frac{1}{\widetilde{D}(G_n,G_0)} \cdot \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\pi^n_i K^n_{ij}(1,1,1,1,1) \to 0. \tag{38}$$

Additionally, since $|T_{2,0,0}(j)|/\widetilde{D}(G_n,G_0) \to 0$ for all $\widetilde{k}+1 \leq j \leq k_0$, we have that

$$\frac{1}{\widetilde{D}(G_n,G_0)} \cdot \sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\pi^n_i|\Delta a^n_{ij}|^2 \to 0. \tag{39}$$

Putting the results in equations (35), (38) and (39) together with the formulation of $\widetilde{D}(G_n, G_0)$ in equation (13), we obtain that

$$\frac{1}{\widetilde{D}(G_n,G_0)}\Bigg[\sum_{\substack{j:|\mathcal{A}_j|>1\\ j\in[\widetilde{k}]}}\sum_{i\in\mathcal{A}_j}\pi_i^n K_{ij}^n\Big(\widetilde{r}(|\mathcal{A}_j|),\frac{\widetilde{r}(|\mathcal{A}_j|)}{2},\frac{\widetilde{r}(|\mathcal{A}_j|)}{2},\widetilde{r}(|\mathcal{A}_j|),\frac{\widetilde{r}(|\mathcal{A}_j|)}{2}\Big) + \sum_{\substack{j:|\mathcal{A}_j|>1\\ \widetilde{k}+1\leq j\leq k_0}}\sum_{i\in\mathcal{A}_j}\pi_i^n K_{-3,ij}^n\Big(\widetilde{r}(|\mathcal{A}_j|),\frac{\widetilde{r}(|\mathcal{A}_j|)}{2},\widetilde{r}(|\mathcal{A}_j|),\frac{\widetilde{r}(|\mathcal{A}_j|)}{2}\Big)\Bigg] \to 1, \tag{40}$$

where $K_{-3,ij}^n(\kappa_1,\kappa_2,\kappa_3,\kappa_4,\kappa_5) := |\Delta c_{ij}^n|^{\kappa_1} + |\Delta\Gamma_{ij}^n|^{\kappa_2} + |\Delta b_{ij}^n|^{\kappa_4} + |\Delta\nu_{ij}^n|^{\kappa_5}$. Now, we will divide our arguments into two main scenarios based on the above limit:

Case 1: $\dfrac{\sum_{\substack{j:|\mathcal{A}_j|>1\\ j\in[\widetilde{k}]}}\sum_{i\in\mathcal{A}_j}\pi_i^n K_{ij}^n\Big(\widetilde{r}(|\mathcal{A}_j|),\frac{\widetilde{r}(|\mathcal{A}_j|)}{2},\frac{\widetilde{r}(|\mathcal{A}_j|)}{2},\widetilde{r}(|\mathcal{A}_j|),\frac{\widetilde{r}(|\mathcal{A}_j|)}{2}\Big)}{\widetilde{D}(G_n,G_0)} \not\to 0.$

This assumption indicates that we can find an index $j^* \in [\widetilde{k}] : |\mathcal{A}_{j^*}| > 1$ such that

$$\frac{1}{\widetilde{D}(G_n,G_0)}\cdot\sum_{i\in\mathcal{A}_{j^*}}\pi_i^n K_{ij^*}^n\Big(\widetilde{r}(|\mathcal{A}_{j^*}|),\frac{\widetilde{r}(|\mathcal{A}_{j^*}|)}{2},\frac{\widetilde{r}(|\mathcal{A}_{j^*}|)}{2},\widetilde{r}(|\mathcal{A}_{j^*}|),\frac{\widetilde{r}(|\mathcal{A}_{j^*}|)}{2}\Big) \not\to 0.$$

WLOG, we may assume that $j^* = 1$ throughout this case. Recall that $S_{\ell_1,\ell_2}^n(1)/\widetilde{D}(G_n,G_0) \to 0$ for all pairs (ℓ_1,ℓ_2) such that $0 \leq \ell_1+\ell_2 \leq 2\widetilde{r}(|\mathcal{A}_1|)$. Combine this result with the assumption of case 1, we obtain

$$\frac{S_{\ell_1,\ell_2}^n(1)}{D_1(G_n,G_0)} = \frac{S_{\ell_1,\ell_2}^n(1)}{\widetilde{D}(G_n,G_0)}\cdot\frac{\widetilde{D}(G_n,G_0)}{D_1(G_n,G_0)} \to 0,$$

where $D_1(G_n,G_0) := \sum_{i\in\mathcal{A}_1}\pi_i^n K_{i1}^n\Big(\widetilde{r}(|\mathcal{A}_1|),\frac{\widetilde{r}(|\mathcal{A}_1|)}{2},\frac{\widetilde{r}(|\mathcal{A}_1|)}{2},\widetilde{r}(|\mathcal{A}_1|),\frac{\widetilde{r}(|\mathcal{A}_1|)}{2}\Big)$. By expanding the formulations of $S_{\ell_1,\ell_2}^n(1)$ and $D_1(G_n,G_0)$, we have that

$$\frac{\sum_{i\in\mathcal{A}_1}\sum_{\alpha\in\mathcal{J}_{\ell_1,\ell_2}}\frac{\pi_i^n}{(\Gamma_1^0)^{\alpha_1+2\alpha_2}\ 2^{\alpha_2+\alpha_5}\alpha!}(\Delta c_{i1}^n)^{\alpha_1}(\Delta\Gamma_{i1}^n)^{\alpha_2}(\Delta a_{i1}^n)^{\alpha_3}(\Delta b_{i1}^n)^{\alpha_4}(\Delta\nu_{i1}^n)^{\alpha_5}}{\sum_{i\in\mathcal{A}_1}\pi_i^n\Big[|\Delta c_{i1}^n|^{\widetilde{r}(|\mathcal{A}_1|)} + |\Delta\Gamma_{i1}^n|^{\frac{\widetilde{r}(|\mathcal{A}_1|)}{2}} + |\Delta a_{i1}^n|^{\frac{\widetilde{r}(|\mathcal{A}_1|)}{2}} + |\Delta b_{i1}^n|^{\widetilde{r}(|\mathcal{A}_1|)} + |\Delta\nu_{i1}^n|^{\frac{\widetilde{r}(|\mathcal{A}_1|)}{2}}\Big]} \to 0. \tag{41}$$

Next, we define $\overline{M}_n := \max\{|\Delta c_{i1}^n|, |\Delta\Gamma_{i1}^n|^{1/2}, |\Delta a_{i1}^n|^{1/2}, |\Delta b_{i1}^n|, |\Delta\nu_{i1}^n|^{1/2} : i\in\mathcal{A}_1\}$ and $\overline{\pi}_n := \max_{i\in\mathcal{A}_1}\pi_i^n$. For any $i\in\mathcal{A}_1$, since the sequence $(\pi_i^n/\overline{\pi}_n)_{i\in\mathcal{A}_1}$ is bounded, we can substitute it with its subsequence that admits a non-negative limit $p_i^2 = \lim_{n\to\infty}\pi_i^n/\overline{\pi}_n$.

Additionally, we define $(\Delta c_{i1}^n)/(\Gamma_j^0\cdot\overline{M}_n) \to q_{1i}$, $(\Delta\Gamma_{i1}^n)/[2(\Gamma_j^0)^2\cdot\overline{M}_n^2] \to q_{2i}$, $(\Delta a_{i1}^n)/\overline{M}_n^2 \to q_{3i}$, $(\Delta b_{i1}^n)/\overline{M}_n \to q_{4i}$ and $(\Delta\nu_{i1}^n)/2\overline{M}_n^2 \to q_{5i}$. It can be seen from the formulation of $\mathcal{O}_{k,\beta}(\Theta)$ that $\pi_i^n \geq \delta$, therefore, p_i's will not vanish and at least one of them is equal to 1. Similarly, at least one of the limits $q_{1i}, q_{2i}, \ldots, q_{5i}$ will be equal to either 1 or -1.

Since

$$\frac{\sum_{i\in\mathcal{A}_1}\pi_i^n\Big[|\Delta c_{i1}^n|^{\widetilde{r}(|\mathcal{A}_1|)} + |\Delta\Gamma_{i1}^n|^{\frac{\widetilde{r}(|\mathcal{A}_1|)}{2}} + |\Delta a_{i1}^n|^{\frac{\widetilde{r}(|\mathcal{A}_1|)}{2}} + |\Delta b_{i1}^n|^{\widetilde{r}(|\mathcal{A}_1|)} + |\Delta\nu_{i1}^n|^{\frac{\widetilde{r}(|\mathcal{A}_1|)}{2}}\Big]}{\overline{\pi}_n\overline{M}_n^{\ell_1+\ell_2}} \not\to 0,$$

for all pairs (ℓ_1,ℓ_2) such that $\ell_1+\ell_2 \in [\widetilde{r}(|\mathcal{A}_1|)]$, we can divide both the numerator and the denominator in equation (41) by $\overline{\pi}_n\overline{M}_n^{\ell_1+\ell_2}$, and then let $n\to\infty$ to achieve the following system of polynomial equations:

$$\sum_{i\in\mathcal{A}_1}\sum_{\alpha\in\mathcal{J}_{\ell_1,\ell_2}}\frac{p_i^2\ q_{1i}^{\alpha_1}\ q_{2i}^{\alpha_2}\ q_{3i}^{\alpha_3}\ q_{4i}^{\alpha_4}\ q_{5i}^{\alpha_5}}{\alpha_1!\ \alpha_2!\ \alpha_3!\ \alpha_4!\ \alpha_5!} = 0,$$

for all pairs (ℓ_1, ℓ_2) such that $0 \leq \ell_1 + \ell_2 \leq \widetilde{r}(|\mathcal{A}_1|)$. Nevertheless, according to the definition of $\widetilde{r}(|\mathcal{A}_1|)$, the above system cannot admit any non-trivial solutions, which is a contradiction. Thus, case 1 does not hold.

Case 2: $\dfrac{1}{\widetilde{D}(G_n, G_0)} \cdot \sum_{j:|\mathcal{A}_j|>1} \sum_{\widetilde{k}+1 \leq j \leq k_0} \sum_{i \in \mathcal{A}_j} \pi_i^n K_{-3,ij}^n \Big(\widetilde{r}(|\mathcal{A}_j|), \frac{\widetilde{r}(|\mathcal{A}_j|)}{2}, \widetilde{r}(|\mathcal{A}_j|), \frac{\widetilde{r}(|\mathcal{A}_j|)}{2}\Big) \not\to 0.$

This assumption implies that there exists an index $\widetilde{k} + 1 \leq j^* \leq k_0 : |\mathcal{A}_{j^*}| > 1$ such that

$$\frac{1}{\widetilde{D}(G_n, G_0)} \cdot \sum_{i \in \mathcal{A}_{j^*}} \pi_i^n K_{-3,ij^*}^n \Big(\widetilde{r}(|\mathcal{A}_{j^*}|), \frac{\widetilde{r}(|\mathcal{A}_{j^*}|)}{2}, \widetilde{r}(|\mathcal{A}_{j^*}|), \frac{\widetilde{r}(|\mathcal{A}_{j^*}|)}{2}\Big) \not\to 0. \tag{42}$$

By applying similar arguments for equation (22) in the proof of Theorem 1 to equation (42), we are able to point out that equation (42) cannot happen, which is a contradiction. As a result, case 2 cannot happen either.

Collect the results of the above two scenarios, we realize that the limit in equation (40) does not hold true, which is a contradiction. As a consequence, not all the coefficients of elements in the set $\mathcal{H}$, defined in equation (34), in the representations of $A_n/\widetilde{D}(G_n, G_0)$, $B_n/\widetilde{D}(G_n, G_0)$ and $E_n/\widetilde{D}(G_n, G_0)$ go to zero as $n \to \infty$.

Step 3 - Application of Fatou's lemma: Next, we denote by m_n the maximum of the absolute values of those coefficients, which means that

$$m_n := \max\left\{\max_{(j,\ell_1,\ell_2)\in\mathcal{Q}} \frac{|Q_{\ell_1,\ell_2}^n(j)|}{\widetilde{D}(G_n, G_0)}, \max_{(j,\alpha_3,\tau_1,\tau_2)\in\mathcal{T}} \frac{|T_{\alpha_3,\tau_1,\tau_2}^n(j)|}{\widetilde{D}(G_n, G_0)}\right\}.$$

In addition, let us define $Q_{\ell_1,\ell_2}^n(j)/m_n \to \zeta_{\ell_1,\ell_2}(j)$ for $(j, \ell_1, \ell_2) \in \mathcal{Q}$ and $T_{\alpha_3,\tau_1,\tau_2}^n(j) \to \xi_{\alpha_3,\tau_1,\tau_2}(j)$ for $(j, \alpha_3, \tau_1, \tau_2) \in \mathcal{T}$ as $n \to \infty$. As not all the coefficients of elements of $\mathcal{H}$ in the representations of $A_n/\widetilde{D}(G_n, G_0)$, $B_n/\widetilde{D}(G_n, G_0)$ and $E_n/\widetilde{D}(G_n, G_0)$ vanish as $n \to \infty$, at least one among $\zeta_{\ell_1,\ell_2}(j)$ and $\xi_{\alpha_3,\tau_1,\tau_2}(j')$ is different from zero and $m_n \not\to 0$. By invoking the Fatou's lemma, we get that

$$0 = \lim_{n\to\infty} \frac{1}{m_n} \cdot \frac{2V(p_{G_n}, p_G)}{\widetilde{D}(G_n, G_0)} \geq \int \liminf_{n\to\infty} \frac{1}{m_n} \cdot \frac{|p_{G_n}(X,Y) - p_G(X,Y)|}{\widetilde{D}(G_n, G_0)} \, \mathrm{d}(X,Y) \geq 0.$$

Furthermore, we have that

$$\begin{aligned}
&\frac{1}{m_n} \cdot \frac{p_{G_n}(X,Y) - p_G(X,Y)}{\widetilde{D}(G_n, G_0)} \\
&\quad \to \sum_{(j,\ell_1,\ell_2)\in\mathcal{Q}} \zeta_{\ell_1,\ell_2}(j) X^{\ell_1} \cdot \frac{\partial^{\ell_2} f_{\mathcal{D}}}{\partial h_1^{\ell_2}}(Y|a_j^0 X + b_j^0, \nu_j^0) \cdot f_{\mathcal{L}}(X|c_j^0, \Gamma_j^0) \\
&\quad + \sum_{(j,\alpha_3,\tau_1,\tau_2)\in\mathcal{T}} \xi_{\alpha_3,\tau_1,\tau_2}(j) X^{\alpha_3} \cdot \frac{\partial^{\tau_1} f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_j^0, \Gamma_j^0) \cdot \frac{\partial^{\alpha_3+\tau_2} f_{\mathcal{D}}}{\partial h_1^{\alpha_3+\tau_2}}(Y|a_j^0 X + b_j^0, \nu_j^0).
\end{aligned}$$

Consequently, we achieve that

$$\begin{aligned}
&\sum_{(j,\ell_1,\ell_2)\in\mathcal{Q}} \zeta_{\ell_1,\ell_2}(j) X^{\ell_1} \cdot \frac{\partial^{\ell_2} f_{\mathcal{D}}}{\partial h_1^{\ell_2}}(Y|a_j^0 X + b_j^0, \nu_j^0) \cdot f_{\mathcal{L}}(X|c_j^0, \Gamma_j^0) \\
&+ \sum_{(j,\alpha_3,\tau_1,\tau_2)\in\mathcal{T}} \xi_{\alpha_3,\tau_1,\tau_2}(j) X^{\alpha_3} \frac{\partial^{\tau_1} f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_j^0, \Gamma_j^0) \frac{\partial^{\alpha_3+\tau_2} f_{\mathcal{D}}}{\partial h_1^{\alpha_3+\tau_2}}(Y|a_j^0 X + b_j^0, \nu_j^0) = 0,
\end{aligned}$$

for almost surely (X, Y). Since elements of the set $\mathcal{H}$ defined in equation (34) are linearly independent (proof of this claim is deferred to the end of this proof), the above equation indicates that $\zeta_{j,\ell_1,\ell_2}(j) = \xi_{\alpha_3,\tau_1,\tau_2}(j') = 0$ for all $(j, \ell_1, \ell_2) \in \mathcal{Q}$ and $(j', \alpha_3, \tau_1, \tau_2) \in \mathcal{T}$, which contradicts the fact that at least one among $\zeta_{j,\ell_1,\ell_2}(j)$, $\xi_{\alpha_3,\tau_1,\tau_2}(j')$ is different from zero. Hence, we reach the conclusion in equation (31).

Linear independence of elements in $\mathcal{H}$: For completion, we will show that elements of the set $\mathcal{F}$ defined in equation (34) are linearly independent by definition. In particular, assume that there exist real numbers $\zeta_{\ell_1,\ell_2}(j)$

and $\xi_{\alpha_3,\tau_1,\tau_2}(j')$, where $(j,\ell_1,\ell_2)\in\mathcal{Q}$ and $(j',\alpha_3,\tau_1,\tau_2)\in\mathcal{T}$, such that the following equation holds for almost surely (X,Y):

$$\sum_{(j,\ell_1,\ell_2)\in\mathcal{Q}}\zeta_{\ell_1,\ell_2}(j)X^{\ell_1}\cdot\frac{\partial^{\ell_2}f_{\mathcal{D}}}{\partial h_1^{\ell_2}}(Y|a_j^0X+b_j^0,\nu_j^0)\cdot f_{\mathcal{L}}(X|c_j^0,\Gamma_j^0)$$
$$+\sum_{(j',\alpha_3,\tau_1,\tau_2)\in\mathcal{T}}\xi_{\alpha_3,\tau_1,\tau_2}(j')X^{\alpha_3}\frac{\partial^{\tau_1}f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_{j'}^0,\Gamma_{j'}^0)\frac{\partial^{\alpha_3+\tau_2}f_{\mathcal{D}}}{\partial h_1^{\alpha_3+\tau_2}}(Y|a_{j'}^0X+b_{j'}^0,\nu_{j'}^0)=0,$$

Now, we rewrite the above equation as follows:

$$\sum_{j=1}^{k_0}\sum_{\omega=0}^{2\widetilde{r}(|\mathcal{A}_j|)}\Bigg[\sum_{\alpha_3+\tau_2=\omega}\sum_{\tau_1=0}^{2(\widetilde{r}(|\mathcal{A}_j|)-\alpha_3)-\tau_2}\xi_{\alpha_3,\tau_1,\tau_2}(j)X^{\alpha_3}\cdot\frac{\partial^{\tau_1}f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_j^0,\Gamma_j^0)\cdot\mathbf{1}_{\{\widetilde{k}+1\leq j\leq k_0\}}$$
$$+\sum_{\ell_1=0}^{2\widetilde{r}(|\mathcal{A}_j|)-\omega}\zeta_{\ell_1,\omega}(j)X^{\ell_1}\cdot f_{\mathcal{L}}(X|c_j^0,\Gamma_j^0)\cdot\mathbf{1}_{\{j\in[\widetilde{k}]\}}\Bigg]\frac{\partial^{\omega}f_{\mathcal{D}}}{\partial h_1^{\omega}}(Y|a_j^0X+b_j^0,\nu_j^0)=0, \tag{43}$$

for almost surely (X,Y). As (a_j^0,b_j^0,ν_j^0) for $j\in[k_0]$ are k_0 distinct tuples, we deduce that $((a_j^0)^\top X+b_j^0,\nu_j^0)$ for $j\in[k_0]$ are also k_0 distinct tuples for almost surely X. Thus, for almost surely X, one has $\frac{\partial^{\omega}f_{\mathcal{D}}}{\partial h_1^{\omega}}(Y|(a_j^0)^\top X+b_j^0,\nu_j^0)$ for $j\in[k_0]$ and $0\leq\omega\leq 2\widetilde{r}(|\mathcal{A}_j|)$ are linearly independent with respect to Y. Given that result, the equation (43) indicates that for almost surely X,

$$\sum_{\alpha_3+\tau_2=\omega}\sum_{\tau_1=0}^{2(\widetilde{r}(|\mathcal{A}_j|)-\alpha_3)-\tau_2}\xi_{\alpha_3,\tau_1,\tau_2}(j)X^{\alpha_3}\cdot\frac{\partial^{\tau_1}f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_j^0,\Gamma_j^0)\cdot\mathbf{1}_{\{\widetilde{k}+1\leq j\leq k_0\}}$$
$$+\sum_{\ell_1=0}^{2\widetilde{r}(|\mathcal{A}_j|)-\omega}\zeta_{\ell_1,\omega}(j)X^{\ell_1}\cdot f_{\mathcal{L}}(X|c_j^0,\Gamma_j^0)\cdot\mathbf{1}_{\{j\in[\widetilde{k}]\}}=0.$$

for all $j\in[k_0]$ and $0\leq\omega\leq 2\widetilde{r}(|\mathcal{A}_j|)$. This equation is equivalent to

$$\sum_{\ell_1=0}^{2\widetilde{r}(|\mathcal{A}_j|)-\omega}\zeta_{\ell_1,\omega}(j)X^{\ell_1}\cdot f_{\mathcal{L}}(X|c_j^0,\Gamma_j^0)=0, \tag{44}$$
$$\sum_{\alpha_3+\tau_2=\omega'}\sum_{\tau_1=0}^{2(\widetilde{r}(|\mathcal{A}_{j'}|)-\alpha_3)-\tau_2}\xi_{\alpha_3,\tau_1,\tau_2}(j')X^{\alpha_3}\cdot\frac{\partial^{\tau_1}f_{\mathcal{L}}}{\partial c^{\tau_1}}(X|c_{j'}^0,\Gamma_{j'}^0)=0, \tag{45}$$

for all $j\in[\widetilde{k}]$, $0\leq\omega\leq 2\widetilde{r}(|\mathcal{A}_j|)$ and $\widetilde{k}+1\leq j^{'}\leq k_0$, $0\leq\omega\leq 2\widetilde{r}(|\mathcal{A}_{j'}|)$. We can treat the left hand side of equation (44) as a polynomial of the random vector $X\in\mathcal{X}$, which is a compact set in $\mathbb{R}$. Meanwhile, the left hand side of equation (45) can be viewed as another polynomial of X and $X-c_{j'}^0$, where $c_{j'}^0\neq 0$. As a result, the above equations hold when $\zeta_{\ell_1,\omega}(j)=0$ for all $j\in[\widetilde{k}]$, $0\leq\omega\leq 2\widetilde{r}(|\mathcal{A}_j|)$, $0\leq\ell_1\leq 2\widetilde{r}(|\mathcal{A}_j|)-\omega$, and $\xi_{\alpha_3,\tau_1,\tau_2}(j')=0$ for all $\widetilde{k}+1\leq j'\leq k_0$, $0\leq\omega'\leq 2\widetilde{r}(|\mathcal{A}_{j'}|)$, $\alpha_3+\tau_2=\omega'$ and $0\leq\tau_1\leq 2(\widetilde{r}(|\mathcal{A}_j|)-\alpha_3)-\tau_2$. This result is equivalent to $\zeta_{\ell_1,\ell_2}(j)=0$, for all $(j,\ell_1,\ell_2)\in\mathcal{Q}$ and $\xi_{\alpha_3,\tau_1,\tau_2}(j')=0$ for all $(j',\alpha_3,\tau_1,\tau_2)\in\mathcal{T}$.

Hence, the elements of $\mathcal{H}$ are linearly independent, which completes the proof.

C PROOF OF REMAINING RESULTS

In this appendix, we provide proofs for Proposition 1, Proposition 2 and Lemma 2 in that order.

C.1 Proof of Proposition 1

For any two mixing measures $G = \sum_{i=1}^{k} \pi_i \delta_{(c_i,\Gamma_i,a_i,b_i,\nu_i)}$ and $G' = \sum_{i=1}^{k'} \pi'_i \delta_{(c'_i,\Gamma'_i,a'_i,b'_i,\nu'_i)}$, we assume that $p_G(X,Y) = p_{G'}(X,Y)$ holds true for almost surely $(X,Y) \in \mathcal{X} \times \mathcal{Y}$, or equivalently,

$$\sum_{i=1}^{k} \pi_i f_{\mathcal{L}}(X|c_i,\Gamma_i) f_{\mathcal{D}}(Y|(a_i)^\top X + b_i, \nu_i) = \sum_{i=1}^{k'} \pi'_i f_{\mathcal{L}}(X|c'_i,\Gamma'_i) f_{\mathcal{D}}(Y|(a'_i)^\top X + b'_i, \nu'_i). \tag{46}$$

Recall that if $Y|X \sim \mathcal{N}_1(a^\top X + b, \nu)$ and $X \sim \mathcal{N}_d(c,\Gamma)$, then

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_{d+1}\left(\begin{pmatrix} c \\ a^\top c + b \end{pmatrix}, \begin{pmatrix} \Gamma & \Gamma a \\ a^\top \Gamma & a^\top \Gamma a + \nu \end{pmatrix}\right).$$

Let us denote

$$\psi_i := \begin{pmatrix} c_i \\ (a_i)^\top c_i + b_i \end{pmatrix}, \quad \Sigma_i := \begin{pmatrix} \Gamma_i & \Gamma_i a_i \\ (a_i)^\top \Gamma_i & (a_i)^\top \Gamma_i a_i + \nu_i \end{pmatrix},$$
$$\psi'_i := \begin{pmatrix} c'_i \\ (a'_i)^\top c'_i + b'_i \end{pmatrix}, \quad \Sigma'_i := \begin{pmatrix} \Gamma'_i & \Gamma'_i a'_i \\ (a'_i)^\top \Gamma'_i & (a_i)^\top \Gamma'_i a'_i + \nu'_i \end{pmatrix}.$$

Then, equation (46) can be rewritten as

$$\sum_{i=1}^{k} \pi_i f(X,Y|\psi_i,\Sigma_i) = \sum_{i=1}^{k'} \pi'_i f(X,Y|\psi'_i,\Sigma'_i), \tag{47}$$

for almost surely (X,Y), where f belongs to the family of $(d+1)$-dimensional Gaussian density functions. Since the location-scale Gaussian mixtures are identifiable, it follows from the above equation that $k = k'$ and $\{\pi_1, \pi_2, \ldots, \pi_k\} \equiv \{\pi'_1, \pi'_2, \ldots, \pi'_k\}$. WLOG, we may assume that $\pi_i = \pi'_i$ for any $i \in [k]$.

Subsequently, we construct a partition of the set $[k]$, denoted by $P_1, P_2, \ldots, P_m$ that satisfies the following properties:

(i) $\pi_i = \pi'_i$ for any $i \in P_\ell$ and $\ell \in [m]$;

(ii) $\pi_i \neq \pi'_j$ if i and j are not in the same set P_ℓ for any $\ell \in [m]$.

Given this partition, we represent equation (47) as follows:

$$\sum_{\ell=1}^{m} \sum_{i \in P_\ell} \pi_i f(X,Y|\psi_i,\Sigma_i) = \sum_{\ell=1}^{m} \sum_{i \in P_\ell} \pi'_i f(X,Y|\psi'_i,\Sigma'_i),$$

for almost surely (X,Y). Consequently, for each $\ell \in [m]$, we obtain that

$$\{(\psi_i,\Sigma_i) : i \in P_\ell\} \equiv \{(\psi'_i,\Sigma'_i) : i \in P_\ell\}.$$

WLOG, we may assume that $(\psi_i,\Sigma_i) = (\psi'_i,\Sigma'_i)$ for any $i \in P_\ell$. Given this result, by some simple algebraic derivations, we achieve that $(c_i,\Gamma_i,a_i,b_i,\nu_i) = (c'_i,\Gamma'_i,a'_i,b'_i,\nu'_i)$ for any $i \in P_\ell$ and $\ell \in [m]$. As a result, it follows that

$$G = \sum_{\ell=1}^{m} \sum_{i \in P_\ell} \pi_i \delta_{(c_i,\Gamma_i,a_i,b_i,\nu_i)} = \sum_{\ell=1}^{m} \sum_{i \in P_\ell} \pi'_i \delta_{(c'_i,\Gamma'_i,a'_i,b'_i,\nu'_i)} = G'.$$

Hence, the proof is completed.

C.2 Proof of Proposition 2

Prior to presenting the proof of Proposition 2, let us review fundamental background on density estimation for M-estimators, which is covered in van de Geer (2000). First of all, we define $\mathcal{P}_{k,\beta}(\Theta) := \{p_G(X,Y) : G \in \mathcal{O}_{k,\beta}(\Theta)\}$ as the set of joint densities of all mixing measure in $\mathcal{O}_{k,\beta}(\Theta)$. In addition, we denote

$$\mathcal{Q}_{k,\beta}(\Theta) := \{p_{(G+G_0)/2}(X,Y) : G \in \mathcal{O}_{k,\beta}(\Theta)\},$$
$$\mathcal{Q}^{1/2}_{k,\beta}(\Theta) := \{p^{1/2}_{(G+G_0)/2}(X,Y) : G \in \mathcal{O}_{k,\beta}(\Theta)\}.$$

Subsequently, for any $\delta > 0$, the Hellinger ball centered around the density $p_{G_0}(X,Y)$ and intersected with the set $\mathcal{Q}^{1/2}_{k,\beta}(\Theta)$ is defined as

$$\mathcal{Q}^{1/2}_{k,\beta}(\Theta,\delta) := \{g^{1/2} \in \mathcal{Q}^{1/2}_{k,\beta}(\Theta) : h(g,p_{G_0}) \leq \delta\}.$$

Additionally, Geer et al. van de Geer (2000) introduce the following quantity to capture the size of the above Hellinger ball:

$$\mathcal{J}_B(\delta, \mathcal{Q}^{1/2}_{k,\beta}(\Theta)) := \int_{\delta^2/2^{13}}^{\delta} H^{1/2}_B\Big(u, \mathcal{Q}^{1/2}_{k,\beta}(\Theta,u), \|\cdot\|\Big)\mathrm{d}u \vee \delta, \tag{48}$$

where $H^{1/2}_B\Big(u, \mathcal{Q}^{1/2}_{k,\beta}(\Theta,u), \|\cdot\|\Big)$ denotes the bracketing entropy of $\mathcal{Q}^{1/2}_{k,\beta}(\Theta,u)$ under the Euclidean distance, and $u \vee \delta := \max\{u,\delta\}$. Given these notations, let us state the result regarding the joint density estimation rate presented in Theorem 7.4 in van de Geer (2000).

Lemma 3 (Theorem 7.4, van de Geer (2000))**.** *Take $\Psi(\delta) \geq \mathcal{J}_B(\delta, \mathcal{Q}^{1/2}_{k,\beta}(\Theta))$ such that $\Psi(\delta)/\delta^2$ is a non-increasing function of δ. Then, for a universal constant c and a sequence (δ_n) that satisfies $\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n)$, we obtain that*

$$\mathbb{P}\Big(h(p_{\widehat{G}_n}, p_{G_0}) > \delta\Big) \leq c\exp\left(-\frac{n\delta^2}{c^2}\right),$$

for any $\delta \geq \delta_n$.

Proof of Lemma 3 is provided in van de Geer (2000). Next, we introduce the upper bounds of the covering number (under the sup norm) $N(\varepsilon, \mathcal{P}_{k,\beta}(\Theta), \|\cdot\|_\infty)$, and the bracketing entropy (under the Hellinger distance) $H_B(\varepsilon, \mathcal{P}_{k,\beta}(\Theta), h)$ of the metric space $\mathcal{P}_{k,\beta}(\Theta)$. For further detail about the definitions of these terms, readers are referred to van de Geer (2000).

Lemma 4. *Given a bounded set Θ, we have for any $\varepsilon \in [0,1/2]$ that*

(i) $\log N(\varepsilon, \mathcal{P}_{k,\beta}(\Theta), \|\cdot\|_\infty) \lesssim \log(1/\varepsilon)$;

(ii) $H_B(\varepsilon, \mathcal{P}_{k,\beta}(\Theta), h) \lesssim \log(1/\varepsilon)$.

Proof of Lemma 4 is relegated to Appendix C.2.2. Now, we already have all necessary ingredients to provide the proof for Proposition 2 in Appendix C.2.1

C.2.1 Proof of Proposition 2

Note that for any $u > 0$, we have

$$H_B(u, \mathcal{Q}^{1/2}_{k,\beta}(\Theta), \|\cdot\|) \leq H_B(u, \mathcal{P}_{k,\beta}(\Theta), h) \leq \log(1/u),$$

where the second inequality is induced by part (ii) of Lemma 4. Then, it follows from equation (48) that

$$\mathcal{J}_B(\delta, \mathcal{Q}^{1/2}_{k,\beta}(\Theta)) \leq \int_{\delta^2/2^{13}}^{\delta} \log(1/u)\mathrm{d}u \vee \delta. \tag{49}$$

By choosing $\Psi(\delta) := \delta \cdot [\log(1/\delta)]^{1/2}$, we get that $\Psi(\delta)/\delta^2$ is a non-increasing function of δ and $\Psi(\delta) \geq \mathcal{J}_B(\delta, \mathcal{Q}_{k,\beta}^{1/2}(\Theta))$ from equation (49). Let $\delta_n := \sqrt{\log(n)/n}$, we achieve that $\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n)$ for some universal constant c. As a result, Lemma 3 gives us that

$$\mathbb{P}(h(p_{\widehat{G}_n}, p_{G_0}) > C_1\sqrt{\log(n)/n}) \lesssim \exp(-C_2\log(n)) = n^{-C_2},$$

where C_1 and C_2 are some universal constants. Finally, since the Total Variation is upper bounded by the Hellinger distance, we obtain the desired conclusion.

C.2.2 Proof of Lemma 4

Part (i). Given some $\varepsilon > 0$, since Θ is a compact set, we can find an ε-cover of Θ, denoted by Θ_ε. Additionally, let Δ_ε be an ε-cover of an $(k-1)$-dimensional simplex. Assume that $|\Theta_\varepsilon| = T$ and $|\Delta_\varepsilon| = S$. Note that $\Theta \subset \mathbb{R}^d \times \mathcal{S}_d^+ \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+$ is a subspace of $\mathbb{R}^{d^2+4d}$, then it can be checked that $T = \mathcal{O}(\varepsilon^{-(d^2+4d)k})$ and $S = \mathcal{O}(\varepsilon^{-(k-1)})$. Next, we define

$$\mathcal{G} := \{p_G \in \mathcal{P}_{k,\beta}(\Theta) : (\pi_1, \pi_2, \ldots, \pi_k) \in \Delta_\varepsilon, (c_i, \Gamma_i, a_i, b_i, \nu_i) \in \Theta_\varepsilon\}.$$

Given some mixing measure $G = \sum_{i=1}^{k'} \pi_i \delta_{\theta_i} \in \mathcal{O}_{k,\beta}(\Theta)$ with $k' \leq k$ and $\theta_i := (c_i, \Gamma_i, a_i, b_i, \nu_i) \in \Theta$, let us consider $\overline{G} = \sum_{i=1}^{k'} \pi_i \delta_{\widetilde{\theta}_i}$ where $\widetilde{\theta}_i := (\widetilde{c}_i, \widetilde{\Gamma}_i, \widetilde{a}_i, \widetilde{b}_i, \widetilde{\nu}_i) \in \Theta_\varepsilon$ such that $\|\widetilde{\theta}_i - \theta_i\| \leq \varepsilon$ for any $i \in [k']$. In addition, we also take into account another mixing measure $\widetilde{G} = \sum_{i=1}^{k'} \widetilde{\pi}_i \delta_{\widetilde{\theta}_i}$ where $(\widetilde{\pi}_1, \widetilde{\pi}_2, \ldots, \widetilde{\pi}_{k'}, 0, \ldots, 0) \in \Delta_\varepsilon$ such that $\|(\widetilde{\pi}_i)_{i=1}^{k'} - (\pi_i)_{i=1}^{k'}\| \leq \varepsilon$. From the definition of $\mathcal{G}$, we get that $p_{\widetilde{G}} \in \mathcal{G}$. Since $\|(\widetilde{\pi}_i)_{i=1}^{k'} - (\pi_i)_{i=1}^{k'}\| \leq \varepsilon$, we can deduce that

$$\|p_{\overline{G}} - p_{\widetilde{G}}\|_\infty \leq \sum_{i=1}^{k'} |\widetilde{\pi}_i - \pi_i| \cdot \|f_{\mathcal{L}}(X|\widetilde{c}_i, \widetilde{\Gamma}_i) f_{\mathcal{D}}(Y|(\widetilde{a}_i)^\top X + \widetilde{b}_i, \widetilde{\nu}_i)\|_\infty \lesssim \varepsilon.$$

Next, we consider

$$\|p_G - p_{\overline{G}}\|_\infty \leq \sum_{i=1}^{k'} \pi_i \|F(\theta_i|X,Y) - F(\widetilde{\theta}_i|X,Y)\|_\infty,$$

where we denote $F(\theta|X,Y) := f_{\mathcal{L}}(X|c,\Gamma) f_{\mathcal{D}}(Y|a^\top X + b, \nu)$. As F is twice differentiable with respect to θ and $\mathcal{X}$ is a bounded set, we achieve the following inequality:

$$\sum_{i=1}^{k'} \pi_i \|F(\theta_i|X,Y) - F(\widetilde{\theta}_i|X,Y)\|_\infty \leq \sum_{i=1}^{k'} \pi \|\widetilde{\theta}_i - \theta_i\| \lesssim \varepsilon,$$

which leads to $\|p_G - p_{\overline{G}}\|_\infty \leq \varepsilon$. As a consequence, by the triangle inequality, we have

$$\|p_G - p_{\widetilde{G}}\|_\infty \leq \|p_G - p_{\overline{G}}\|_\infty + \|p_{\overline{G}} - p_{\widetilde{G}}\|_\infty \lesssim \varepsilon.$$

Given this result, it follows that $\mathcal{G}$ is an ε-cover of $\mathcal{P}_{k,\beta}(\Theta)$, therefore,

$$N\Big(\varepsilon, \mathcal{P}_{k,\beta}(\Theta), \|\cdot\|_\infty\Big) \leq |\mathcal{G}| = S \times T = \mathcal{O}(\varepsilon^{-(d^2+4d)k}) \times \mathcal{O}(\varepsilon^{-(k-1)}) = \mathcal{O}(\varepsilon^{-(d^2+4d+1)k+1}),$$

which implies that $\log N\Big(\varepsilon, \mathcal{P}_{k,\beta}(\Theta), \|\cdot\|_\infty\Big) \lesssim \log(1/\varepsilon)$.

Part (ii). We begin with finding an upper bound for the density $f_{\mathcal{L}}(X|c,\Gamma) f_{\mathcal{D}}(Y|a^\top X + b, \nu)$. Since $\mathcal{X}$ and Θ are bounded sets, we can find positive constants $u, u_1, u_2, u_3, l_1, l_3$ such that $\|c\| \leq u$, $l_1 \leq \lambda_{\min}(\Gamma) \leq \lambda_{\max}(\Gamma) \leq u_1$, $-u_2 \leq a^\top X + b \leq u_2$ and $l_3 \leq \nu \leq u_3$, where $\lambda_{\min}(\Gamma)$ and $\lambda_{\max}(\Gamma)$ are the smallest and the largest eigenvalues of Γ, respectively. Firstly, it is clear that

$$f_{\mathcal{L}}(X|c,\Gamma) = \frac{1}{\sqrt{(2\pi)^d \det(\Gamma)}} \exp\Big(-\frac{1}{2}(x-c)^\top \Gamma^{-1}(x-c)\Big) \leq \frac{1}{(2\pi l_1)^{d/2}}.$$

Additionally, note that

$$(X-c)^\top \Gamma^{-1}(x-c) \geq \lambda_{\min}(\Gamma^{-1})\|X-c\|^2 = \frac{1}{\lambda_{\max}(\Gamma)}\|X-c\|^2.$$

Moreover, for any $\|X\| \geq 2u$, by the Cauchy-Schwartz inequality, we get

$$4\|X-c\|^2 - \|X\|^2 = 3\|X\|^2 - 8X^\top c + 4\|c\|^2 \geq 3\|X\|^2 - 8\|X\| \cdot \|c\| + 4\|c\|^2 \geq 0,$$

which implies that $(X-c)^\top \Gamma^{-1}(x-c) \geq \frac{1}{4u_1}\|X\|^2$. As a result,

$$f_{\mathcal{L}}(X|c,\Gamma) = \frac{1}{\sqrt{(2\pi)^d \det(\Gamma)}} \exp\Big(-\frac{1}{2}(x-c)^\top \Gamma^{-1}(x-c)\Big) \leq \frac{1}{(2\pi l_1)^{d/2}} \exp\left(-\frac{\|X\|^2}{8u_1}\right),$$

for any $\|X\| \geq 2u$. Combine this result with the previous bound, we obtain that $f_{\mathcal{L}}(X|c,\Gamma) \leq G_1(X)$, where

$$G_1(X) := \begin{cases} \dfrac{1}{(2\pi l_1)^{d/2}} \exp\left(-\dfrac{\|X\|^2}{8u_1}\right), & \|X\| \geq 2u, \\ \dfrac{1}{(2\pi l_1)^{d/2}}, & \|X\| < 2u. \end{cases}$$

By arguing in a similar fashion, we also have $f_{\mathcal{D}}(Y|a^\top X + b, \nu) \leq G_2(X,Y)$ where

$$G_2(X,Y) := \begin{cases} \dfrac{1}{\sqrt{2\pi l_3}} \exp\left(-\dfrac{Y^2}{8u_3}\right), & |Y| \geq 2u_2 \\ \dfrac{1}{\sqrt{2\pi l_3}}, & |Y| < 2u_2. \end{cases}$$

Consequently, we achieve that $f_{\mathcal{L}}(X|c,\Gamma) f_{\mathcal{D}}(Y|a^\top X + b, \nu) \leq G(X,Y) := G_1(X)G_2(X,Y)$.

Next, given some $\eta > 0$ that we will choose later, we consider an η-cover of $\mathcal{P}_{k,\beta}(\Theta)$ which is assumed to have N elements denoted by $f_1, f_2, \ldots, f_N$. For any $i \in [N]$, we define

$$L_i(X,Y) := \max\{f_i(X,Y) - \eta, 0\}, \quad U_i(X,Y) := \{f_i(X,Y) + \eta, G(X,Y)\}.$$

Then, we can validate that $\mathcal{P}_{k,\beta}(\Theta) \subset \cup_{i=1}^N [L_i(X,Y), U_i(X,Y)]$ and $U_i(X,Y) - L_i(X,Y) \leq \min\{2\eta, G(X,Y)\}$. Furthermore, we also deduce that

$$\begin{aligned} \|U_i - L_i\|_1 &= \int (U_i(X,Y) - L_i(X,Y))\mathrm{d}(X,Y) \\ &= \int_{|Y|<2u_2} (U_i(X,Y) - L_i(X,Y))\mathrm{d}(X,Y) + \int_{|Y|\geq 2u_2} (U_i(X,Y) - L_i(X,Y))\mathrm{d}(X,Y) \\ &\leq c_1\eta + \exp(-c_1^2/(2u_3)) \leq c_2\eta, \end{aligned}$$

where $c_1 = \max\{2u_2, \sqrt{8}u_3\}\log(1/\eta)$ and $c_2 > 0$ is some universal constant. This means that each bracket $[L_i(X,Y), U_i(X,Y)]$ is of size $c_2\eta$. Recall that the bracketing entropy is the logarithm of the smallest number of brackets to cover $\mathcal{P}_{k,\beta}(\Theta)$, it follows that

$$H_B(c_2\eta, \mathcal{P}_{k,\beta}(\Theta), \|\cdot\|_1) \leq \log N(\eta, \mathcal{P}_{k,\beta}(\Theta), \|\cdot\|_1) \leq \log N(\eta, \mathcal{P}_{k,\beta}(\Theta), \|\cdot\|_\infty) \lesssim \log(1/\eta),$$

where the second inequality occurs since $\|\cdot\|_\infty \leq \|\cdot\|_1$, while the last inequality is due to the result in part (i). Moreover, as the Hellinger distance is upper bounded by the $L1$-norm $\|\cdot\|_1$, we get that

$$H_B(c_2\eta, \mathcal{P}_{k,\beta}(\Theta), h) \leq H_B(c_2\eta, \mathcal{P}_{k,\beta}(\Theta), \|\cdot\|_1) \lesssim \log(1/\eta).$$

Here, if we choose $\eta = \varepsilon/c_2$, we can conclude that $H_B(\varepsilon, \mathcal{P}_{k,\beta}(\Theta), h) \leq \log(1/\varepsilon)$.

C.3 Proof of Lemma 2

We begin with recalling the system of interest here:

$$\sum_{l=1}^{m}\sum_{\alpha\in\mathcal{J}_{\ell_1,\ell_2}}\frac{p_l^2\, q_{1l}^{\alpha_1}\, q_{2l}^{\alpha_2}\, q_{3l}^{\alpha_3}\, q_{4l}^{\alpha_4}\, q_{5l}^{\alpha_5}}{\alpha_1!\,\alpha_2!\,\alpha_3!\,\alpha_4!\,\alpha_5!}=0, \tag{50}$$

with unknown variables $\{(p_l, q_{1l}, q_{2l}, q_{3l}, q_{4l}, q_{5l})\}_{l=1}^{m}\subset\mathbb{R}^5$ for all $\ell_1\geq 0$ and $\ell_2\geq 0$ that satisfy $1\leq\ell_1+\ell_2\leq r$, where

$$\mathcal{J}_{\ell_1,\ell_2}:=\{\alpha=(\alpha_i)_{i=1}^5\in\mathbb{N}^5:\alpha_1+2\alpha_2+\alpha_3=\ell_1,\ \alpha_3+\alpha_4+2\alpha_5=\ell_2\}.$$

Let us consider only a part of the above system when $\ell_1=0$ as follows:

$$\sum_{l=1}^{m}\sum_{\substack{\alpha_4,\alpha_5\in\mathbb{N}\\ \alpha_4+2\alpha_5=\ell_2}}\frac{p_l^2\, q_{4l}^{\alpha_4}\, q_{5l}^{\alpha_5}}{\alpha_4!\,\alpha_5!}=0, \tag{51}$$

for all $1\leq\ell_2\leq r$, which takes the same form as the system in equation (9). Thus, it follows from Lemma 1 that the smallest positive integer r such that the system (51) does not admit any non-trivial solutions is $\bar{r}(m)$. Therefore, we obtain that $\widetilde{r}(m)\leq\bar{r}(m)$.

Next, we will respectively show that $\widetilde{r}(2)=4$ and $\widetilde{r}(3)=6$.

When $m=2$: In this case, it follows from the above result that $\widetilde{r}(m)\leq\bar{r}(m)=4$. Thus, it is sufficient to demonstrate $\widetilde{r}(m)>3$, i.e. pointing out a non-trivial solution for the system (50) when $r=3$, which is given by

$$\begin{gathered}
\sum_{l=1}^{m}p_l^2 q_{1l}=0,\quad \sum_{l=1}^{m}p_l^2 q_{4l}=0,\\
\sum_{l=1}^{m}p_l^2\Big(\frac{1}{2!}q_{1l}^2+q_{2l}\Big)=0,\quad \sum_{l=1}^{m}p_l^2\Big(q_{1l}q_{4l}+q_{3l}\Big)=0,\quad \sum_{l=1}^{m}p_l^2\Big(\frac{1}{2!}q_{4l}^2+q_{5l}\Big)=0,\\
\sum_{l=1}^{m}p_l^2\Big(\frac{1}{3!}q_{1l}^3+q_{1l}q_{2l}\Big)=0,\quad \sum_{l=1}^{m}p_l^2\Big(\frac{1}{2!}q_{1l}^2q_{4l}+q_{1l}q_{3l}+q_{2l}q_{4l}\Big)=0,\\
\sum_{l=1}^{m}p_l^2\Big(\frac{1}{2!}q_{1l}q_{4l}^2+q_{1l}q_{5l}+q_{3l}q_{4l}\Big)=0,\quad \sum_{l=1}^{m}p_l^2\Big(\frac{1}{3!}q_{4l}^3+q_{4l}q_{5l}\Big)=0.
\end{gathered} \tag{52}$$

We can check that the following is a non-trivial solution of the system (52):

$$\begin{gathered}
p_l=0,\quad q_{1l}=q_{2l}=q_{3l}=0,\quad \forall l\in[m],\\
q_{41}=1, q_{42}=-1,\quad q_{51}=q_{52}=-\frac{1}{2}.
\end{gathered}$$

Hence, we conclude that $\widetilde{r}(m)=4$.

When $m=3$: Again, according to Lemma 1, we have $\widetilde{r}(m)\leq\bar{r}(m)=6$. Therefore, it suffices to show a non-trivial solution of the system (50) for $r=5$, which is a combination of the system (52) and the following

system:

$$\sum_{l=1}^{m} p_l^2\Big(\frac{1}{4!}q_{1l}^4+\frac{1}{2!}q_{1l}^2q_{2l}+\frac{1}{2!}q_{2l}^2\Big)=0,\quad \sum_{l=1}^{m} p_l^2\Big(\frac{1}{4!}q_{4l}^2+\frac{1}{2!}q_{4l}^2q_{5l}+\frac{1}{2!}q_{5l}^2\Big)=0,$$

$$\sum_{l=1}^{m} p_l^2\Big(\frac{1}{3!}q_{1l}^3q_{4l}+\frac{1}{2!}q_{1l}q_{2l}^2+\frac{1}{2!}q_{1l}^2q_{2l}+q_{2l}q_{3l}\Big)=0,$$

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

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

$$\sum_{l=1}^{m} p_l^2\Big(\frac{1}{5!}q_{1l}^5+\frac{1}{3!}q_{1l}^3q_{2l}+\frac{1}{2!}q_{1l}q_{2l}^2\Big)=0,\quad \sum_{l=1}^{m} p_l^2\Big(\frac{1}{5!}q_{4l}^5+\frac{1}{3!}q_{4l}^3q_{5l}+\frac{1}{2!}q_{4l}q_{5l}^2\Big)=0,$$

$$\sum_{l=1}^{m} p_l^2\Big(\frac{1}{4!}q_{1l}^4q_{4l}+\frac{1}{2!}q_{1l}^2q_{2l}+\frac{1}{2!}q_{2l}^2q_{4l}+\frac{1}{3!}q_{1l}^3q_{3l}+q_{1l}q_{2l}q_{3l}\Big)=0,$$

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

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

$$\sum_{l=1}^{m} p_l^2\Big(\frac{1}{2!3!}q_{1l}^2q_{4l}^3+q_{1l}^2q_{4l}q_{5l}+\frac{1}{3!}q_{2l}q_{3l}^3+q_{2l}q_{4l}q_{5l}+\frac{1}{2!}q_{1l}q_{3l}q_{4l}^2+q_{1l}q_{3l}q_{5l}+\frac{1}{2!}q_{3l}^2q_{4l}\Big)=0.$$

It can be verified that the following is a non-trivial of this system:

$$p_l=0,\quad q_{1l}=q_{2l}=q_{3l}=0,\quad \forall l\in[m],$$
$$q_{41}=\frac{\sqrt{3}}{3},q_{42}=-\frac{\sqrt{3}}{3},q_{43}=0,\quad q_{51}=q_{52}=-\frac{1}{6},q_{53}=0.$$

As a consequence, we obtain that $\widetilde{r}(m)>5$, which implies the desired conclusion that $\widetilde{r}(m)=6$.