Convergence Rates for Gaussian Mixtures of Experts

Nhat Ho*, Chiao-Yu Yang†, Michael I. Jordan†,*

Department of Electrical Engineering and Computer Sciences*
Department of Statistics†
UC Berkeley, Berkeley, CA 94720

July 9, 2019

Abstract

We provide a theoretical treatment of over-specified Gaussian mixtures of experts with covariate-free gating networks. We establish the convergence rates of the maximum likelihood estimation (MLE) for these models. Our proof technique is based on a novel notion of *algebraic independence* of the expert functions. Drawing on optimal transport theory, we establish a connection between the algebraic independence and a certain class of partial differential equations (PDEs). Exploiting this connection allows us to derive convergence rates and minimax lower bounds for parameter estimation.

1 Introduction

Gaussian mixtures of experts, a class of piece-wise regression models introduced by [9, 14, 15], have found applications in many fields including social science [8, 7, 2], speech recognition [21, 19], natural language processing [3, 24, 18, 19], and system identification [22]. Gaussian mixtures of experts differ from classical finite Gaussian mixture models in two ways. First, the mixture components (the "experts") are regression models, linking the location and scale of a Gaussian model of the response variable to a covariate vector X via parametric models $h_1(X, \theta_1)$ and $h_2(X, \theta_2)$, where θ_1, θ_2 are parameters. Second, the mixing proportions (the "gating network") are also functions of the covariate vector X, via a parametric model $\pi(X, \gamma)$ that maps X to a probability distribution over the labels of the experts. The overall model can be viewed as a covariate-dependent finite mixture. Despite their popularity in applications, the theoretical understanding of Gaussian mixtures of experts has proved challenging and lagged behind that of finite mixture models. The inclusion of covariates X in the experts and the gating networks leads to complex interactions of their parameters, which complicates the theoretical analysis.

In the setting of finite mixture models, while the early literature focused on identifiability issues [25, 26, 27, 17], recent work has provided a substantive inferential theory; see for example [23, 20, 5, 6]. Chen [1] set the stage for these recent developments by establishing a convergence rate of $n^{-1/4}$ for parameter estimation in the univariate setting of over-specified mixture models. Later, Nguyen [20] used the Wasserstein metric to analyze the posterior convergence rates of parameter estimation for both finite and infinite mixtures. Recently, Ho et al. [6] provided a unified framework to rigorously characterize the convergence rates of parameter estimation based on the singularity structures of finite mixture models. Their results demonstrated that there is a connection between the singularities of these models and the algebraic-geometric structure of the parameter space.

Moving to Gaussian mixtures of experts, a classical line of research focused on the identifiability in these models [13] and on parameter estimation in the setting of exact-fitted models where the true number of components is assumed known [11, 10, 12]. This assumption is, however, overly strong for most applications; the true number of components is rarely known in practice. There are two common practical approaches to deal with this issue. The first approach relies on model selection, most notably the BIC penalty [16]. This approach is, however, computationally expensive as we need to search for the optimal number of components over all the possible values. Furthermore, the sample size may not be large enough to support this form of inference. The second approach is to over-specify the true model, by using rough prior knowledge to specify more components than is necessary. However, theoretical analysis is challenging in this setting, given the complicated interaction among the parameters of the expert functions, a phenomenon that does not occur in the exact-fitted setting of Gaussian mixtures of experts. Another challenge arises from *inhomogeneity*—some parameters tend to have faster convergence rates than other parameters. This inhomogeneity makes it nontrivial to develop an appropriate distance for characterizing convergence rates.

In the current paper we focused on a simplified setting in which the expert functions are covariate-dependent, but the gating network is not. We refer to this as the *Gaussian mixture of experts with covariate-free gating functions* (GMCF) model. Although simplified, this model captures the core of the mixtures-of-experts problem, which is the interactions among the different mixture components. We believe that the general techniques that we develop here can be extended to the full mixtures-of-experts model—in particular by an appropriate generalization of the transportation distance to capture the variation of parameters from the gating networks—but we leave the development of that direction to future work.

1.1 Setting

We propose a general theoretical framework for analyzing the statistical performance of maximum likelihood estimation (MLE) for parameters in the setting of over-specified Gaussian mixtures of experts with covariate-free gating functions. In particular, we assume that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. samples from a Gaussian mixture of experts with covariate-free gating functions (GMCF) of order k_0, with conditional density function $g_{G_0}(Y|X)$:

$$g_{G_0}(Y|X) := \sum_{i=1}^{k_0} \pi_i^0 f(Y|h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0)), \tag{1}$$

where $G_0 := \sum_{i=1}^{k_0} \pi_i^0 \delta_{(\theta_{1i}^0, \theta_{2i}^0)}$ is a true but unknown probability measure (mixing measure) and $\theta_{ji}^0 \in \Omega_j \subset \mathbb{R}^{q_j}$ for all i, j. We over-specify the true model by choosing $k > k_0$ components.

We estimate G_0 under the over-specified GMCF model via maximum likelihood estimation (MLE). We denote the MLE as $\widehat{G}_n$. Our results reveal a fundamental connection between the algebraic structure of the expert functions h_1 and h_2 and the convergence rates of the MLE through a general version of the optimal transport distance, which refers to as the *generalized transportation distance*. A similar distance has been used to study the effect of algebraic singularities on parameter estimation in classical finite mixtures [6].

1.2 Generalized transportation distance

In contrast to the traditional Wasserstein metric [29], the generalized transportation distance assigns different orders to each parameter. This special property of generalized transportation

distance provides us with a tool to capture the inhomogeneity of parameter estimation in Gaussian mixtures of experts. In order to define the generalized transportation distance, we first define the semi-metric $d_\kappa(.,.)$ for any vector $\kappa = (\kappa_1, \ldots, \kappa_{q_1+q_2}) \in \mathbb{N}^{q_1+q_2}$ as follows:

$$d_\kappa(\theta_1, \theta_2) := \Bigg(\sum_{i=1}^{q_1+q_2} |\theta_1^{(i)} - \theta_2^{(i)}|^{\kappa_i} \Bigg)^{1/\|\kappa\|_\infty},$$

for any $\theta_i = \left(\theta_i^{(1)}, \ldots, \theta_i^{(q_1+q_2)}\right) \in \mathbb{R}^{q_1+q_2}$. Generally, $d_\kappa(.,.)$ does not satisfy the standard triangle inequality. More precisely, when not all κ_i are identical, d_κ satisfies a triangle inequality only up to some positive constant less than one. When all κ_i are identical, d_κ becomes a metric.

Now, we let $G = \sum_{i=1}^k \pi_i \delta_{(\theta_{1i}, \theta_{2i})}$ be some probability measure. The generalized transportation distance between G and G_0 with respect to $\kappa = (\kappa_1, \ldots, \kappa_{q_1+q_2}) \in \mathbb{N}^{q_1+q_2}$ is given by:

$$\widetilde{W}_\kappa(G, G_0) := \Bigg(\inf \sum_{i,j} q_{ij} d_\kappa^{\|\kappa\|_\infty}(\eta_i, \eta_j^0) \Bigg)^{1/\|\kappa\|_\infty}, \tag{2}$$

where the infimum is taken over all couplings $\boldsymbol{q}$ between $\boldsymbol{\pi}$ and $\boldsymbol{\pi}^0$; i.e., where $\sum_j q_{ij} = \pi_i$ and $\sum_i q_{ij} = \pi_j^0$. Additionally, $\eta_i = (\theta_{1i}, \theta_{2i})$ and $\eta_j^0 = (\theta_{1j}^0, \theta_{2j}^0)$ for all i, j.

In general, the convergence rates of mixing measures under generalized Wasserstein distance translate directly to the convergence rates of their associated atoms or parameters. More precisely, assume that there exist a sequence $\{G_n\}$ and a vector $\kappa = (\kappa_1, \ldots, \kappa_{q_1+q_2}) \in \mathbb{N}^{q_1+q_2}$ such that $\widetilde{W}_\kappa(G_n, G_0) \to 0$ at rate $\omega_n = o(1)$ as $n \to \infty$. Then, we can find a sub-sequence of G_n such that each atom (support) $(\theta_{1i}^0, \theta_{2i}^0)$ of G_0 is the limit point of atoms of G_n. Additionally, the convergence rates for estimating $(\theta_{1i}^0)^{(u)}$, the uth component of θ_{1i}^0, are $\omega_n^{\|\kappa\|_\infty/\kappa_u}$ while those for estimating $(\theta_{2i}^0)^{(v)}$ are $\omega_n^{\|\kappa\|_\infty/\kappa_{q_1+v}}$ for $1 \leq u \leq q_1$ and $1 \leq v \leq q_2$. Furthermore, the convergence rates for estimating the weights associated with these parameters are $\omega_n^{\|\kappa\|_\infty}$. Finally, there may exist some atoms of G_n that converge to limit points outside the atoms of G_0. The convergence rates of these limit points are also similar to those for estimating the atoms of G_0.

1.3 Main contribution

The generalized transportation distance in (2) allows us to introduce a notion of *algebraic independence* between expert functions h_1 and h_2 that is expressed in the language of partial differential equations (PDEs). Using this notion, we are able to characterize the convergence rates of parameter estimation for several choices of expert functions h_1 and h_2 when they are either algebraically independent or not. Our overall contributions in the paper can be summarized as follows:

- **Algebraically independent settings:** When the expert functions h_1 and h_2 are algebraically independent, we establish the best possible convergence rate of order $n^{-1/4}$ for $\widetilde{W}_\kappa(\widehat{G}_n, G_0)$ (up to a logarithmic factor) where $\kappa = (2, \ldots, 2)$. Furthermore, we demonstrate that this convergence rate is minimax. That result directly translates to a convergence rate of $n^{-1/4}$ for the support of $\widehat{G}_n$.

- **Algebraically dependent settings:** When the expert functions h_1 and h_2 are algebraically dependent, we prove that the convergence rates of parameter estimation are very slow and inhomogeneous. More precisely, the rates of convergence are either determined by the solvability of a system of polynomial equations or by the admissibility of a system of polynomial limits. The formulations of these systems depend on the PDEs that capture the interactions among the parameters for the expert functions. Furthermore, we show that the inhomogeneity of parameter estimation can be characterized based on the generalized transportation distance.

Organization. The remainder of the paper is organized as follows. In Section 2, we introduce the problem setup for Gaussian mixtures of experts with covariate-free gating functions. Section 3 establishes convergence rates for parameter estimation and provides global miximax lower bounds under the algebraically independent setting. In Section 4, we consider various settings in which the expert functions are algebraically dependent and establish the convergence rates of parameter estimation under these settings. We provide proofs for a few key results in Section 5 while deferring the majority of the proofs to the Appendices. Finally, we conclude in Section 6.

Notation. For any vector $x \in \mathbb{R}^d$, we use superscript and subscript notation interchangeably, letting $x = (x^{(1)}, \ldots, x^{(d)})$ or $x = (x_1, \ldots, x_d)$. Thus, either $x^{(i)}$ or x_i is the i-th component of x. Additionally, for each $x \in \mathbb{R}^d$, we denote $x^{\kappa} = \prod_{i=1}^{d} (x^{(i)})^{\kappa_i}$ for any $\kappa = (\kappa^{(1)}, \ldots, \kappa^{(d)}) \in \mathbb{N}^d$. Finally, for any two vectors $x, y \in \mathbb{R}^d$, we write $x \preceq y$ if $x^{(i)} \leq y^{(i)}$ for all $1 \leq i \leq d$ and $x \prec y$ if $x \preceq y$ and $x \neq y$.

For any two density functions p, q (with respect to the Lebesgue measure μ), the total variation distance is given by $V(p, q) = \frac{1}{2} \int |p(x) - q(x)| \mathrm{d}\mu(x)$. The squared Hellinger distance is defined as $h^2(p, q) = \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 \mathrm{d}\mu(x)$.

2 Background

In this section, we provide the necessary background for our analysis of the convergence rates of the MLE under over-specified Gaussian mixtures of experts with covariate-free gating functions. In particular, in Section 2.1, we define the over-specified Gaussian mixture of experts with covariate-free gating functions, and in Section 2.2, we establish identifiability and smoothness properties for these models as well as establishing the convergence rates of density estimation.

2.1 Problem setup

Let $Y \in \mathcal{Y} \subset \mathbb{R}$ be a response variable of interest and let $X \in \mathcal{X} \subset \mathbb{R}^d$ be a vector of covariates believed to have an effect on Y. We start with a definition of identifiable expert functions.

Definition 1. *Given $\Theta \subset \mathbb{R}^q$ for some $q \geq 1$. We say that an expert function $h_1 : \mathcal{X} \times \Theta \to \mathbb{R}$ is* identifiable *if $h_1(X, \eta_1) = h_1(X, \eta_2)$ almost surely for X implies $\eta_1 = \eta_2$.*

Recall that we focus on Gaussian mixtures of experts [9, 14, 15] for which the gating functions are independent of covariate X. In particular, we denote $\{f(\cdot|\theta, \sigma)\}$ as the family of location-scale univariate Gaussian distributions and define our models of interest as follows.

Definition 2. *Assume that we are given two identifiable expert functions* $h_1 : \mathcal{X} \times \Omega_1 \to \Theta_1 \subset \mathbb{R}$ *and* $h_2 : \mathcal{X} \times \Omega_2 \to \Theta_2 \subset \mathbb{R}_+$ *where* $\Omega_i \subset \mathbb{R}^{q_i}$ *for given dimensions* $q_i \geq 1$ *as* $1 \leq i \leq 2$. *Let* $\{\pi_i\}_{i=1}^k$ *denote* k *weights with* $\sum_{i=1}^k \pi_i = 1$. *We say that* (X, Y) *follows a* Gaussian mixtures of experts with covariate-free gating functions *(GMCF) of order* k, *with respect to expert functions* h_1, h_2 *and gating functions* π_i, *if the conditional density function of* Y *given* X *has the following form*

$$\begin{aligned} g_G(Y|X) &:= \int f\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right) dG(\theta_1,\theta_2) \\ &= \sum_{i=1}^k \pi_i f(Y|h_1(X,\theta_{1i}), h_2(X,\theta_{2i})), \end{aligned}$$

where $G = \sum_{i=1}^k \pi_i \delta_{(\theta_{1i},\theta_{2i})}$ *is a discrete probability measure that has exactly* k *atoms on* $\Omega := \Omega_1 \times \Omega_2$.

As an example, when $q_1 = q_2 = d + 1$, generalized linear expert functions take the form $h_1(X, \theta_1) = \theta_1^\top [1, X]$ and $h_2(X, \theta_2) = \exp\left(\theta_2^\top [1, X]\right)$.

Over-specified GMCF Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. draws from a GMCF of order k_0 with conditional density function $g_{G_0}(Y|X)$ where $G_0 := \sum_{i=1}^{k_0} \pi_i^0 \delta_{(\theta_{1i}^0, \theta_{2i}^0)}$ is a true but unknown probability measure (mixing measure). Since k_0 is generally unknown in practice, one popular approach to estimate the mixing measure G_0 is based on over-specifying the true number of components k_0. In particular, we fit the true model with $k > k_0$ number of components where k is a given threshold that is chosen based on prior domain knowledge. We refer to this setting as the *over-specified GMCF*.

Maximum likelihood estimation (MLE) To obtain an estimate of G_0, we define the MLE as follows:

$$\widehat{G}_n := \arg\max_{G \in \mathcal{G}} \sum_{i=1}^n \log(g_G(Y_i|X_i)), \tag{3}$$

where $\mathcal{G}$ is some subset of $\mathcal{O}_k(\Omega) := \{G = \sum_{i=1}^l \pi_i \delta_{(\theta_{1i},\theta_{2i})} : 1 \leq l \leq k\}$, namely, the set of all discrete probability measures with at most k components. Detailed formulations of $\mathcal{G}$ will be given later based on the specific structures of expert functions h_1 and h_2.

Universal assumptions and notation Throughout this paper, we assume that Ω_1 and Ω_2 are compact subsets of $\mathbb{R}^{q_1}$ and $\mathbb{R}^{q_2}$ respectively. Additionally, $\Omega := \Omega_1 \times \Omega_2$ and X is a random vector and has a given prior density function $\overline{f}(X)$, which is independent of the choices of expert functions h_1, h_2. Furthermore, $\mathcal{X}$ is a fixed compact set of $\mathbb{R}^d$. Finally we denote

$$p_G(X, Y) := g_G(Y|X)\overline{f}(X)$$

as the joint distribution (or equivalently mixing density) of X and Y for any $G \in \mathcal{O}_k(\Omega)$.

2.2 General identifiability, smoothness condition, and density estimation

In order to establish the convergence rates of $\widehat{G}_n$, our analysis relies on three main ingredients: general identifiability of the GMCF, Hölder continuity of the GMCF up to any order $r \geq 1$, and parametric convergence rates for density estimation under the over-specified GMCF. We begin with the following result regarding the identifiability of GMCF.

Proposition 1. *For given identifiable expert functions h_1 and h_2, the GMCF is identifiable with respect to h_1 and h_2, namely, whenever there are finite discrete probability measures G and G' on Ω such that $p_G(X,Y) = p_{G'}(X,Y)$ almost surely $(X,Y) \in \mathcal{X} \times \mathcal{Y}$, then it follows that $G \equiv G'$.*

A second result that plays a central role in analyzing convergence of the MLE in over-specified GMCF is the uniform Hölder continuity, formulated as follows:

Proposition 2. *For any $r \geq 1$, the GMCF admits the uniform Hölder continuity up to the rth order, with respect to the expert functions h_1, h_2 and prior density function $\overline{f}$:*

$$\sum_{|\kappa|=r} \overline{f}(x) \Big| \Big(\frac{\partial^{|\kappa|} f}{\partial \theta_1^{\kappa_1} \partial \theta_2^{\kappa_2}} \left(y|h_1(x,\theta_1), h_2(x,\theta_2)\right) - \frac{\partial^{|\kappa|} f}{\partial \theta_1^{\kappa_1} \partial \theta_2^{\kappa_2}} f\left(y|h_1(x,\theta_1'), h_2(x,\theta_2')\right) \Big) \gamma^{\kappa} \Big| \leq C \|(\theta_1,\theta_2) - (\theta_1',\theta_2')\|^{\delta} \|\gamma\|^r,$$

for any $\gamma \in \mathbb{R}^{q_1+q_2}$ and for some positive constants δ and C that are independent of x, y and $(\theta_1,\theta_2), (\theta_1',\theta_2') \in \Omega$. Here, $\kappa = (\kappa_1,\kappa_2) \in \mathbb{N}^{q_1+q_2}$ where $\kappa_i \in \mathbb{N}^{q_i}$ for any $1 \leq i \leq 2$.

Finally, when the expert functions h_1 and h_2 are sufficiently smooth in terms of their parameters, we can guarantee the parametric convergence rate of density estimation.

Proposition 3. *Assume that the expert functions h_1 and h_2 are twice differentiable with respect to their parameters. Additionally, assume that there exist positive constants $a, \underline{\gamma}, \overline{\gamma}$ such that $h_1(X,\theta_1) \in [-a,a]$, $h_2(X,\theta_2) \in [\underline{\gamma}, \overline{\gamma}]$ for all $X \in \mathcal{X}, \theta_1 \in \Omega_1, \theta_2 \in \Omega_2$. Then, the following holds:*

$$\mathbb{P}(h(p_{\widehat{G}_n}, p_{G_0}) > C(\log n/n)^{1/2}) \precsim \exp(-c \log n) \tag{4}$$

for universal positive constants C and c that depend only on Ω.

The proof of Proposition 3 is provided in Appendix C.

3 Algebraically independent expert functions

In this section, we consider the MLE in (3) over the entire parameter space $\mathcal{O}_k(\Omega)$. That is, we let $\mathcal{G} = \mathcal{O}_k(\Omega)$. To analyze the convergence rates of MLE under over-specified GMCF we capture the algebraic interaction among the expert functions h_1 and h_2 via the following definition.

Definition 3. *We say that the expert functions h_1, h_2 are* algebraically independent *if they are twice differentiable with respect to their parameters θ_1 and θ_2 and the following holds:*

(O.1) *For any (θ_1, θ_2), if we have $\alpha_i, \beta_{uv} \in \mathbb{R}$ (for $1 \leq i \leq q_2$, and $1 \leq u, v \leq q_1$) such that $\beta_{uv} = \beta_{vu}$ and*

$$\sum_{i=1}^{q_2} \alpha_i \frac{\partial h_2^2}{\partial \theta_2^{(i)}}(X, \theta_2) + \sum_{1 \leq u,v \leq q_1} \beta_{uv} \frac{\partial h_1}{\partial \theta_1^{(u)}}(X, \theta_1) \frac{\partial h_1}{\partial \theta_1^{(v)}}(X, \theta_1) = 0,$$

almost surely in X, then we must also have $\alpha_i = \beta_{uv} = 0$ for all $1 \leq i \leq q_2$ and $1 \leq u, v \leq q_1$.

Note that in this definition we use the convention that if $\dfrac{\partial h_2^2}{\partial \theta_2^{(i)}}(X, \theta_2) = 0$ almost surely for some $1 \leq i \leq q_2$, then we have $\alpha_i = 0$. The same convention goes for other derivatives in Condition (O.1). An equivalent way to express the algebraic independence notion in Definition 3 is that the elements in a set of partial derivatives,

$$\left\{ \frac{\partial h_1}{\partial \theta_1^{(u)}}(X, \theta_1) \frac{\partial h_1}{\partial \theta_1^{(v)}}(X, \theta_1), \frac{\partial h_2^2}{\partial \theta_2^{(i)}}(X, \theta_2) : \ 1 \leq i \leq q_2, 1 \leq u, v \leq q_1 \right\},$$

are linearly independent with respect to X. To exemplify Definition 3, consider the following simple examples of expert functions h_1 and h_2 that are algebraically independent.

Example 3.1. *(a) Let $\mathcal{X} \subset \mathbb{R}$. If we choose expert functions $h_1(X, \theta_1) = \theta_1 X$ and $h_2^2(X, \theta_2) = \theta_2$ for all $\theta_1 \in \Omega_1 \subset \mathbb{R}$ and $\theta_2 \in \Omega_2 \subset \mathbb{R}_+$, then h_1 and h_2 are algebraically independent.*

(b) Let $\mathcal{X} \subset \mathbb{R}_+$. If we choose expert functions $h_1(X, \theta_1) = (\theta_1^{(1)} + \theta_1^{(2)} X)^m$ for all $\theta_1 = (\theta_1^{(1)}, \theta_1^{(2)}) \in \Omega_1 \subset \mathbb{R}^2$, where $m > 1$ and $h_2^2(X, \theta_2) = \theta_2 X$ for all $\theta_2 \in \Omega_2 \subset \mathbb{R}_+$, then h_1, h_2 are algebraically independent.

Under the algebraic independence condition for the expert functions h_1 and h_2, we have the following result regarding the convergence rates of parameter estimation $\widehat{G}_n$ as well as their corresponding minimax lower bound under the over-specified GMCF model.

Theorem 1. *Assume that expert functions h_1 and h_2 are algebraically independent. Then, we have:*

(a) (Maximum likelihood estimation) There exists a positive constant C_0 depending on G_0 and Ω such that

$$\mathbb{P}(\widetilde{W}_\kappa(\widehat{G}_n, G_0) > C_0 (\log n / n)^{1/4}) \precsim \exp(-c \log n),$$

where $\kappa = (2, \ldots, 2) \in \mathbb{N}^{q_1 + q_2}$ and c is a positive constant depending only on Ω.

(b) (Minimax lower bound) For any κ' such that $(1, \ldots, 1) \preceq \kappa' \prec \kappa = (2, \ldots, 2)$,

$$\inf_{\overline{G}_n} \sup_{G \in \mathcal{O}_k(\Omega) \setminus \mathcal{O}_{k_0 - 1}(\Omega)} \mathbb{E}_{p_G} \left(\widetilde{W}_{\kappa'}(\overline{G}_n, G) \right) \geq c' n^{-1/(2 \| \kappa' \|_\infty)}.$$

Here, the infimum is taken over all sequences of estimates $\overline{G}_n \in \mathcal{O}_k(\Omega)$. Furthermore, $\mathbb{E}_{p_G}$ denotes the expectation taken with respect to the product measure with mixture density p_G^n, and c' stands for a universal constant depending only on Ω.

The proof of Theorem 1 is deferred to Section 5.1.

Remark: First, part (a) of Theorem 1 establishes a convergence rate of $n^{-1/4}$ (up to a logarithmic factor) of $\widehat{G}_n$ to G_0 under the generalized transportation distance $\widetilde{W}_\kappa$ while part (b) of the theorem indicates that this convergence rate is minimax. The convergence rate $n^{-1/4}$ of $\widehat{G}_n$ suggests that the rate of estimating individual components $(\beta_{1i}^0)^{(u)}$ and $(\beta_{2i}^0)^{(v)}$ is $n^{-1/4}$ for $1 \le u \le q_1$ and $1 \le v \le q_2$. The main reason for these slow convergence rates is the singularity of Fisher information matrix for these components. Such a singularity phenomenon is caused by the effect of fitting the true model by larger model, a phenomenon which has been observed previously in traditional mixture models settings under strong identifiability [1, 20].

Second, we would like to emphasize that Theorem 1 is not only of theoretical interest. Indeed, it provides insight into the choice of expert functions that are likely to have favorable convergence in practice. When the expert functions are not algebraically independent, we demonstrate in the next section that the convergence rates of parameter estimation in over-specified GMCF are very slow and depend on a notion of complexity level of over-specification.

4 Algebraically dependent expert functions

In the previous section we established a convergence rate $n^{-1/4}$ for the MLE as well as a minimax lower bound when the expert functions h_1 and h_2 are algebraically independent. In many scenarios, however, the expert functions are taken to be *algebraically dependent*. Here we show that in these settings the convergence rates of the MLE can be much slower than $n^{-1/4}$.

To simplify our proofs in the algebraically-dependent case, we focus on the case in which the MLE is restrained to a parameter space $\mathcal{G}$ that has the following structure:

$$\mathcal{G} = \mathcal{O}_{k,\bar{c}_0}(\Omega) = \{G = \sum_{i=1}^{l} \pi_i \delta_{(\theta_{1i},\theta_{2i})} : 1 \le l \le k \text{ and } \pi_i \ge \bar{c}_0 \ \forall i\}.$$

That is, we consider the set of discrete probability measures with at most k components such that their weights are lower bounded by $\bar{c}_0$ for some given sufficiently small positive number $\bar{c}_0$. Under this assumption, the true but unknown mixing measure $G_0 = \sum_{i=1}^{k_0} \pi_i^0 \delta_{(\theta_{1i}^0,\theta_{2i}^0)} \in \mathcal{E}_{k_0}(\Omega)$ is assumed to have $\pi_i^0 \ge \bar{c}_0$ for $1 \le i \le k_0$.

4.1 Linear expert functions and uniform convergence rates of the MLE

In this section, we consider a few representative examples involving expert functions h_1 and h_2 that are algebraically dependent. We establish the corresponding convergence rates of the MLE for these examples. Our analysis will be divided into two distinct choices for h_2: when h_2 is covariate independent and when h_2 depends on the covariate.

4.1.1 Covariate-independent expert function h_2

We first consider an algebraic dependence setting where the expert function h_2 is independent of the covariate X.

Example 4.1. *Let the expert functions be* $h_1(X|\theta_1) = \theta_1^{(1)} + \theta_1^{(2)} X$ *for all* $\theta_1 = (\theta_1^{(1)}, \theta_1^{(2)}) \in \Omega_1 \subset \mathbb{R}^2$ *and* $h_2^2(X|\theta_2) = \theta_2$ *for all* $\theta_2 \in \Omega_2 \subset \mathbb{R}_+$. *These expert functions* h_1 *and* h_2 *are*

algebraically dependent, as characterized via the following PDE relating h_1 and h_2

$$\left(\frac{\partial h_1}{\partial \theta_1^{(1)}}(X,\theta_1)\right)^2 = \frac{\partial h_2^2}{\partial \theta_2}(X,\theta_2), \tag{5}$$

for all θ_1 and θ_2.

Let $\overline{r} := \overline{r}(k-k_0+1)$ be the minimum value of r such that the following system of polynomial equations:

$$\sum_{j=1}^{k-k_0+1} \sum_{n_1,n_2} \frac{c_j^2 a_j^{n_1} b_j^{n_2}}{n_1! n_2!} = 0 \text{ for each } \alpha = 1,\ldots,r, \tag{6}$$

does not have any nontrivial solution for the unknown variables $(a_j, b_j, c_j)_{j=1}^{k-k_0+1}$. The ranges of n_1 and n_2 in the second sum consist of all natural pairs satisfying the equation $n_1+2n_2=\alpha$. A solution to the above system is considered *nontrivial* if all of variables c_j are non-zeroes, while at least one of the a_j is non-zero.

Our use of the parameter $\overline{r}$ builds on earlier work by [6] who used it to establish convergence rates in the setting of over-specified Gaussian mixtures. The following theorem shows that $\overline{r}$ plays a role in our setting in both the upper bound for the convergence of the MLE and in the minimax lower bound.

Theorem 2. *Given expert functions $h_1(X|\theta_1) = \theta_1^{(1)} + \theta_1^{(2)} X$ for $\theta_1 = (\theta_1^{(1)}, \theta_1^{(2)}) \in \Omega_1 \subset \mathbb{R}^2$ and $h_2^2(X|\theta_2) = \theta_2$ for $\theta_2 \in \Omega_2 \subset \mathbb{R}_+$, the following holds:*

(a) (Maximum likelihood estimation) There exists a positive constant C_0 depending only on G_0 and Ω such that

$$\mathbb{P}\left(\widetilde{W}_\kappa(\widehat{G}_n, G_0) > C_0(\log n/n)^{1/2\overline{r}}\right) \precsim \exp(-c\log n),$$

where $\kappa = (\overline{r}, 2, \lceil \overline{r}/2 \rceil)$ and $\overline{r}$ is defined in (6). Here, c is a positive constant depending only on Ω.

(b) (Minimax lower bound) For any κ' such that $(1,1,1) \preceq \kappa' \prec \kappa = (\overline{r}, 2, \lceil \overline{r}/2 \rceil)$,

$$\inf_{\overline{G}_n} \sup_{G \in \mathcal{O}_k(\Omega) \setminus \mathcal{O}_{k_0-1}(\Omega)} \mathbb{E}_{p_G}\left(\widetilde{W}_{\kappa'}(\overline{G}_n, G)\right) \succsim n^{-1/(2\|\kappa'\|_\infty)}.$$

The proof of Theorem 2 is in Section 5.2.

Remark: First, the convergence rates of MLE in part (a) of Theorem 2 demonstrate that the best possible convergence rates of estimating $(\theta_{1i}^0)^{(1)}$, $(\theta_{1i}^0)^{(2)}$, and θ_{2i}^0 are not uniform. In particular, the rates for estimating $(\theta_{1i}^0)^{(1)}$ and $(\theta_{1i}^0)^{(2)}$ are $n^{-1/2\overline{r}}$ and $n^{-1/4}$, respectively, while the rate for estimating θ_{2i}^0 is $n^{-1/2\lceil \overline{r}/2 \rceil}$ (up to a logarithmic factor) for all $1 \leq i \leq k_0$. Therefore, estimation of the second component of θ_{1i}^0 is generally much faster than estimation of the first component of θ_{1i}^0 and θ_{2i}^0. As is seen in the proof, the slow convergence of $(\theta_{1i}^0)^{(1)}$ and θ_{2i}^0 arises from the way in which the structure of the PDE (5) captures the statistically relevant dependence of the expert functions h_1 and h_2. In particular, the PDE shows that $(\theta_{1i}^0)^{(1)}$ and θ_{2i}^0 are linearly dependent, but, since the second component of θ_{2i}^0 is associated with the covariate X, it does not have any interaction with θ_{2i}^0, which explains why it enjoys a much faster convergence rate than the other parameters.

Second, if we choose expert functions $h_1(X,\theta_1) = \theta_1^{(1)} + \theta_1^{(2)}X + \ldots + \theta_1^{(q_1)}X^{q_1}$ for any $q_1 \geq 2$ and $h_2^2(X,\theta_2) = \theta_2$ where $\theta_1 = (\theta_1^{(1)},\ldots,\theta_1^{(q_1)})$, then with a similar argument we obtain that the best possible convergence rates for estimating $(\theta_{1i}^0)^{(j)}$ for $j \neq 1$ are $n^{-1/4}$ for all $1 \leq i \leq k_0$ while those for $(\theta_{1i}^0)^{(1)}$ and θ_{2i}^0 are $n^{-1/2\overline{r}}$ and $n^{-1/2\lceil \overline{r}/2 \rceil}$, respectively (up to a logarithmic factor).

4.1.2 Covariate-dependent expert function h_2

We now turn to the setting of algebraic dependence between the parameters associated with covariate X in h_1 and the parameters of h_2.

Example 4.2. *Define expert functions $h_1(X,\theta_1) = \theta_1^{(1)} + \theta_1^{(2)}X$ for all $\theta_1 = (\theta_1^{(1)},\theta_1^{(2)}) \in \Omega_1 \subset \mathbb{R}^2$ and $h_2^2(X,\theta_2) = \theta_2^{(1)} + \theta_2^{(2)}X^2$, for all $\theta_2 = (\theta_2^{(1)},\theta_2^{(2)}) \in \Omega_2 \subset \mathbb{R}^2$ such that $\theta_2^{(1)},\theta_2^{(2)} \geq 0$ and $\theta_2^{(1)} + \theta_2^{(2)} \geq \overline{\gamma}$ for some positive constant $\overline{\gamma}$. We have the following PDE for these expert functions:*

$$\left(\frac{\partial h_1}{\partial \theta_1^{(1)}}(X,\theta_1)\right)^2 = \frac{\partial h_2^2}{\partial \theta_2^{(1)}}(X,\theta_2), \tag{7}$$

$$\left(\frac{\partial h_1}{\partial \theta_1^{(2)}}(X,\theta_1)\right)^2 = \frac{\partial h_2^2}{\partial \theta_2^{(2)}}(X,\theta_2), \tag{8}$$

which shows that h_1 and h_2 are algebraically dependent.

The main distinction between Example 4.2 and Example 4.1 is that we have the covariate X^2 in the formulation of the expert function h_2 in Example 4.2. This inclusion leads to a rather rich spectrum of convergence rates for the MLE. To illustrate these convergence rates, we consider two distinct cases for the expert function h_2:

- **without offset**: $\theta_2^{(1)} = 0$, i.e., $h_2^2(X,\theta_2) = \theta_2^{(2)}X^2$.
- **with offset**: $\theta_2^{(1)}$ is taken into account; i.e., $h_2^2(X,\theta_2) = \theta_2^{(1)} + \theta_2^{(2)}X^2$.

Theorem 3. *(Without offset) Let $\overline{r}$ be defined as in (6). Given expert functions $h_1(X,\theta_1) = \theta_1^{(1)} + \theta_2^{(2)}X$ for $\theta_1 = (\theta_1^{(1)},\theta_1^{(2)}) \in \Omega_1 \subset \mathbb{R}^2$ and $h_2^2(X,\theta_2) = \theta_2 X^2$ for $\theta_2 \in \Omega_2 \subset \mathbb{R}_+$, we have:*

(a) (Maximum likelihood estimation) There exists a positive constant C_0 depending only on G_0 and Ω such that

$$\mathbb{P}\left(\widetilde{W}_\kappa(\widehat{G}_n, G_0) > C_0(\log n/n)^{1/2\overline{r}}\right) \precsim \exp(-c\log n),$$

where $\kappa = (2,\overline{r},\lceil \overline{r}/2 \rceil)$. Here, c is a positive constant depending only on Ω.

(b) (Minimax lower bound) For any κ' such that $(1,1,1) \preceq \kappa' \prec \kappa = (2,\overline{r},\lceil \overline{r}/2 \rceil)$,

$$\inf_{\overline{G}_n} \sup_{G \in \mathcal{O}_k(\Omega)\setminus\mathcal{O}_{k_0-1}(\Omega)} \mathbb{E}_{p_G}\left(\widetilde{W}_{\kappa'}(\overline{G}_n, G)\right) \succsim n^{-1/(2\|\kappa'\|_\infty)}.$$

The proof of Theorem 3 is deferred to Appendix A.1.

In contrast to the setting of Theorem 2, the expert function h_2 is now a function of X^2. The convergence rate of $\widehat{G}_n$ in Theorem 3 demonstrates that the convergence rates for estimating

$(\theta_{1i}^0)^{(1)}$, $(\theta_{1i}^0)^{(2)}$, and θ_{2i}^0 are $n^{-1/4}$, $n^{-1/2\overline{r}}$, and $n^{-1/2\lceil\overline{r}/2\rceil}$, respectively, for all $1 \le i \le k_0$. Therefore, with the formulation of expert functions given in Theorem 3, estimation of the first component of θ_{1i}^0 is much faster than estimation of the second component of θ_{1i}^0. This is in contrast to the results in Theorem 2. A high-level explanation for this phenomenon is again obtained by considering the PDE structure, which in this case is given by (8):

$$\left(\frac{\partial h_1}{\partial \theta_1^{(2)}}(X, \theta_1)\right)^2 = \frac{\partial h_2^2}{\partial \theta_2}(X, \theta_2).$$

Such a structure implies the dependence of the second component of θ_{1i}^0 and θ_{2i}^0; therefore, there exists a strong interaction between $(\theta_{1i}^0)^{(2)}$ and θ_{2i}^0 in terms of their convergence rates. On the other hand, the first component of θ_{1i}^0 and θ_2^0 are linearly independent, which implies that there is virtually no interaction between these two terms. As a consequence, $(\theta_{1i}^0)^{(1)}$ will enjoy much faster convergence rates than $(\theta_{1i}^0)^{(2)}$ and θ_{2i}^0.

In contrast to the setting without an offset term in the expert function h_2, the convergence rate of the MLE under the setting with the offset term in h_2 suffers from two ways: one which is captured by the PDE structure with respect to $\theta_1^{(1)}$ and $\theta_2^{(1)}$ in (7) and another from the PDE structure with respect to $\theta_1^{(2)}$ and $\theta_2^{(2)}$ in (8).

Theorem 4. *(With offset) Let $\overline{r}$ be defined as in (6). Given expert functions $h_1(X|\theta_1) = \theta_1^{(1)} + \theta_1^{(2)} X$ for $\theta_1 = (\theta_1^{(1)}, \theta_2^{(2)}) \in \Omega_1 \subset \mathbb{R}^2$ and $h_2^2(X|\theta_2) = \theta_2^{(1)} + \theta_2^{(2)} X^2$ for $\theta_2 = (\theta_2^{(1)}, \theta_2^{(2)}) \in \Omega_2 \subset \mathbb{R}^2$ such that $\theta_2^{(1)}, \theta_2^{(2)} \ge 0$ and $\theta_2^{(1)} + \theta_2^{(2)} \ge \overline{\gamma}$ for some given positive $\overline{\gamma}$, we have:*

(a) (Maximum likelihood estimation) There exists a positive constant C_0 depending only on G_0 and Ω such that

$$\mathbb{P}\left(\widetilde{W}_\kappa(\widehat{G}_n, G_0) > C_0(\log n/n)^{1/2\overline{r}}\right) \precsim \exp(-c \log n), \tag{9}$$

where $\kappa = (\overline{r}, \overline{r}, \lceil\overline{r}/2\rceil, \lceil\overline{r}/2\rceil)$. Here, c is a positive constant depending only on Ω.

(b) (Minimax lower bound) For any κ' such that $(1,1,1,1) \preceq \kappa' \prec \kappa = (\overline{r}, \overline{r}, \lceil\overline{r}/2\rceil, \lceil\overline{r}/2\rceil)$,

$$\inf_{\overline{G}_n} \sup_{G \in \mathcal{O}_k(\Omega) \setminus \mathcal{O}_{k_0-1}(\Omega)} \mathbb{E}_{p_G}\left(\widetilde{W}_{\kappa'}(\overline{G}_n, G)\right) \gtrsim n^{-1/(2\|\kappa'\|_\infty)}.$$

The proof of Theorem 4 is given in Appendix A.2.

Note that when there is an offset term in the expert function h_2, the convergence rate of $\widehat{G}_n$ suggests that the convergence rates for estimating $(\theta_{1i}^0)^{(1)}$, $(\theta_{1i}^0)^{(2)}$, $(\theta_{2i}^0)^{(1)}$, and $(\theta_{2i}^0)^{(2)}$ are $n^{-1/2\overline{r}}$, $n^{-1/2\overline{r}}$, $n^{-1/2\lceil\overline{r}/2\rceil}$, and $n^{-1/2\lceil\overline{r}/2\rceil}$, respectively, for all $1 \le i \le k_0$. In comparison to the convergence rate $n^{-1/4}$ for estimating $(\theta_{1i}^0)^{(2)}$ under the setting without covariate X^2 in h_2 in Theorem 2, the convergence rate $n^{-1/2\overline{r}}$ for estimating $(\theta_{1i}^0)^{(2)}$ under the setting of Theorem 4 is much slower. Furthermore, the convergence rate $n^{-1/2\overline{r}}$ for estimating $(\theta_{1i}^0)^{(2)}$ in the setting of Theorem 4 is much slower than the corresponding rate $n^{-1/4}$ for estimating $(\theta_{1i}^0)^{(2)}$ in the setting of Theorem 3.

Note also that if we choose more general expert functions, $h_1(X, \theta_1) = \theta_1^{(1)} + \theta_1^{(2)} X + \ldots + \theta_1^{(q_1)} X^{q_1}$, for any $q_1 \ge 1$ and $h_2^2(X, \theta_2) = \theta_2^{(1)} + \theta_2^{(2)} X^2 + \ldots + \theta_2^{(q_1)} X^{2q_1}$, where $\theta_1 = (\theta_1^{(1)}, \ldots, \theta_1^{(q_1)})$ and $\theta_2 = (\theta_2^{(1)}, \ldots, \theta_2^{(q_1)})$, i.e., letting $q_2 = q_1$, then we also obtain that the best possible convergence rates for estimating $(\theta_{1i}^0)^{(j)}$ are $n^{-1/2\overline{r}}$ while those for estimating

$(\theta_{2i}^0)^{(j)}$ are $n^{-1/2\lceil \overline{r}/2 \rceil}$ for all $1 \leq i \leq k_0$ and $1 \leq j \leq q_1$. Such results can be explained by the following system of PDEs characterizing the dependence between $\theta_1^{(i)}$ and $\theta_2^{(i)}$ for $1 \leq i \leq q_1$:

$$\left(\frac{\partial h_1}{\partial \theta_1^{(i)}}(X, \theta_1) \right)^2 = \frac{\partial h_2^2}{\partial \theta_2^{(i)}}(X, \theta_2), \text{ for all } 1 \leq i \leq q_1,$$

for any (θ_1, θ_2).

4.2 Nonlinear expert functions and non-uniform convergence rates of MLE

Thus far we have considered various algebraic dependence settings for linear expert functions h_1 and h_2 with respect to their parameters. Under these settings, the convergence rates of the MLE are uniform; i.e., they are independent of the values of the true mixing measure G_0. In this section, we demonstrate that in the case of nonlinear expert functions h_1 and h_2 that are algebraically dependent, the convergence rates of $\widehat{G}_n$ strongly depend on the values of G_0.

The specific setting that we consider is when h_1 is nonlinear in terms of its parameter θ_1 while h_2 is independent of the covariate X. In that setting, we have the following simple example of algebraically dependent expert functions:

$$h_1(X, \theta_1) = \left(\theta_1^{(1)} + \theta_1^{(2)} X \right)^2, \; h_2^2(X|\theta_2) = \theta_2, \tag{10}$$

for all $\theta_1 = (\theta_1^{(1)}, \theta_1^{(2)}) \in \Omega_1 = [0, \overline{\tau}_1] \times [0, \overline{\tau}_2]$ and $\theta_2 \in \Omega_2 \subset \mathbb{R}_+$ where $\overline{\tau}_1$, $\overline{\tau}_2$ are given positive numbers. Here, the choice regarding the ranges of θ_1 is to ensure that the expert function h_1 is identifiable with respect to its parameter θ_1. The following result shows that the expert functions h_1 and h_2 are algebraically dependent.

Proposition 4. *Assume that the expert functions h_1 and h_2 take the forms in (10). Then the expert functions h_1 and h_2 are algebraically dependent, as captured in the following PDE that relates h_1 and h_2:*

$$\left(\frac{\partial h_1}{\partial \theta_1^{(1)}}(X, \theta_1) \right)^2 = 4(\theta_1^{(1)})^2 \frac{\partial h_2^2}{\partial \theta_2}(X, \theta_2), \tag{11}$$

for all $\theta_1 = (\theta_1^{(1)}, 0)$ and θ_2.

Unlike the previous PDEs in (5), (7), and (8), which hold for all (θ_1, θ_2), the PDE in (11) holds only under a special structure for θ_1; namely, $\theta_1 = (\theta^{(1)}, 0)$, where the second component of θ_1 needs to be zero. Such a special structure of the PDE leads to an interesting phase transition regarding the convergence rates of the MLE under specific values of true mixing measure G_0. In order to capture this phase transition precisely, we distinguish two separate settings of G_0:

- **Nonlinearity setting I**: As long as there exists $(\theta_{1i}^0)^{(2)} = 0$ for some $1 \leq i \leq k_0$, we have $(\theta_{1i}^0)^{(1)} = 0$.
- **Nonlinearity setting II**: There exists θ_{1i}^0 such that $(\theta_{1i}^0)^{(1)} \neq 0$ and $(\theta_{1i}^0)^{(2)} = 0$ for some index $1 \leq i \leq k_0$.

4.2.1 Nonlinearity setting I

Under the nonlinearity setting I for the true mixing measure G_0, we have the following result regarding the convergence rate of the MLE.

Theorem 5. *Let the expert functions h_1 and h_2 be defined as in (10). Under the nonlinearity setting I for G_0, the following holds:*

(a) (Maximum likelihood estimation) There exists a positive constant C_0 depending only on G_0 and Ω such that

$$\mathbb{P}(\widetilde{W}_{\kappa}(\widehat{G}_n, G_0) > C_0(\log n/n)^{1/4}) \precsim \exp(-c \log n),$$

where $\kappa = (2, 2, 2)$. Here, c is a positive constant depending only on Ω.

(b) (Minimax lower bound) For any κ' such that $(1, 1, 1) \preceq \kappa' \prec \kappa = (2, 2, 2)$,

$$\inf_{\overline{G}_n \in \mathcal{G}_1} \sup_{G \in \mathcal{G}_1} \mathbb{E}_{p_G}\left(\widetilde{W}_{\kappa'}(\overline{G}_n, G)\right) \gtrsim n^{-1/(2\|\kappa'\|_\infty)},$$

where the structure of the parameter space $\mathcal{G}_1 \subset \mathcal{O}_k(\Omega) \backslash \mathcal{O}_{k_0-1}(\Omega)$ is given by

$$\mathcal{G}_1 = \Big\{ G = \sum_{i=1}^{k'} \pi_i \delta_{(\theta_{1i}, \theta_{2i})} : k_0 \leq k' \leq k \text{ and as long as } \theta_{1i}^{(2)} = 0 \text{ for some } 1 \leq i \leq k', \text{ then } \theta_{1i}^{(1)} = 0 \Big\}.$$

The proof of Theorem 5 is provided in Appendix A.4.

Under the nonlinearity setting I for G_0, the result of Theorem 5 suggests that the convergence rates for estimating $(\theta_{1i}^0)^{(1)}$, $(\theta_{1i}^0)^{(2)}$, and θ_{2i}^0 are $n^{-1/4}$. These convergence rates match those under the settings in which the expert functions h_1 and h_2 are algebraically independent. This phenomenon arises because there is no linkage between θ_{1i}^0 and θ_{2i}^0 in the PDE for the nonlinearity setting I.

4.2.2 Nonlinearity setting II

Unlike the nonlinearity setting I of G_0, the convergence rate of MLE under nonlinearity setting II is more complicated to analyze due to the existence of the zero-valued coefficient $(\theta_{1i}^0)^{(2)}$ for some $1 \leq i \leq k_0$. To simplify the presentation, we first start with a result regarding the structure of the partial derivatives of f when the second component of θ_1 is zero. We then define an *inhomogeneous system of polynomial limits* based on this structural assumption to analyze the behavior of the MLE. Finally, we state a formal convergence rate result of the MLE under the general nonlinearity setting II for G_0.

Partial derivative structures Since there exists a zero-valued coefficient $(\theta_{1i}^0)^{(2)}$ for some $1 \leq i \leq k_0$ under the nonlinearity setting II of G_0, we will focus on understanding the partial derivatives of f when the second component of θ_1 is 0, i.e., $\theta_1^{(2)} = 0$. To facilitate the

discussion, we firstly consider a few specific simple examples of these derivatives:

$$\frac{\partial f}{\partial \theta_1^{(1)}} = 2\theta_1^{(1)}\frac{\partial f}{\partial h_1}, \qquad \frac{\partial f}{\partial \theta_1^{(2)}} = 2\theta_1^{(1)}X\frac{\partial f}{\partial h_1}, \qquad \frac{\partial f}{\partial \theta_2} = \frac{\partial f}{\partial h_2^2} = \frac{1}{2}\frac{\partial^2 f}{\partial h_1^2},$$
$$\frac{\partial^2 f}{\partial (\theta_1^{(1)})^2} = 2\frac{\partial f}{\partial h_1} + 4(\theta_1^{(1)})^2\frac{\partial^2 f}{\partial h_1^2}, \qquad \frac{\partial f}{\partial (\theta_1^{(2)})^2} = 2X^2\frac{\partial f}{\partial h_1} + 4(\theta_1^{(1)})^2X^2\frac{\partial^2 f}{\partial h_1^2},$$
$$\frac{\partial^2 f}{\partial \theta_2^2} = \frac{\partial^2 f}{\partial h_2^4} = \frac{1}{4}\frac{\partial^4 f}{\partial h_1^4}, \qquad \frac{\partial^3 f}{\partial (\theta_1^{(1)})^3} = 12\theta_1^{(1)}\frac{\partial^2 f}{\partial h_1^2} + 8(\theta_1^{(1)})^3\frac{\partial^3 f}{\partial h_1^3}.$$

Here, we suppress the condition on $h_1(X, \theta_1)$ and $h_2(X, \theta_2)$ in the notation to simplify the presentation. From this computation, it is clear that $\frac{\partial f}{\partial \theta_1^{(1)}}$, $\frac{\partial f}{\partial \theta_2}$, and $\frac{\partial^2 f}{\partial (\theta_1^{(1)})^2}$ are not linearly independent with respect to X and Y. This dependence among these partial derivatives underlies the complex behavior of the MLE in this setting.

By iterating this computation of partial derivatives of f up to a high order, we obtain the following key lemma generalizing the structure of partial derivatives of f with respect to $\theta_1^{(1)}$ and θ_2.

Lemma 1. *Assume that $\theta_1^{(2)} = 0$. For any value of $\theta_1^{(1)} \neq 0$, θ_2, and $\gamma = (\gamma_1, \gamma_2) \in \mathbb{N}^2$, the following holds:*

(a) When γ_1 is an odd number, we have:

$$\frac{\partial^{|\gamma|} f}{\partial (\theta_1^{(1)})^{\gamma_1} \partial \theta_2^{\gamma_2}}(Y|h_1(X,\theta_1), h_2(X,\theta_2))$$
$$= \frac{1}{2^{\gamma_2}}\Bigg(\sum_{u=0}^{(\gamma_1-1)/2} P_u^{(\gamma_1)}(\theta_1^{(1)}) \frac{\partial^{\frac{\gamma_1+1}{2}+u+2\gamma_2} f}{\partial h_1^{\frac{\gamma_1+1}{2}+u+2\gamma_2}}(Y|h_1(X,\theta_1), h_2(X,\theta_2)) \Bigg).$$

(b) When γ_1 is an even number, then:

$$\frac{\partial^{|\gamma|} f}{\partial (\theta_1^{(1)})^{\gamma_1} \partial \theta_2^{\gamma_2}}(Y|h_1(X,\theta_1), h_2(X,\theta_2))$$
$$= \frac{1}{2^{\gamma_2}}\Bigg(\sum_{u=0}^{\gamma_1/2} P_u^{(\gamma_1)}(\theta_1^{(1)}) \frac{\partial^{\frac{\gamma_1}{2}+u+2\gamma_2} f}{\partial h_1^{\frac{\gamma_1}{2}+u+2\gamma_2}}(Y|h_1(X,\theta_1), h_2(X,\theta_2)) \Bigg).$$

Here, $P_u^{(\gamma_1)}(\theta_1^{(1)})$ are polynomials in terms of $\theta_1^{(1)}$ that satisfy the following iterative equations:

$$P_0^{(1)}(\theta_1^{(1)}) := 2\theta_1^{(1)}, \;\; P_0^{(\gamma_1+1)}(\theta_1^{(1)}) := \frac{\partial P_0^{(\gamma_1)}}{\partial \theta_1^{(1)}}(\theta_1^{(1)}),$$
$$P_\tau^{(\gamma_1+1)}(\theta_1^{(1)}) := 2\theta_1^{(1)} P_{\tau-1}^{(\gamma_1)}(\theta_1^{(1)}) + \frac{\partial P_{\tau-1}^{(\gamma_1)}}{\partial \theta_1^{(1)}}(\theta_1^{(1)}),$$

for any $1 \leq u \leq (\gamma_1 - 1)/2$ when γ_1 is an odd number or for any $1 \leq u \leq (\gamma_1 - 2)/2$ such that γ_1 is an even number. Additionally, $P_{(\gamma_1+1)/2}^{(\gamma_1+1)}(\theta_1^{(1)}) = 2\theta_1^{(1)} P_{(\gamma_1-1)/2}^{(\gamma_1)}(\theta_1^{(1)})$ if γ_1 is an odd number while $P_{\gamma_1/2}^{(\gamma_1+1)}(\theta_1^{(1)}) = 2\theta_1^{(1)} P_{(\gamma_1-2)/2}^{(\gamma_1)}(\theta_1^{(1)})$ when $\gamma_1 \geq 2$ is an even number.

Inhomogeneous system of polynomial limits Given the specifications of the polynomials $P_\tau^{(\gamma_1)}(\theta_1^{(1)})$ in Lemma 1, we define a system of polynomial limits that is useful for studying convergence rates under the nonlinearity setting II as follows. Assume that we are given $s \in \mathbb{N}$ and $3s$ sequences $\{a_{i,n}\}_{n\geq 1}$, $\{b_{i,n}\}_{n\geq 1}$, and $\{c_{i,n}\}_{n\geq 1}$ such that $a_{i,n} \to 0$, $b_{i,n} \to 0$ as $n \to \infty$ for $1 \leq i \leq s$, while $c_{i,n} \geq 0$ as $1 \leq i \leq s$ and $\sum_{i=1}^{s} c_{i,n} \leq \bar{c}$ for some given $\bar{c} > 0$. For each $\theta_1^{(1)}$ and $r \in \mathbb{N}$, we denote the following inhomogeneous system of polynomial limits:

$$\frac{\sum_{\gamma_1,\gamma_2,u} \frac{P_u^{(\gamma_1)}(\theta_1^{(1)})}{2^{\gamma_2}} \left(\sum_{i=1}^{s} c_{i,n} \frac{a_{i,n}^{\gamma_1} b_{i,n}^{\gamma_2}}{\gamma_1!\gamma_2!} \right)}{\sum_{i=1}^{s} c_{i,n} \left(|a_{i,n}|^r + |b_{i,n}|^{\lceil r/2 \rceil} \right)} \to 0, \tag{12}$$

as $n \to \infty$ for all $1 \leq l \leq 2r$ where the summation with respect to γ_1, γ_2, u in the numerator satisfies $\gamma_1/2 + u + 2\gamma_2 = l$, $u \leq \gamma_1/2$ when γ_1 is an even number while $(\gamma_1+1)/2 + u + 2\gamma_2 = l$, $u \leq (\gamma_1 - 1)/2$ when γ_1 is an odd number. Additionally, $\gamma_1 + \gamma_2 \leq r$.

From these conditions, it is clear that the system of polynomial limits (12) contains exactly $2r$ polynomial limits. For example, when $r = 2$, the system of polynomial limits contains four polynomial limits, which take the following form:

$$\begin{aligned}
\left(\sum_{i=1}^{s} c_{i,n} a_{i,n}^2 + 2\theta_1^{(1)} \sum_{i=1}^{s} c_{i,n} a_{i,n} \right) \Big/ \left(\sum_{i=1}^{s} c_{i,n} \left(|a_{i,n}|^2 + |b_{i,n}| \right) \right) &\to 0, \\
\left(4(\theta_1^{(1)})^2 \left(\sum_{i=1}^{s} c_{i,n} a_{i,n}^2 \right) + \sum_{i=1}^{s} c_{i,n} b_{i,n} \right) \Big/ \left(\sum_{i=1}^{s} c_{i,n} \left(|a_{i,n}|^2 + |b_{i,n}| \right) \right) &\to 0, \\
\theta_1^{(1)} \left(\sum_{i=1}^{s} c_{i,n} a_{i,n} b_{i,n} \right) \Big/ \left(\sum_{i=1}^{s} c_{i,n} \left(|a_{i,n}|^2 + |b_{i,n}| \right) \right) &\to 0, \\
\left(\sum_{i=1}^{s} c_{i,n} b_{i,n}^2 \right) \Big/ \left(\sum_{i=1}^{s} c_{i,n} \left(|a_{i,n}|^2 + |b_{i,n}| \right) \right) &\to 0.
\end{aligned}$$

Studying system of polynomial limits In general, when r is large, the system of polynomial limits (12) does not have a solution; i.e., not all the polynomial limits go to zero. We can therefore find a smallest value of r such that this system of polynomial limits has no solution. This motivates the following definition that plays a key role in obtaining a convergence rate for the MLE.

Definition 4. *For any $s \geq 1$ and $\theta_1^{(1)}$, define $\widetilde{\mathrm{r}}(\theta_1^{(1)}, s)$ as the smallest positive integer r such that system of polynomial limits (12) does not hold for any choices of sequences $\{a_{i,n}\}_{n\geq 1}$, $\{b_{i,n}\}_{n\geq 1}$, and $\{c_{i,n}\}_{n\geq 1}$.*

In general, determining the exact value of $\widetilde{\mathrm{r}}(\theta_1^{(1)}, s)$ is difficult as the system of polynomial limits (12) is intricate. In the following lemma, we demonstrate that we can obtain an upper bound of $\widetilde{\mathrm{r}}(\theta_1^{(1)}, s)$ based on the system of polynomial equations (6), for any $s \geq 1$ and $\theta_1^{(1)} \neq 0$.

Lemma 2. *For a general value of $s \geq 2$ and for all $\theta_1^{(1)} \neq 0$, we have*

$$3 \leq \widetilde{\mathrm{r}}(\theta_1^{(1)}, s) \leq \overline{r}(s),$$

where $\overline{r}(s)$ is defined as in (6).

Convergence rates of MLE Equipped with the definition of $\widetilde{\mathrm{r}}(\theta_1^{(1)}, s)$, we have the following result for the convergence rate of the MLE under the nonlinearity setting II.

Theorem 6. *Given the nonlinearity setting II for G_0 and the expert functions h_1 and h_2 in (10), we define $\mathcal{A} := \{i \in [k_0] : \ (\theta_{1i}^0)^{(1)} \neq 0$ and $(\theta_{1i}^0)^{(2)} = 0\}$ and*

$$i_{max} := \arg\max_{i \in \mathcal{A}} \widetilde{\mathrm{r}}((\theta_{1i}^0)^{(1)}, k - k_0 + 1).$$

Additionally, we denote $\widetilde{\mathrm{r}}_{sin} := \widetilde{\mathrm{r}}((\theta_{1i_{max}}^0)^{(1)}, k - k_0 + 1)$. Then, there exists a positive constant C_0 depending only on G_0 and Ω such that

$$\mathbb{P}(\widetilde{W}_\kappa(\widehat{G}_n, G_0) > C_0(\log n/n)^{1/2\widetilde{\mathrm{r}}_{sin}}) \precsim \exp(-c \log n),$$

where $\kappa = (\widetilde{\mathrm{r}}_{sin}, 2, \lceil \widetilde{\mathrm{r}}_{sin}/2 \rceil)$.

The proof of Theorem 6 is in Appendix A.5.

A few comments are in order. First, the result of Theorem 6 indicates that the convergence rates for estimating $(\theta_{1i}^0)^{(1)}, (\theta_{1i}^0)^{(2)}, \theta_{2i}^0$ are $n^{-1/2\widetilde{\mathrm{r}}_{\mathrm{sin}}}, n^{-1/4}$, and $n^{-1/2\lceil \widetilde{\mathrm{r}}_{\mathrm{sin}}/2 \rceil}$, respectively, for $1 \leq i \leq k_0$. The slow convergence rates of estimating $(\theta_1^0)^{(1)}$ and θ_2^0 under nonlinearity setting II is captured by the PDE (11), which indicates that $(\theta_{1i}^0)^{(1)}$ and θ_{2i}^0 are linearly dependent when the second component of θ_{1i}^0 is zero.

Second, since $\widetilde{\mathrm{r}}_{\mathrm{sin}} \leq \overline{r} = \overline{r}(k - k_0 + 1)$, the convergence rates for estimating $(\theta_{1i}^0)^{(1)}$ and θ_{2i}^0 under the settings of expert functions h_1 and h_2 in (10) may be faster than those of $(\theta_{1i}^0)^{(1)}$ and θ_{2i}^0 under the choice of expert functions h_1 and h_2 in Example 4.1, i.e., $h_1(X|\theta_1) = \theta_1^{(1)} + \theta_1^{(2)} X$ and $h_2^2(X|\theta_2) = \theta_2$. Therefore, parameter estimation when h_1 is quadratic in terms of $\theta_1^{(1)} + \theta_1^{(2)} X$ is generally easier than when h_1 is linear in terms of $\theta_1^{(1)} + \theta_1^{(2)} X$.

General picture In general, if we have an expert function $h_1(X|\theta_1) = (\theta_1^{(1)} + \theta_1^{(2)} X)^m$ for some positive integer $m \geq 1$, and expert function h_2 is independent of covariate X as in (10), then we also have that h_1 and h_2 are algebraically dependent. The corresponding PDE strucure is the following:

$$\left(\frac{\partial h_1}{\partial \theta_1^{(1)}}(X, \theta_1)\right)^2 = m^2 (\theta_1^{(1)})^{2(m-1)} \frac{\partial h_2^2}{\partial \theta_2}(X, \theta_2), \tag{13}$$

for all $\theta_1 = (\theta_1^{(1)}, 0)$ and θ_2. This PDE structure captures a phase transition between nonlinearity setting I and nonlinearity setting II. More precisely, we can check that the convergence rate of $\widehat{G}_n$ will be $n^{-1/4}$, which is similar to that in Theorem 5 under the nonlinearity setting I. Under the nonlinearity setting II, the convergence rates of $\widehat{G}_n$ are again determined by a system of polynomial limits, which is dependent on m and much more complicated than that in (12). A useful insight that arises from these systems is that the convergence rates of $(\theta_{1i}^0)^{(1)}$ and θ_{2i}^0 are better than $n^{-1/2\overline{r}}$ and $n^{-1/2\lceil \overline{r}/2 \rceil}$ respectively while that of $(\theta_{1i}^0)^{(2)}$ is $n^{-1/4}$ for any $1 \leq i \leq k_0$. As a consequence, the convergence rates for parameter estimation when $m \geq 2$ are always better than $m = 1$.

5 Proofs of key results

In this section, we provide the proofs of the key theoretical results in the paper while deferring the rest to the Appendices. Our proof techniques build on previous work for establishing the

convergence rates for parameter estimation under traditional finite mixture models [1, 5, 6] and are based on using a generalized transportation distance to provide controls on various Taylor expansions. We begin with a lemma that presents a general strategy for obtaining convergence rates and minimax lower bounds.

Lemma 3. *(a) (MLE estimation) Assume that there exists some $\kappa \in \mathbb{N}^{q_1+q_2}$ such that*

$$\inf_{G\in\mathcal{G}} h(p_G, p_{G_0})/\widetilde{W}_{\kappa}^{\|\kappa\|_\infty}(G, G_0) > 0, \tag{14}$$

where $\mathcal{G}$ is a subset of $\mathcal{O}_k(\Omega)$ for the over-fitted setting of the GMCF model. Then there exists some positive constant C_0 depending only on G_0 and Ω such that

$$\mathbb{P}(\widetilde{W}_{\kappa}(\widehat{G}_n, G_0) > C_0(\log n/n)^{1/2\|\kappa\|_\infty}) \lesssim \exp(-c\log n),$$

where c is a positive constant depending only on Ω.
(b) (Minimax lower bound) Assume that inequality (14) holds for any $G_0 \in \mathcal{G}$. Furthermore, as long as $G_0 \in \mathcal{G}$, the following holds

$$\inf_{G\in\mathcal{G}} h(p_G, p_{G_0})/W_{\kappa'}^{\|\kappa'\|_\infty}(G, G_0) = 0 \tag{15}$$

for all $\kappa' \prec \kappa$. Then, for any κ' such that $(1, \ldots, 1) \preceq \kappa' \prec \kappa$,

$$\inf_{\overline{G}_n\in\mathcal{G}} \sup_{G\in\mathcal{G}\setminus\mathcal{O}_{k_0-1}(\Omega)} \mathbb{E}_{p_G}\left(\widetilde{W}_{\kappa'}(\overline{G}_n, G)\right) \geq c' n^{-1/(2\|\kappa'\|_\infty)}.$$

Here, $\mathbb{E}_{p_G}$ denotes the expectation taken with respect to product measure with mixture density p_G^n, and c' stands for a universal constant depending on Ω.

The proof of part (a) in Lemma 3 is straightforward from the parametric convergence rate of $h(p_{\widehat{G}_n}, p_{G_0})$ established in Proposition 3. The proof of part (b) in Lemma 3 is a standard application of Le Cam's Lemma utilized in previous work [6]. We therefore omit the proof of Lemma 3 for the brevity of presentation.

5.1 Proof of Theorem 1

Given Lemma 3, we obtain the conclusion of Theorem 1, by demonstrating the following results:

$$\inf_{G\in\mathcal{O}_k(\Omega)} h(p_G, p_{G_0})/\widetilde{W}_{\kappa}^{\|\kappa\|_\infty}(G, G_0) > 0, \tag{16}$$

$$\inf_{G\in\mathcal{O}_k(\Omega)} h(p_G, p_{G_0})/\widetilde{W}_{\kappa'}^{\|\kappa'\|_\infty}(G, G_0) = 0 \tag{17}$$

for any $\kappa' \prec \kappa$ where $\kappa = (2, \ldots, 2)$.

5.1.1 Proof for inequality (16)

The proof of (16) is divided into two parts: local structure and global structure.

Local structure We first demonstrate that inequality (16) holds when $\widetilde{W}_\kappa(G, G_0)$ is sufficiently small. It is equivalent to verify that

$$\lim_{\epsilon\to 0} \inf_{G\in\mathcal{O}_k(\Omega):\widetilde{W}_\kappa(G,G_0)\le\epsilon} h(p_G, p_{G_0})/\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G, G_0) > 0.$$

Due to the standard lower bound $h \ge V$, it is sufficient to show that

$$\lim_{\epsilon\to 0} \inf_{G\in\mathcal{O}_k(\Omega):\widetilde{W}_\kappa(G,G_0)\le\epsilon} V(p_G, p_{G_0})/\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G, G_0) > 0.$$

Assume that the above statement does not hold. This implies that we can find a sequence $G_n \in \mathcal{O}_k(\Omega)$ such that $V(p_{G_n}, p_{G_0})/\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G_n, G_0) \to 0$ and $\widetilde{W}_\kappa(G_n, G_0) \to 0$ as $n \to \infty$. As being demonstrated in Lemma 4 in Appendix B, we can assume the sequence G_n has exactly $\bar{k}$ atoms, where $k_0 \le \bar{k} \le k$, and can be represented as follows:

$$G_n = \sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} p_{ij}^n \delta_{(\theta_{1ij}^n, \theta_{2ij}^n)},$$

where $\bar{l} \ge 0$ is some nonnegative integer and $s_i \ge 1$ for $1 \le i \le k_0 + \bar{l}$ such that $\sum_{i=1}^{k_0+\bar{l}} s_i = \bar{k}$. Additionally, $(\theta_{1ij}^n, \theta_{2ij}^n) \to (\theta_{1i}^0, \theta_{2i}^0)$ and $\sum_{j=1}^{s_i} p_{ij}^n \to \pi_i^0$ for all $1 \le i \le k_0 + \bar{l}$. Here, $\pi_i^0 = 0$ as $k_0 + 1 \le i \le \bar{k}$ while $(\theta_{1i}^0, \ldots, \theta_{di}^0)$ are possible extra limit points from the convergence of components of G_n as $k_0 + 1 \le i \le \bar{k}$.

Now, according to Lemma 5 in Appendix B, we have

$$\begin{aligned}
\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G, G_0) &\gtrsim \sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} p_{ij}^n d_\kappa^{\|\kappa\|_\infty}\left(\eta_{ij}^n, \eta_i^0\right) + \sum_{i=1}^{k_0+\bar{l}} \left|\sum_{j=1}^{s_i} p_{ij}^n - \pi_i^0\right| \\
&= \sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} p_{ij}^n \left(\left\|\theta_{1ij}^n - \theta_{1i}^0\right\|_2^2 + \left\|\theta_{2ij}^n - \theta_{2i}^0\right\|_2^2\right) + \sum_{i=1}^{k_0+\bar{l}} \left|\sum_{j=1}^{s_i} p_{ij}^n - \pi_i^0\right| := D_\kappa(G_n, G_0),
\end{aligned}$$

where $\kappa = (2, \ldots, 2)$, $\eta_i^0 = \left(\theta_{1i}^0, \theta_{2i}^0\right)$ and $\eta_{ij}^n = \left(\theta_{1ij}^n, \theta_{2ij}^n\right)$ for $1 \le i \le k_0$ and $1 \le j \le s_i$. For the simplicity of presentation, we introduce the following notation: $\Delta\theta_{1ij}^n := \theta_{1ij}^n - \theta_{1i}^0$, $\Delta\theta_{2ij}^n := \theta_{2ij}^n - \theta_{2i}^0$ for $1 \le i \le k_0 + \bar{l}$ and $1 \le j \le s_i$. Additionally, we denote $\Delta\theta_{1ij}^n := \left(\left(\Delta\theta_{1ij}^n\right)^{(1)}, \ldots, \left(\Delta\theta_{1ij}^n\right)^{(q_1)}\right)$ and $\Delta\theta_{2ij}^n := \left(\left(\Delta\theta_{2ij}^n\right)^{(1)}, \ldots, \left(\Delta\theta_{2ij}^n\right)^{(q_2)}\right)$ for all i, j.

Since $V(p_{G_n}, p_{G_0})/\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G_n, G_0) \to 0$ as $n \to \infty$, we obtain that $V(p_{G_n}, p_{G_0})/D_\kappa(G_n, G_0) \to 0$. To facilitate the proof argument, we divide it into several steps.

Step 1 - Structure of Taylor expansion By means of a Taylor expansion up to the second order, for any $1 \le i \le k_0 + \bar{l}$ and $1 \le j \le s_i$, the following holds:

$$\begin{aligned}
&f\left(Y|h_1(X, \theta_{1ij}^n), h_2(X, \theta_{2ij}^n)\right) - f\left(Y|h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0)\right) \\
&= \sum_{1\le|\alpha|+|\beta|\le 2} \frac{1}{\alpha!\beta!} \prod_{u=1}^{q_1}\left\{\left(\Delta\theta_{1ij}^n\right)^{(u)}\right\}^{\alpha_u} \prod_{u=1}^{q_1}\left\{\left(\Delta\theta_{2ij}^n\right)^{(v)}\right\}^{\beta_v} \frac{\partial^{|\alpha|+|\beta|} f}{\partial\theta_1^\alpha \partial\theta_2^\beta}\left(Y|h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0)\right) \\
&\qquad + R_{ij}(X, Y),
\end{aligned}$$

where $\alpha = (\alpha_1, \ldots, \alpha_{q_1})$, $\beta = (\beta_1, \ldots, \beta_{q_2})$, $|\alpha| = \alpha_1 + \ldots + \alpha_{q_1}$, and $|\beta| = \beta_1 + \ldots + \beta_{q_2}$. $R_{ij}(X, Y)$ is the remainder from the Taylor expansion and it satisfies

$$R_{ij}(X,Y)\overline{f}(X) = \mathcal{O}\left(\|\Delta\theta_{1ij}\|_2^{2+\gamma} + \|\Delta\theta_{2ij}\|_2^{2+\gamma}\right),$$

for some universal constant $\gamma > 0$ for all $1 \le i \le k_0$ and $1 \le j \le s_i$. We thus have:

$$\begin{aligned}
&p_{G_n}(X,Y) - p_{G_0}(X,Y) \\
&\quad= \sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} p_{ij}^n \left[f(Y|h_1(X,\theta_{1ij}^n), h_2(X,\theta_{2ij}^n) - f(Y|h_1(X,\theta_{1i}^0), h_2(X,\theta_{2i}^0)\right] \overline{f}(X) \\
&\qquad+ \sum_{i=1}^{k_0+\bar{l}} \left(\sum_{j=1}^{s_i} p_{ij}^n - p_i^0 \right) f(Y|h_1(X,\theta_{1i}^0), h_2(X,\theta_{2i}^0))\overline{f}(X) \\
&\quad= \sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} p_{ij}^n \sum_{1 \le |\alpha|+|\beta| \le 2} \frac{1}{\alpha!\beta!} \prod_{u=1}^{q_1} \left\{ (\Delta\theta_{1ij}^n)^{(u)} \right\}^{\alpha_u} \prod_{u=1}^{q_1} \left\{ (\Delta\theta_{2ij}^n)^{(v)} \right\}^{\beta_v} \\
&\qquad\qquad\qquad\qquad \times \frac{\partial^{|\alpha|+|\beta|} f}{\partial\theta_1^\alpha \partial\theta_2^\beta} \left(Y|h_1(X,\theta_{1i}^0), h_2(X,\theta_{2i}^0)\right) \overline{f}(X) \\
&\qquad+ \sum_{i=1}^{k_0+\bar{l}} \left(\sum_{j=1}^{s_i} p_{ij}^n - p_i^0 \right) f(Y|h_1(X,\theta_{1i}^0), h_2(X,\theta_{2i}^0))\overline{f}(X) + R(X,Y) \\
&\quad:= A_n + B_n + R(X,Y),
\end{aligned}$$

where $R(X,Y) = \left(\sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} R_{ij}(X,Y) \right) \overline{f}(X) = \mathcal{O}\left(\sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} p_{ij}^n \left[\|\Delta\theta_{1ij}\|_2^{2+\gamma} + \|\Delta\theta_{2ij}\|_2^{2+\gamma} \right] \right)$. From the formulation of $D_\kappa(G_n, G_0)$, it is clear that

$$R(X,Y)/D_\kappa(G_n,G_0) \precsim \sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} \left[\|\Delta\theta_{1ij}\|_2^\gamma + \|\Delta\theta_{2ij}\|_2^\gamma \right] \to 0 \tag{18}$$

as $n \to \infty$. For the univariate location-scale Gaussian distribution, we have the following characteristic PDE:

$$\frac{\partial^2 f}{\partial\mu^2}(x,\mu,\sigma) = 2\frac{\partial f}{\partial\sigma^2}(x,\mu,\sigma), \tag{19}$$

where μ and σ respectively stand for the location and scale parameter in a location-scale Gaussian distribution. Governed by that PDE, we find that

$$\frac{\partial^2 f}{\partial h_1^2}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right) = 2\frac{\partial f}{\partial h_2^2}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right), \tag{20}$$

for all (θ_1, θ_2). Therefore, for any (θ_1, θ_2), a straightforward calculation yields the following:

$$\begin{aligned}
\frac{\partial f}{\partial\theta_1^{(u)}}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right) &= \frac{\partial h_1}{\partial\theta_1^{(u)}}(X,\theta_1)\frac{\partial f}{\partial h_1}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right), \\
\frac{\partial f}{\partial\theta_2^{(v)}}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right) &= \frac{\partial h_2^2}{\partial\theta_2^{(v)}}(X,\theta_2)\frac{\partial f}{\partial h_2^2}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right) \\
&= \frac{1}{2}\frac{\partial h_2^2}{\partial\theta_2^{(v)}}(X,\theta_2)\frac{\partial^2 f}{\partial h_1^2}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right),
\end{aligned}$$

for all $1 \leq u \leq q_1$ and $1 \leq v \leq q_2$. Similarly, the PDE structure (18) leads to

$$\begin{aligned}
\frac{\partial^2 f}{\partial \theta_1^{(u)} \partial \theta_1^{(v)}}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right) &= \frac{\partial^2 h_1}{\partial \theta_1^{(u)} \partial \theta_1^{(v)}}(X,\theta_1)\frac{\partial f}{\partial h_1}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right) \\
&\quad + \frac{\partial h_1}{\partial \theta_1^{(u)}}(X,\theta_1)\frac{\partial h_1}{\partial \theta_1^{(v)}}(X,\theta_1)\frac{\partial^2 f}{\partial h_1^2}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right), \\
\frac{\partial^2 f}{\partial \theta_2^{(u)} \partial \theta_2^{(v)}}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right) &= \frac{\partial^2 h_2^2}{\partial \theta_2^{(u)} \partial \theta_2^{(v)}}(X,\theta_2)\frac{\partial f}{\partial h_2^2}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right) \\
&\quad + \frac{\partial h_2^2}{\partial \theta_2^{(u)}}(X,\theta_2)\frac{\partial h_2^2}{\partial \theta_2^{(v)}}(X,\theta_2)\frac{\partial^2 f}{\partial h_2^4}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right) \\
&= \frac{1}{2}\frac{\partial^2 h_2^2}{\partial \theta_2^{(u)} \partial \theta_2^{(v)}}(X,\theta_2)\frac{\partial^2 f}{\partial h_1^2}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right) \\
&\quad + \frac{1}{4}\frac{\partial h_2^2}{\partial \theta_2^{(u)}}(X,\theta_2)\frac{\partial h_2^2}{\partial \theta_2^{(v)}}(X,\theta_2)\frac{\partial^4 f}{\partial h_1^4}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right), \\
\frac{\partial^2 f}{\partial \theta_1^{(u)} \partial \theta_2^{(v)}}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right) &= \frac{\partial h_1}{\partial \theta_1^{(u)}}(X,\theta_1)\frac{\partial h_2^2}{\partial \theta_2^{(v)}}(X,\theta_2)\frac{\partial^2 f}{\partial h_1 \partial h_2^2}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right) \\
&= \frac{1}{2}\frac{\partial h_1}{\partial \theta_1^{(u)}}(X,\theta_1)\frac{\partial h_2^2}{\partial \theta_2^{(v)}}(X,\theta_2)\frac{\partial^3 f}{\partial h_1^3}\left(Y|h_1(X,\theta_1), h_2(X,\theta_2)\right),
\end{aligned}$$

for all u, v.

Equipped with the above equations, we can rewrite A_n as follows

$$A_n = \sum_{i=1}^{k_0+\bar{l}} \sum_{\tau=1}^{4} A_{n,\tau}^{(i)}(X)\frac{\partial^\tau f}{\partial h_1^\tau}\left(Y|h_1(X,\theta_{1i}^0), h_2(X,\theta_{2i}^0)\right)\overline{f}(X) := \sum_{i=1}^{k_0+\bar{l}} \overline{A}_{n,\tau}^{(i)}(X,Y),$$

where the explicit forms of $A_{n,\tau}^{(i)}(X)$ are

$$\begin{aligned}
A_{n,1}^{(i)}(X) &:= \sum_{j=1}^{s_i} p_{ij}^n \Bigg(\sum_{u=1}^{q_1} \left(\Delta\theta_{1ij}^n\right)^{(u)} \frac{\partial h_1}{\partial \theta_1^{(u)}}(X,\theta_{1i}^0) \\
&\qquad + \sum_{1\leq u,v\leq q_1} \frac{\left(\Delta\theta_{1ij}^n\right)^{(u)}\left(\Delta\theta_{1ij}^n\right)^{(v)}}{1+1_{\{u=v\}}}\frac{\partial^2 h_1}{\partial \theta_1^{(u)}\partial \theta_1^{(v)}}(X,\theta_{1i}^0)\Bigg), \\
A_{n,2}^{(i)}(X) &:= \sum_{j=1}^{s_i} p_{ij}^n \Bigg\{ \frac{1}{2}\sum_{u=1}^{q_2} \left(\Delta\theta_{2ij}^n\right)^{(u)} \frac{\partial h_2^2}{\partial \theta_2^{(u)}}(X,\theta_{2i}^0) \\
&\qquad + \sum_{1\leq u,v\leq q_1} \frac{\left(\Delta\theta_{1ij}^n\right)^{(u)}\left(\Delta\theta_{1ij}^n\right)^{(v)}}{1+1_{\{u=v\}}}\frac{\partial h_1}{\partial \theta_1^{(u)}}(X,\theta_{1i}^0)\frac{\partial h_1}{\partial \theta_1^{(v)}}(X,\theta_{1i}^0) \\
&\qquad + \frac{1}{2}\sum_{1\leq u,v\leq q_2} \frac{\left(\Delta\theta_{2ij}^n\right)^{(u)}\left(\Delta\theta_{2ij}^n\right)^{(v)}}{1+1_{\{u=v\}}}\frac{\partial^2 h_2^2}{\partial \theta_2^{(u)}\partial \theta_2^{(v)}}(X,\theta_{2i}^0)\Bigg\},
\end{aligned}$$

$$A_{n,3}^{(i)}(X) := \frac{1}{2}\sum_{j=1}^{s_i} p_{ij}^n \sum_{u=1}^{q_1}\sum_{v=1}^{q_2} \left(\Delta\theta_{1ij}^n\right)^{(u)} \left(\Delta\theta_{2ij}^n\right)^{(v)} \frac{\partial h_1}{\partial \theta_1^{(u)}}(X,\theta_{1i}^0)\frac{\partial h_2^2}{\partial \theta_2^{(v)}}(X,\theta_{2i}^0),$$

$$A_{n,4}^{(i)}(X) := \frac{1}{4}\sum_{j=1}^{s_i} p_{ij}^n \sum_{1\leq u,v\leq q_2} \frac{\left(\Delta\theta_{2ij}^n\right)^{(u)} \left(\Delta\theta_{2ij}^n\right)^{(v)}}{1+1_{\{u=v\}}} \frac{\partial h_2^2}{\partial \theta_2^{(u)}}(X,\theta_{2i}^0)\frac{\partial h_2^2}{\partial \theta_2^{(v)}}(X,\theta_{2i}^0).$$

In view of the above computations, we can treat $\overline{A}_{n,\tau}^{(i)}(X,Y)/D_\kappa(G_n,G_0)$ as a linear combinations of elements from $\mathcal{F}_\tau(i)$ for $1\leq\tau\leq 4$, which can be defined as follows:

$$\begin{aligned}
\mathcal{F}_1(i) :=& \left\{ \frac{\partial h_1}{\partial \theta_1^{(u)}}(X,\theta_{1i}^0)\frac{\partial f}{\partial h_1}\left(Y|h_1(X,\theta_{1i}^0),h_2(X,\theta_{2i}^0)\right)\overline{f}(X):\ 1\leq u\leq q_1\right\}\\
&\cup\left\{ \frac{\partial^2 h_1}{\partial \theta_1^{(u)}\partial \theta_1^{(v)}}(X,\theta_{1i}^0)\frac{\partial^2 f}{\partial h_1^2}\left(Y|h_1(X,\theta_{1i}^0),h_2(X,\theta_{2i}^0)\right)\overline{f}(X):\ 1\leq u,v\leq q_1\right\},\\
\mathcal{F}_2(i) :=& \left\{ \frac{\partial h_2^2}{\partial \theta_2^{(u)}}(X,\theta_{2i}^0)\frac{\partial^2 f}{\partial h_1^2}\left(Y|h_1(X,\theta_{1i}^0),h_2(X,\theta_{2i}^0)\right)\overline{f}(X):\ 1\leq u\leq q_2\right\}\\
&\cup\left\{ \frac{\partial h_1}{\partial \theta_1^{(u)}}(X,\theta_{1i}^0)\frac{\partial h_1}{\partial \theta_1^{(v)}}(X,\theta_{1i}^0)\frac{\partial^2 f}{\partial h_1^2}\left(Y|h_1(X,\theta_{1i}^0),h_2(X,\theta_{2i}^0)\right)\overline{f}(X):\ 1\leq u,v\leq q_1\right\}\\
&\cup\left\{ \frac{\partial^2 h_2^2}{\partial \theta_2^{(u)}\partial \theta_2^{(v)}}(X,\theta_{2i}^0)\frac{\partial^2 f}{\partial h_1^2}\left(Y|h_1(X,\theta_{1i}^0),h_2(X,\theta_{2i}^0)\right)\overline{f}(X):\ 1\leq u,v\leq q_2\right\},\\
\mathcal{F}_3(i) :=& \left\{ \frac{\partial h_1}{\partial \theta_1^{(u)}}(X,\theta_{1i}^0)\frac{\partial h_2^2}{\partial \theta_2^{(v)}}(X,\theta_{2i}^0)\frac{\partial^3 f}{\partial h_1^3}\left(Y|h_1(X,\theta_{1i}^0),h_2(X,\theta_{2i}^0)\right)\overline{f}(X):\right.\\
&\qquad\qquad\left. 1\leq u\leq q_1,\ 1\leq v\leq q_2\right\},\\
\mathcal{F}_4(i) :=& \left\{ \frac{\partial h_2^2}{\partial \theta_2^{(u)}}(X,\theta_{2i}^0)\frac{\partial h_2^2}{\partial \theta_2^{(v)}}(X,\theta_{2i}^0)\frac{\partial^4 f}{\partial h_1^4}\left(Y|h_1(X,\theta_{1i}^0),h_2(X,\theta_{2i}^0)\right)\overline{f}(X):\ 1\leq u,v\leq q_2\right\}.
\end{aligned}$$

Therefore, we can view $A_n/D_\kappa(G_n,G_0)$ as a linear combination of elements from $\mathcal{F} := \cup_{i=1}^{k_0+\bar{l}}\cup_{j=1}^4 \mathcal{F}_j(i)$. Similarly, we can view $B_n/D_\kappa(G_n,G_0)$ as a linear combination of elements of the form $f(Y|h_1(X,\theta_{1i}^0),h_2(X,\theta_{2i}^0))\overline{f}(X)$ for $1\leq i\leq k_0+\bar{l}$.

Step 2 - Non-vanishing coefficients Assume that all of the coefficients in the representation of $A_n/D_\kappa(G_n,G_0)$ and $B_n/D_\kappa(G_n,G_0)$ go to 0 as $n\to\infty$. By taking the summation of the absolute values of the coefficients of $B_n/D_\kappa(G_n,G_0)$, the following limit holds

$$\sum_{i=1}^{k_0+\bar{l}}\left|\sum_{j=1}^{s_i} p_{ij}^n - \pi_i^0\right| \Big/ D_\kappa(G_n,G_0)\to 0.$$

From the expression for $D_\kappa(G_n,G_0)$, this yields:

$$\sum_{i=1}^{k_0+\bar{l}}\sum_{j=1}^{s_i} p_{ij}^n\left(\left\|\Delta\theta_{1ij}^n\right\|_2^2+\left\|\Delta\theta_{2ij}^n\right\|_2^2\right)\Big/ D_\kappa(G_n,G_0)\to 1. \tag{21}$$

On the other hand, according to the formulation of $A_{n,4}^{(i)}(X)$, the coefficients associated with the elements $\left(\frac{\partial h_2^2}{\partial \theta_2^{(u)}}(X,\theta_{2i}^0)\right)^2 \frac{\partial^4 f}{\partial h_1^4}$ in $\mathcal{F}_4(i)$ are $\sum_{j=1}^{s_i} p_{ij}^n \left\{\left(\Delta\theta_{2ij}^n\right)^{(u)}\right\}^2 / \left[8D_\kappa(G_n,G_0)\right]$ as $1 \leq i \leq k_0+\bar{l}$ and $1 \leq u \leq q_2$. According to the hypothesis, these coefficients go to zero; therefore, by taking the summation of all of these coefficients, we obtain that

$$\sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} p_{ij}^n \left\|\Delta\theta_{2ij}^n\right\|_2^2 \Big/ D_\kappa(G_n,G_0) \to 0. \tag{22}$$

Furthermore, from the formulation of $A_{n.2}^{(i)}(X)$, we can check that the coefficients attached to the elements $\left(\frac{\partial h_1}{\partial \theta_1^{(u)}}(X,\theta_{1i}^0)\right)^2 \frac{\partial^2 f}{\partial h_1^2}\left(Y|h_1(X,\theta_{1i}^0),h_2(X,\theta_{2i}^0)\right)\overline{f}(X)$ in $\mathcal{F}_2(i)$ are

$$\sum_{j=1}^{s_i} p_{ij}^n \left\{\left(\Delta\theta_{1ij}^n\right)^{(u)}\right\}^2 \Big/ \left[2D_\kappa(G_n,G_0)\right],$$

as $1 \leq i \leq k_0+\bar{l}$ and $1 \leq u \leq q_1$. As all of these coefficients go to zero, by taking the summation of these coefficients, we obtain the following limit:

$$\sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} p_{ij}^n \left\|\Delta\theta_{1ij}^n\right\|_2^2 \Big/ D_\kappa(G_n,G_0) \to 0. \tag{23}$$

Combining the results from (22) and (23), the following limit holds:

$$\sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} p_{ij}^n \left(\left\|\Delta\theta_{1ij}^n\right\|_2^2 + \left\|\Delta\theta_{2ij}^n\right\|_2^2\right) \Big/ D_\kappa(G_n,G_0) \to 0,$$

which is a contradiction to (21). Therefore, not all the coefficients in the representation of $A_n/D_\kappa(G_n,G_0)$ and $B_n/D_\kappa(G_n,G_0)$ go to zero as $n \to \infty$.

Step 3 - Fatou's argument We denote m_n as the maximum of the absolute values of the coefficients in the representation of $A_n/D_\kappa(G_n,G_0)$ and $B_n/D_\kappa(G_n,G_0)$. From here, we define $d_n := 1/m_n$. Since not all the coefficients of $A_n/D_\kappa(G_n,G_0)$ and $B_n/D_\kappa(G_n,G_0)$ vanish, we have $d_n \not\to \infty$ as $n \to \infty$. From the definition of m_n, we denote

$$\left(\sum_{j=1}^{s_i} p_{ij}^n - p_i^0\right)/m_n \to \alpha(i); \quad \left(\sum_{j=1}^{s_i} p_{ij}^n \left(\Delta\theta_{\tau ij}^n\right)^{(u)}\right)/m_n \to \beta_{\tau u}(i),$$

$$\left(\sum_{j=1}^{s_i} p_{ij}^n \left(\Delta\theta_{\tau ij}^n\right)^{(u)} \left(\Delta\theta_{\tau ij}^n\right)^{(v)}\right)/m_n \to \gamma_{\tau uv}(i),$$

$$\left(\sum_{j=1}^{s_i} p_{ij}^n \left(\Delta\theta_{1ij}^n\right)^{(u)} \left(\Delta\theta_{2ij}^n\right)^{(v)}\right)/m_n \to \eta_{uv}(i),$$

as $n \to \infty$ for all $1 \le i \le k_0 + \bar{l}$ and all u, v. Here, at least one among $\alpha(i), \beta_{\tau u}(i), \gamma_{\tau uv}(i)$, and $\eta_{uv}(i)$ is different from zero for all i, u, v. Invoking Fatou's lemma, we have:

$$0 = \lim_{n\to\infty} d_n \frac{V(p_{G_n}, p_{G_0})}{D_\kappa(G_n, G_0)} \ge \int \liminf_{n\to\infty} d_n \frac{|p_{G_n}(X,Y) - p_{G_0}(X,Y)|}{D_\kappa(G_n, G_0)} d(X,Y). \tag{24}$$

From the definition of $\alpha(i), \beta_{\tau u}(i), \gamma_{\tau uv}(i), \eta_{uv}(i)$, the following holds:

$$d_n \frac{p_{G_n}(X,Y) - p_{G_0}(X,Y)}{D_\kappa(G_n, G_0)} \to \sum_{i=1}^{k_0+\bar{l}} \sum_{\tau=0}^{4} E_\tau^{(i)}(X) \frac{\partial^\tau f}{\partial h_1^\tau}\left(Y | h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0)\right) \overline{f}(X), \tag{25}$$

for all (X, Y) where the expressions for $E_\tau^{(i)}(X)$ are:

$$\begin{aligned}
E_0^{(i)}(X) &:= \alpha(i),\ E_1^{(i)}(X) = \sum_{u=1}^{q_1} \beta_{1u}(i) \frac{\partial h_1}{\partial \theta_1^{(u)}}(X, \theta_{1i}^0) + \sum_{1\le u,v\le q_1} \frac{\gamma_{1uv}(i)}{1 + 1_{\{u=v\}}} \frac{\partial^2 h_1}{\partial \theta_1^{(u)} \partial \theta_1^{(v)}}(X, \theta_{1i}^0), \\
E_2^{(i)}(X) &:= \frac{1}{2} \sum_{u=1}^{q_2} \beta_{2u}(i) \frac{\partial h_2^2}{\partial \theta_2^{(u)}}(X, \theta_{2i}^0) + \sum_{1\le u,v\le q_1} \frac{\gamma_{1uv}(i)}{1 + 1_{\{u=v\}}} \frac{\partial h_1}{\partial \theta_1^{(u)}}(X, \theta_{1i}^0) \frac{\partial h_1}{\partial \theta_1^{(v)}}(X, \theta_{1i}^0) \\
&\quad + \frac{1}{2} \sum_{1\le u,v\le q_2} \frac{\gamma_{2uv}(i)}{1 + 1_{\{u=v\}}} \frac{\partial^2 h_2^2}{\partial \theta_2^{(u)} \partial \theta_2^{(v)}}(X, \theta_{2i}^0), \\
E_3^{(i)}(X) &:= \frac{1}{2} \sum_{u=1}^{q_1} \sum_{v=1}^{q_2} \eta_{uv}(i) \frac{\partial h_1}{\partial \theta_1^{(u)}}(X, \theta_{1i}^0) \frac{\partial h_2^2}{\partial \theta_2^{(v)}}(X, \theta_{2i}^0), \\
E_4^{(i)}(X) &:= \frac{1}{4} \sum_{1\le u,v\le q_2} \frac{\gamma_{2uv}(i)}{1 + 1_{\{u=v\}}} \frac{\partial h_2^2}{\partial \theta_2^{(u)}}(X, \theta_{2i}^0) \frac{\partial h_2^2}{\partial \theta_2^{(v)}}(X, \theta_{2i}^0).
\end{aligned}$$

Combining the results from (24) and (25), the following equation holds

$$\sum_{i=1}^{k_0+\bar{l}} \sum_{\tau=0}^{4} E_\tau^{(i)}(X) \frac{\partial^\tau f}{\partial h_1^\tau}\left(Y | h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0)\right) \overline{f}(X) = 0,$$

almost surely (X, Y). For each X, the set

$$\left\{ \frac{\partial^\tau f}{\partial h_1^\tau}\left(Y | h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0)\right) : 0 \le \tau \le 4 \right\}$$

is linearly independent with respect to Y. Therefore, the above equation eventually leads to $E_\tau^{(i)}(X) = 0$ almost surely X for $0 \le \tau \le 4$ and $1 \le i \le k_0 + \bar{l}$. When $\tau = 0$, it is clear that the equation $E_\tau^{(i)}(X) = 0$ almost surely X demonstrates that $\alpha(i) = 0$ for all i. When $\tau \ge 3$, the equations $E_\tau^{(i)}(X) = 0$ almost surely X lead to $\gamma_{2uv} = 0$ and $\eta_{uv} = 0$ for all (u, v). Invoking the fact that the expert functions h_1 and h_2 are algebraically independent and the result that $\gamma_{2uv} = 0$ for all (u, v), the equation $E_2^{(i)}(X) = 0$ almost surely X implies that $\beta_{2u}(i) = 0$ and $\gamma_{1uv}(i) = 0$ for all i and (u, v). Collecting the previous results, the equation $E_1^{(i)}(X) = 0$ almost surely X leads to $\beta_{1u}(i) = 0$ for all i and u. Therefore, all the coefficients $\alpha(i), \beta_{\tau u}(i), \gamma_{\tau uv}(i)$, and $\eta_{uv}(i)$ are equal to zero for all i and u, v, which is a contradiction.

As a consequence, we can find some $\epsilon_0 > 0$ such that

$$\inf_{G \in \mathcal{O}_k(\Omega): \widetilde{W}_\kappa(G, G_0) \le \epsilon_0} h(p_G, p_{G_0}) / \widetilde{W}_\kappa^{\|\kappa\|_\infty}(G, G_0) > 0.$$

Global structure Given the local bound that we have just established, to obtain the conclusion of inequality (16), it is sufficient to demonstrate that

$$\inf_{G\in\mathcal{O}_k(\Omega):\widetilde{W}_\kappa(G,G_0)>\epsilon_0} h(p_G,p_{G_0})/\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G,G_0)>0.$$

Assume that the above result does not hold. This indicates that we can find a sequence $\overline{G}_n\in\mathcal{O}_k(\Omega)$ such that $h(p_{\overline{G}_n},p_{G_0})/\widetilde{W}_\kappa^{\|\kappa\|_\infty}(\overline{G}_n,G_0)\to 0$ as $n\to\infty$ while $\widetilde{W}_\kappa(\overline{G}_n,G_0)>\epsilon_0$ for all $n\geq 1$. Since the set Ω is bounded, there exists a subsequence of G_n such that $G_n\to G'$ for some mixing measure $G'\in\mathcal{O}_k(\Omega)$. To facilitate the discussion, we replace this subsequence by the whole sequence of G_n. Then, as $\widetilde{W}_\kappa(\overline{G}_n,G_0)>\epsilon_0$ for all $n\geq 1$, this implies that $\widetilde{W}_\kappa(G',G_0)\geq\epsilon_0$. Combining the previous bound with $h(p_{\overline{G}_n},p_{G_0})/\widetilde{W}_\kappa^{\|\kappa\|_\infty}(\overline{G}_n,G_0)\to 0$, we obtain that $h(p_{\overline{G}_n},p_{G_0})\to 0$ as $n\to\infty$. Invoking Fatou's lemma, the following inequality holds:

$$\begin{aligned}0=\lim_{n\to\infty}h^2(p_{\overline{G}_n},p_{G_0})&\geq\frac{1}{2}\int\liminf_{n\to\infty}\left(\sqrt{p_{\overline{G}_n}(X,Y)}-\sqrt{p_{G_0}(X,Y)}\right)^2d(X,Y)\\&=\frac{1}{2}\int\left(\sqrt{p_{G'}(X,Y)}-\sqrt{p_{G_0}(X,Y)}\right)^2d(X,Y).\end{aligned}$$

This inequality leads to $p_{G'}(X,Y)=p_{G_0}(X,Y)$ for almost surely X,Y. Due to the identifiability of GMCF, this leads to $G'\equiv G_0$, which is a contradiction to the result that $\widetilde{W}_\kappa(G',G_0)\geq\epsilon_0>0$. Hence, we achieve the conclusion of inequality (16).

5.1.2 Proof for equality (17)

To achieve the conclusion of equality (17), it is equivalent to find a sequence $G_n\in\mathcal{O}_k(\Omega)$ such that $h(p_{G_n},p_{G_0})/\widetilde{W}_{\kappa'}^{\|\kappa'\|_\infty}(G_n,G_0)\to 0$ as $n\to\infty$ for every $\kappa'\prec\kappa$. In fact, for any $\kappa'\prec\kappa$, we have $\min_{1\leq i\leq k_0}\kappa'^{(i)}<2$. Without loss of generality, we assume $\kappa'^{(1)}=\min_{1\leq i\leq k_0}\kappa'^{(i)}<2$. Now, we construct a sequence of mixing measures, $G_n=\sum_{i=1}^{k_0+1}\pi_i^n\delta_{(\theta_{1i}^n,\theta_{2i}^n)}$, with k_0+1 components as follows: $(\pi_i^n,\theta_{1i}^n,\theta_{2i}^n)\equiv(\pi_{i-1}^0,\theta_{1(i-1)}^0,\theta_{2(i-1)}^0)$ for $3\leq i\leq k_0+1$. Additionally, $\pi_1^n=\pi_2^n=1/2$, $(\theta_{11}^n,\theta_{21}^n)\equiv(\theta_{11}^0-\mathbf{1}_{q_1}/n,\theta_{21}^0-\mathbf{1}_{q_2}/n)$, and $(\theta_{12}^n,\theta_{22}^n)\equiv(\theta_{11}^0+\mathbf{1}_{q_1}/n,\theta_{21}^0+\mathbf{1}_{q_2}/n)$. Now, by means of Taylor expansion up to the first order, we have

$$\begin{aligned}p_{G_n}(X,Y)-p_{G_0}(X,Y)&=\sum_{i=1}^2\pi_i^n\left(f(Y|h_1(X,\theta_{1i}^n,\theta_{2i}^n)-f(Y|h_1(X,\theta_{11}^0,\theta_{21}^0)\right)\overline{f}(X)\\&=\sum_{i=1}^2\pi_i^n\sum_{|\alpha|+|\beta|=1}\frac{1}{\alpha!\beta!}\prod_{u=1}^{q_1}\left\{(\Delta\theta_{1i}^n)^{(u)}\right\}^{\alpha_u}\prod_{u=1}^{q_1}\left\{(\Delta\theta_{2i}^n)^{(v)}\right\}^{\beta_v}\\&\times\frac{\partial f}{\partial\theta_1^\alpha\partial\theta_2^\beta}\left(Y|h_1(X,\theta_{11}^0),h_2(X,\theta_{21}^0)\right)\overline{f}(X)+\overline{R}(X,Y),\end{aligned}$$

where $\Delta\theta_{1i}^n=\theta_{1i}^n-\theta_{11}^0$ and $\Delta\theta_{2i}^n=\theta_{2i}^n-\theta_{21}^0$ for $1\leq i\leq 2$. Here $\overline{R}(X,Y)$ is a Taylor remainder from the above expansion. With the choice of π_i^n,θ_{1i}^n, and θ_{2i}^n for $1\leq i\leq 2$, we can verify that:

$$\sum_{i=1}^2\pi_i^n\prod_{u=1}^{q_1}\left\{(\Delta\theta_{1i}^n)^{(u)}\right\}^{\alpha_u}\prod_{u=1}^{q_1}\left\{(\Delta\theta_{2i}^n)^{(v)}\right\}^{\beta_v}=0,$$

for all $|\alpha|+|\beta|=1$. Therefore, we have the following representation

$$p_{G_n}(X,Y)-p_{G_0}(X,Y)=\overline{R}(X,Y),$$

where the explicit form of the Taylor remainder $\overline{R}(X,Y)$ is as follows:

$$\begin{aligned}\overline{R}(X,Y)=&\sum_{i=1}^{2}\pi_i^n\sum_{|\alpha|+|\beta|=2}\frac{2}{\alpha!\beta!}\prod_{u=1}^{q_1}\Big\{(\Delta\theta_{1i}^n)^{(u)}\Big\}^{\alpha_u}\prod_{u=1}^{q_1}\Big\{(\Delta\theta_{2i}^n)^{(v)}\Big\}^{\beta_v}\\&\times\int_0^1(1-t)\frac{\partial^2 f}{\partial\theta_1^{\alpha}\partial\theta_2^{\beta}}\left(Y|h_1(X,\theta_{11}^0+t\Delta\theta_{1i}^n),h_2(X,\theta_{21}^0+t\Delta\theta_{2i}^n)\right)\overline{f}(X)dt.\end{aligned}$$

From the properties of a univariate location-scale Gaussian distribution, it is not hard to verify that

$$T_{\alpha,\beta}=\sup_{t\in[0,1]}\int\frac{\left(\frac{\partial^2 f}{\partial\theta_1^{\alpha}\partial\theta_2^{\beta}}\left(Y|h_1(X,\theta_{11}^0+t\Delta\theta_{1i}^n),h_2(X,\theta_{21}^0+t\Delta\theta_{2i}^n)\right)\right)^2}{f(Y|h_1(X,\theta_{11}^0),h_2(X,\theta_{21}^0)}d(X,Y)<\infty,\quad(26)$$

for all $|\alpha|+|\beta|=2$. Additionally, the expressions for θ_{1i}^n and θ_{2i}^n indicate that

$$F_{\alpha,\beta}=\sum_{i=1}^{2}\pi_i^n\frac{2}{\alpha!\beta!}\prod_{u=1}^{q_1}\Big\{(\Delta\theta_{1i}^n)^{(u)}\Big\}^{\alpha_u}\prod_{u=1}^{q_1}\Big\{(\Delta\theta_{2i}^n)^{(v)}\Big\}^{\beta_v}=\mathcal{O}(n^{-2}),\quad(27)$$

for all $|\alpha|+|\beta|=2$. Now a direct computation yields that

$$\begin{aligned}\frac{h^2(p_{G_n},p_{G_0})}{\widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n,G_0)}&=\frac{1}{2}\int\frac{(p_{G_n}(X,Y)-p_{G_0}(X,Y))^2}{\left(\sqrt{p_{G_n}(X,Y)}+\sqrt{p_{G_0}(X,Y)}\right)^2\widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n,G_0)}d(X,Y)\\&\leq\frac{1}{2}\int\frac{\overline{R}^2(X,Y)}{p_{G_0}(X,Y)\widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n,G_0)}d(X,Y).\end{aligned}$$

The Cauchy-Schwartz inequality implies that the following inequality holds:

$$\int\frac{\overline{R}^2(X,Y)}{p_{G_0}(X,Y)}d(X,Y)\precsim\sum_{|\alpha|+|\beta|=2}T_{\alpha,\beta}F_{\alpha,\beta}^2=\mathcal{O}(n^{-4}),$$

where the final bound comes from the bounds on $T_{\alpha,\beta}$ and $F_{\alpha,\beta}$ in (26) and (27). On the other hand, the choice of G_n guarantees that $\widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n,G_0)=\mathcal{O}(n^{-2\kappa'^{(1)}})$ as $\kappa'^{(1)}=\min_{1\leq i\leq k_0}\kappa'^{(i)}$. Since $\kappa'^{(1)}<2$, it is clear that

$$\int\frac{\overline{R}^2(X,Y)}{p_{G_0}(X,Y)\widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n,G_0)}d(X,Y)\to 0,$$

as $n\to\infty$. Therefore, $h^2(p_{G_n},p_{G_0})\Big/\widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n,G_0)\to 0$. As a consequence, we obtain the conclusion of equality (17).

5.2 Proof of Theorem 2

By means of Lemma 3, we prove Theorem 2 by establishing the following results:

$$\inf_{G\in\mathcal{O}_{k,\bar{c}_0}(\Omega)} h(p_G, p_{G_0})/\widetilde{W}_{\kappa}^{\|\kappa\|_\infty}(G, G_0) > 0, \tag{28}$$

$$\inf_{G\in\mathcal{O}_{k}(\Omega)} h(p_G, p_{G_0})/\widetilde{W}_{\kappa'}^{\|\kappa'\|_\infty}(G, G_0) = 0, \tag{29}$$

for any $\kappa' \prec \kappa$ where $\kappa = (\overline{r}, 2, \lceil \overline{r}/2 \rceil)$. To simplify the presentation, we assume that $\overline{r}$ is an even number throughout this proof, which leads to $\kappa = (\overline{r}, 2, \overline{r}/2)$. The proof when $\overline{r}$ is an odd number can be obtained in a similar fashion.

5.2.1 Proof for inequality (28)

To streamline the argument, we provide a proof only for the local structural inequality:

$$\lim_{\epsilon\to 0} \inf_{G\in\mathcal{O}_{k,\bar{c}_0}(\Omega):\widetilde{W}_\kappa(G,G_0)\le\epsilon} V(p_G, p_{G_0})/\widetilde{W}_{\kappa}^{\|\kappa\|_\infty}(G, G_0) > 0;$$

the global structural result, for inequality (28), can be argued in a similar fashion as in the proof of Theorem 1. Assume now that the local structure inequality does not hold. This implies that we can find a sequence $G_n \in \mathcal{O}_{k,\bar{c}_0}(\Omega)$ such that $V(p_{G_n}, p_{G_0})/\widetilde{W}_{\kappa}^{\|\kappa\|_\infty}(G_n, G_0) \to 0$ and $\widetilde{W}_\kappa(G_n, G_0) \to 0$ as $n \to \infty$. Employing the similar argument as in Theorem 1 in Section 5.1.1, we can represent the sequence G_n as follows:

$$G_n = \sum_{i=1}^{k_0}\sum_{j=1}^{s_i} p_{ij}^n \delta_{(\theta_{1ij}^n, \theta_{2ij}^n)}, \tag{30}$$

where $(\theta_{1ij}^n, \theta_{2ij}^n) \to (\theta_{1i}^0, \theta_{2i}^0)$ for all $1 \le i \le k_0, 1 \le j \le s_i$ and $\sum_{j=1}^{s_i} p_{ij}^n \to \pi_i^0$ for all $1 \le i \le k_0$. Note that we do not have $\bar{l}$ in the representation of $G_n \in \mathcal{O}_{k,\bar{c}_0}(\Omega)$, in contrast to the result in Section 5.1.1. The reason is that the weights of G_n are lower bounded by a positive number $\bar{c}_0$, which entails that there exists no extra components $(\theta_{1i}^0, \theta_{2i}^0)$ as the limit points of the components of G_n. In this proof, for the simplicity of presentation, we denote $\Delta\theta_{1ij}^n = \theta_{1ij}^n - \theta_{1i}^0$ and $\Delta\theta_{2ij}^n = \theta_{2ij}^n - \theta_{2i}^0$ for all $1 \le i \le k_0, 1 \le j \le s_i$. Additionally, $\Delta\theta_{1ij}^n = ((\Delta\theta_{1ij}^n)^{(1)}, (\Delta\theta_{1ij}^n)^{(2)})$ for all $1 \le i \le k_0, 1 \le j \le s_i$. Now, according to Lemma 5 in Appendix B, we have:

$$\begin{aligned}\widetilde{W}_{\kappa}^{\|\kappa\|_\infty}(G_n, G_0) \gtrsim \sum_{i=1}^{k_0}\sum_{j=1}^{s_i} p_{ij}^n \left(\left|(\Delta\theta_{1ij}^n)^{(1)}\right|^{\overline{r}} + \left|(\Delta\theta_{1ij}^n)^{(2)}\right|^2 + |\Delta\theta_{2ij}^n|^{\overline{r}/2}\right) \\ + \sum_{i=1}^{k_0} |\sum_{j=1}^{s_i} p_{ij}^n - \pi_i^0| := D_\kappa(G_n, G_0),\end{aligned}$$

where $\kappa = (\overline{r}, 2, \overline{r}/2)$. Since $V(p_{G_n}, p_{G_0})/\widetilde{W}_{\kappa}^{\|\kappa\|_\infty}(G, G_0) \to 0$, we have $V(p_{G_n}, p_{G_0})/D_\kappa(G_n, G_0) \to 0$. We again divide our proof argument into several steps.

Step 1 - Structure of Taylor expansion Using the decomposition $p_{G_n}(X,Y)-p_{G_0}(X,Y)$, as in the proof of inequality (16) in Section 5.1.1, we carry out a Taylor expansion up to the order $\overline{r}$:

$$\begin{aligned} p_{G_n}(X,Y) - p_{G_0}(X,Y) &= \sum_{i=1}^{k_0}\sum_{j=1}^{s_i} p_{ij}^n \sum_{1\leq|\alpha|\leq\overline{r}} \frac{1}{\alpha!}\Big\{(\Delta\theta_{1ij}^n)^{(1)}\Big\}^{\alpha_1}\Big\{(\Delta\theta_{1ij}^n)^{(2)}\Big\}^{\alpha_2}(\Delta\theta_{2ij}^n)^{\alpha_3} \\ &\quad\times\frac{\partial^{|\alpha|}f}{\partial(\theta_1^{(1)})^{\alpha_1}\partial(\theta_1^{(2)})^{\alpha_2}\partial\theta_2^{\alpha_3}}\big(Y|h_1(X,\theta_{1i}^0),h_2(X,\theta_{2i}^0)\big)\,\overline{f}(X) \\ &\quad+\sum_{i=1}^{k_0}\Big(\sum_{j=1}^{s_i}p_{ij}^n-p_i^0\Big)f(Y|h_1(X,\theta_{1i}^0),h_2(X,\theta_{2i}^0))\overline{f}(X)+R(X,Y) \\ &:= A_n + B_n + R(X,Y), \end{aligned} \tag{31}$$

where $R(X,Y)$ is a remainder term. This remainder term is such that $R(X,Y)/D_\kappa(G_n,G_0)\to 0$ as $n\to\infty$, is due to the uniform Hölder continuity of a location-scale Gaussian family with respect to expert functions h_1, h_2, and prior density $\overline{f}$ (cf. Proposition 2).

From the formulation of the expert functions h_1, h_2 as well as the structural form of the PDE for location-scale Gaussian kernel, we obtain the following:

$$\frac{\partial^{|\alpha|}f}{\partial(\theta_1^{(1)})^{\alpha_1}\partial(\theta_1^{(2)})^{\alpha_2}\partial\theta_2^{\alpha_3}}(Y|h_1(X,\theta_1),h_2(X,\theta_2)) = \frac{X^{\alpha_2}}{2^{\alpha_3}}\frac{\partial^{\alpha_1+\alpha_2+2\alpha_3}f}{\partial h_1^{\alpha_1+\alpha_2+2\alpha_3}}(Y|h_1(X,\theta_1),h_2(X,\theta_2)),$$

for any $\alpha_1,\alpha_2,\alpha_3\in\mathbb{N}$, $\theta_1\in\Omega_1$, and $\theta_2\in\Omega_2$. From this equation, we can rewrite A_n as follows:

$$\begin{aligned} A_n &= \sum_{i=1}^{k_0}\sum_{j=1}^{s_i} p_{ij}^n \sum_{1\leq|\alpha|\leq\overline{r}} \frac{1}{\alpha!}\Big\{(\Delta\theta_{1ij}^n)^{(1)}\Big\}^{\alpha_1}\Big\{(\Delta\theta_{1ij}^n)^{(2)}\Big\}^{\alpha_2}(\Delta\theta_{2ij}^n)^{\alpha_3} \\ &\quad\times\frac{X^{\alpha_2}}{2^{\alpha_3}}\frac{\partial^{\alpha_1+\alpha_2+2\alpha_3}f}{\partial h_1^{\alpha_1+\alpha_2+2\alpha_3}}(Y|h_1(X,\theta_{1i}^0),h_2(X,\theta_{2i}^0))\overline{f}(X) \\ &= \sum_{i=1}^{k_0}\sum_{j=1}^{s_i} p_{ij}^n \sum_{\alpha_2=0}^{\overline{r}}\sum_{l=0}^{2(\overline{r}-\alpha_2)}\sum_{\alpha_1,\alpha_3}\frac{1}{2^{\alpha_3}\alpha!}\Big\{(\Delta\theta_{1ij}^n)^{(1)}\Big\}^{\alpha_1}\Big\{(\Delta\theta_{1ij}^n)^{(2)}\Big\}^{\alpha_2}(\Delta\theta_{2ij}^n)^{\alpha_3} \\ &\quad\times X^{\alpha_2}\frac{\partial^{l+\alpha_2}f}{\partial h_1^{l+\alpha_2}}(Y|h_1(X,\theta_{1i}^0),h_2(X,\theta_{2i}^0))\overline{f}(X), \end{aligned} \tag{32}$$

where $\alpha_1,\alpha_3\in\mathbb{N}$ in the sum of the second equation satisfies $\alpha_1+2\alpha_3=l$ and $1-\alpha_2\leq\alpha_1+\alpha_3\leq\overline{r}-\alpha_2$. We define

$$\mathcal{F}:=\Big\{X^{\alpha_2}\frac{\partial^{l+\alpha_2}f}{\partial h_1^{l+\alpha_2}}(Y|h_1(X,\theta_{1i}^0),h_2(X,\theta_{2i}^0))\overline{f}(X):\ 0\leq\alpha_2\leq\overline{r},\ 0\leq l\leq 2(\overline{r}-\alpha_2),\ 1\leq i\leq k_0\Big\}.$$

We claim that the elements of $\mathcal{F}$ are linearly independent with respect to X and Y. We prove this claim at the end of this proof. Assume that this claim is given at the moment. Inspecting the explicit form of $\mathcal{F}$, we can treat $A_n/D_\kappa(G_n,G_0)$, $B_n/D_\kappa(G_n,G_0)$ as a linear combination of elements of $\mathcal{F}$.

Step 2 - Non-vanishing coefficients To simplify the proof, we denote $E_{\alpha_2,l}(\theta_{1i}^0, \theta_{2i}^0)$ as the coefficient of $X^{\alpha_2}\dfrac{\partial^{l+\alpha_2} f}{\partial h_1^{l+\alpha_2}}(Y|h_1(X|\theta_{1i}^0), h_2(X|\theta_{2i}^0))\overline{f}(X)$ in A_n and B_n for any $0 \leq \alpha_2 \leq \overline{r}$, $0 \leq l \leq 2(\overline{r} - \alpha_2)$, and $1 \leq i \leq k_0$. Then, the coefficients associated with $X^{\alpha_2}\dfrac{\partial^{l+\alpha_2} f}{\partial h_1^{l+\alpha_2}}(Y|h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0))\overline{f}(X)$ in $A_n/D_\kappa(G_n, G_0)$ and $B_n/D_\kappa(G_n, G_0)$ take the form $E_{\alpha_2,l}(\theta_{1i}^0, \theta_{2i}^0)/D_\kappa(G_n, G_0)$.

Assume that all of the coefficients in the representation of $A_n/D_\kappa(G_n, G_0)$, $B_n/D_\kappa(G_n, G_0)$ go to zero as $n \to \infty$. By taking the summation of $|E_{0,0}(\theta_{1i}^0, \theta_{2i}^0)/D_\kappa(G_n, G_0)|$ for all $1 \leq i \leq k_0$, we obtain that

$$\left(\sum_{i=1}^{k_0} |\sum_{j=1}^{s_i} p_{ij}^n - \pi_i^0|\right)/D_\kappa(G_n, G_0) \to 0.$$

Additionally, according to equation (32), we can verify that

$$\frac{E_{2,0}(\theta_{1i}^0, \theta_{2i}^0)}{D_\kappa(G_n, G_0)} = \frac{\sum_{j=1}^{s_i} p_{ij}^n \left|(\Delta\theta_{1ij}^n)^{(2)}\right|^2}{D_\kappa(G_n, G_0)} \to 0,$$

for all $1 \leq i \leq k_0$. From the formulation of $D_\kappa(G_n, G_0)$, the above limits lead to

$$\left\{\sum_{i=1}^{k_0}\sum_{j=1}^{s_i} p_{ij}^n \left(\left|(\Delta\theta_{1ij}^n)^{(1)}\right|^{\overline{r}} + |\Delta\theta_{2ij}^n|^{\overline{r}/2}\right)\right\}/D_\kappa(G_n, G_0) \to 1 \text{ as } n \to \infty.$$

Therefore, we can find an index $i^* \in \{1, \ldots, k_0\}$ such that

$$L = \left\{\sum_{j=1}^{s_i} p_{i^*j}^n \left(\left|(\Delta\theta_{1i^*j}^n)^{(1)}\right|^{\overline{r}} + |\Delta\theta_{2i^*j}^n|^{\overline{r}/2}\right)\right\}/D_\kappa(G_n, G_0) \not\to 0,$$

as $n \to \infty$. Without loss of generality, we assume that $i^* = 1$. Now, since we have $E_{\alpha_2,l}(\theta_{1i}^0, \theta_{2i}^0)/D_\kappa(G_n, G_0) \to 0$ for all values of α_2, l, i, we obtain that

$$M_{\alpha_2,l}(\theta_{1i}^0, \theta_{2i}^0) = \frac{E_{\alpha_2,l}(\theta_{1i}^0, \theta_{2i}^0)}{\sum_{j=1}^{s_i} p_{1j}^n \left(\left|(\Delta\theta_{11j}^n)^{(1)}\right|^{\overline{r}} + |\Delta\theta_{21j}^n|^{\overline{r}/2}\right)} = \frac{1}{L}\frac{E_{\alpha_2,l}(\theta_{1i}^0, \theta_{2i}^0)}{D_\kappa(G_n, G_0)}, \to 0$$

for any $0 \leq \alpha_2 \leq \overline{r}$, $0 \leq l \leq 2(\overline{r} - \alpha_2)$, and $1 \leq i \leq k_0$. By the representation of A_n in (32), we can verify that

$$M_{0,l}(\theta_{11}^0, \theta_{21}^0) = \frac{\sum_{j=1}^{s_1} p_{1j}^n \sum_{\substack{\alpha_1+2\alpha_3=l \\ \alpha_1+\alpha_3 \leq \overline{r}}} \dfrac{\left\{(\Delta\theta_{11j}^n)^{(1)}\right\}^{\alpha_1} (\Delta\theta_{21j}^n)^{\alpha_3}}{2^{\alpha_3}\alpha_1!\alpha_3!}}{\sum_{j=1}^{s_i} p_{1j}^n \left(\left|(\Delta\theta_{11j}^n)^{(1)}\right|^{\overline{r}} + |\Delta\theta_{21j}^n|^{\overline{r}/2}\right)} \to 0.$$

Step 3 - Understanding the system of polynomial limits The technique for studying the above system of polynomial limits is similar to that of Step 1 in the proof of Proposition 3.3 in [6]. Here, we briefly sketch the proof for completeness. We denote $\overline{M} = \max_{1\le j\le s_1}\left\{|(\Delta\theta^n_{11j})^{(1)}|, |\Delta\theta^n_{21j}|^{1/2}\right\}$ and $\overline{p} = \max_{1\le j\le s_1}\{p_j\}$. Given this notation, let $(\Delta\theta^n_{11j})^{(1)}/\overline{M} \to a_j$, $\Delta\theta^n_{21j}/\overline{M}^2 \to b_j$, and $p^n_{1j}/\overline{p} \to c_j^2$ for all $1 \le j \le s_1$. Since $p^n_{1j} \ge \bar{c}_0$, we will have $c_j > 0$ for all $1 \le j \le s_1$. By dividing both the numerators and the denominators of $M_{0,l}(\theta^0_{11}, \theta^0_{21})$ by $\overline{M}^l$, we obtain the following system of polynomial equations:

$$\sum_{j=1}^{s_1}\sum_{\substack{\alpha_1+2\alpha_3=l\\ \alpha_1+\alpha_3\le\overline{r}}} \frac{c_j^2 a_j^{\alpha_1} b_j^{\alpha_3}}{2^{\alpha_3}\alpha_1!\alpha_3!} = 0$$

for all $1 \le l \le \overline{r}$. Since $s_1 \le k - k_0 + 1$ (as $s_i \ge 1$ for all $1 \le i \le k_0$), this system of polynomial equations will not admit any nontrivial solutions $(a_j, b_j, c_j)_{j=1}^{s_1}$ according to the definition of $\overline{r}$. This is a contradiction. As a consequence, not all the coefficients of $A_n/D_\kappa(G_n, G_0)$ and $B_n/D_\kappa(G_n, G_0)$ go to zero as $n \to \infty$.

Step 4 - Fatou's argument Equipped with the above result, we utilize Fatou's argument in Step 3 of the proof of inequality (16) to obtain a contradiction. We denote

$$m_n = \max_{0\le\alpha_2\le\overline{r},\ 0\le l\le 2(\overline{r}-\alpha_2),\ 1\le i\le k_0} \left|E_{\alpha_2,l}(\theta^0_{1i}, \theta^0_{2i})\right| / D_\kappa(G_n, G_0);$$

i.e., m_n is the maximum of the absolute values of the coefficients in the representation of $A_n/D_\kappa(G_n, G_0)$ and $B_n/D_\kappa(G_n, G_0)$. We now define $E_{\alpha_2,l}(\theta^0_{1i}, \theta^0_{2i})/m_n \to \tau_{\alpha_2,l}(i)$ as $n \to \infty$ for all $1 \le i \le k_0$, $0 \le \alpha_2 \le \overline{r}$, and $0 \le l \le 2(\overline{r} - \alpha_2)$. Here, at least one among $\tau_{\alpha_2,l}(i)$ is different from zero. Armed with Fatou's lemma as in the proof of inequality (16), we obtain the following equation:

$$\sum_{i,\alpha_2,l} \tau_{\alpha_2,l}(i) X^{\alpha_2} \frac{\partial^{l+\alpha_2} f}{\partial h_1^{l+\alpha_2}}(Y|h_1(X, \theta^0_{1i}), h_2(X, \theta^0_{2i}))\overline{f}(X) = 0, \tag{33}$$

almost surely (X, Y) where the ranges of (i, α_2, l) in the sum satisfy $1 \le i \le k_0$, $0 \le \alpha_2 \le \overline{r}$, and $0 \le l \le 2(\overline{r} - \alpha_2)$. According to the claim that the elements of $\mathcal{F}$ are linearly independent with respect to X and Y, equation (33) indicates that $\tau_{\alpha_2,l}(i) = 0$ for all i, α_2, l, which is a contradiction. As a consequence, we prove inequality (28).

Proof for claim that the elements of $\mathcal{F}$ are linearly independent To facilitate the presentation, we reuse the notation from Step 4. In particular, assume that we can find $\tau_{\alpha_2,l}(i) \in \mathbb{R}$ ($1 \le i \le k_0$, $0 \le \alpha_2 \le \overline{r}$, and $0 \le l \le 2(\overline{r} - \alpha_2)$) such that equation (33) holds almost surely X and Y. This equation is equivalent to

$$\sum_{i=1}^{k_0}\sum_{u=0}^{2\overline{r}}\left(\sum_{\alpha_2+l=u} \tau_{\alpha_2,l}(i) X^{\alpha_2}\right)\frac{\partial^u f}{\partial h_1^u}(Y|h_1(X|\theta^0_{1i}), h_2(X|\theta^0_{2i})) = 0 \tag{34}$$

for almost surely X and Y. Since $(\theta^0_{11}, \theta^0_{21}), \ldots, (\theta^0_{1k_0}, \theta^0_{2k_0})$ are k_0 distinct pairs, we also obtain that $(h_1(X|\theta^0_{11}), h_2(X|\theta^0_{21})), \ldots, (h_1(X|\theta^0_{1k_0}), h_2(X|\theta^0_{2k_0}))$ are k_0 distinct pairs for almost surely X. With that result, for X almost surely, we have that $\frac{\partial^u f}{\partial h_1^u}(Y|h_1(X|\theta^0_{1i}), h_2(X|\theta^0_{2i}))$ are

linearly independent with respect to Y for $0 \le u \le 2\overline{r}$. Therefore, equation (34) implies that $\sum_{j+l=u} \tau_{\alpha_2,l}(i) X^j = 0$ for all $1 \le i \le k_0$ and $0 \le u \le 2\overline{r}$. As it is a polynomial of $X \in \mathcal{X}$, which is a bounded subset of $\mathbb{R}$, equation (34) only holds when all the coefficients are zero; i.e., $\tau_{\alpha_2,l}(i) = 0$ for all $\alpha_2 + l = u$, $1 \le i \le k_0$ and $0 \le u \le 2\overline{r}$. Hence, we establish the claim.

5.2.2 Proof for equality (29)

In a manner similar to the proof strategy in Theorem 1, to obtain the conclusion for (29), it is sufficient to construct some sequence $G_n \in \mathcal{O}_k(\Omega)$ such that

$$h(p_{G_n}, p_{G_0}) / \widetilde{W}_{\kappa'}^{\|\kappa'\|_\infty}(G_n, G_0) \to 0,$$

for any $\kappa' \prec \kappa = (\overline{r}, 2, \overline{r}/2)$. The construction for G_n will be carried out under two particular settings of κ'.

Case 1: $\kappa' = (\kappa'^{(1)}, \kappa'^{(2)}, \kappa'^{(3)})$ where $\kappa'^{(2)} < 2$. Under this setting, we construct $G_n = \sum_{i=1}^{k_0+1} \pi_i^n \delta_{(\theta_{1i}^n, \theta_{2i}^n)}$ such that $(\pi_i^n, \theta_{1i}^n, \theta_{2i}^n) \equiv (\pi_{i-1}^0, \theta_{1(i-1)}^0, \theta_{2(i-1)}^0)$ for $3 \le i \le k_0 + 1$. Additionally, $\pi_1^n = \pi_2^n = \pi_1^0/2$, $\big((\theta_{1i}^n)^{(1)}, \theta_{2i}^n\big) = \big((\theta_{11}^0)^{(1)}, \theta_{21}^0\big)$ for $1 \le i \le 2$, and $(\theta_{11}^n)^{(2)} = (\theta_{11}^0)^{(2)} - 1/n$, $(\theta_{12}^n)^{(2)} = (\theta_{11}^0)^{(2)} + 1/n$. From this construction for G_n, we can verify that $\widetilde{W}_{\kappa'}^{\|\kappa'\|_\infty}(G_n, G_0) \asymp n^{-\kappa'^{(2)}}$. Denote $\Delta\theta_{1i}^n = \theta_{1i}^n - \theta_{1i}^0$ for $1 \le i \le 2$. Now, by means of Taylor expansion up to the first order around $(\theta_{11}^0)^{(2)}$, we have

$$\begin{aligned} p_{G_n}(X,Y) - p_{G_0}(X,Y) &= \sum_{i=1}^{2} \pi_i^n \left(f(Y|h_1(X,\theta_{1i}^n), h_2(X,\theta_{2i}^n)) - f(Y|h_1(X,\theta_{11}^0), h_2(X,\theta_{21}^0)) \right) \overline{f}(X) \\ &= \sum_{i=1}^{2} \pi_i^n (\Delta\theta_{1i}^n)^{(2)} \frac{\partial f}{\partial \theta_1^{(2)}} \left(Y|h_1(X,\theta_{11}^0), h_2(X,\theta_{21}^0)\right) \overline{f}(X) + \overline{R}_1(X,Y), \end{aligned}$$

where $(\Delta\theta_{1i}^n)^{(2)} = (\theta_{1i}^n)^{(2)} - (\theta_{1i}^0)^{(2)}$ for $1 \le i \le 2$ and $\overline{R}_1(X,Y)$ is Taylor remainder such that

$$\begin{aligned} \overline{R}_1(X,Y) = \sum_{i=1}^{2} \pi_i^n \left\{ (\Delta\theta_{1i}^n)^{(2)} \right\}^2 \int_0^1 (1-t) \frac{\partial^2 f}{\partial (\theta_1^{(2)})^2} (Y|h_1(X, \theta_{11}^0 + t\Delta\theta_{1i}^n), \\ h_2(X, \theta_{21}^0 + t\Delta\theta_{2i}^n)) \overline{f}(X) dt. \end{aligned}$$

It is not hard to check that $\sum_{i=1}^{2} \pi_i^n \left\{ (\Delta\theta_{1i}^n)^{(2)} \right\}^2 = \mathcal{O}(n^{-2})$ and

$$\sup_{t\in[0,1]} \int \frac{\left(\frac{\partial^2 f}{\partial (\theta_1^{(2)})^2 \partial \theta_2^\beta} \left(Y|h_1(X, \theta_{11}^0 + t\Delta\theta_{1i}^n), h_2(X, \theta_{21}^0 + t\Delta\theta_{2i}^n)\right) \right)^2}{f(Y|h_1(X,\theta_{11}^0), h_2(X,\theta_{21}^0)} d(X,Y) < \infty.$$

Therefore, using the same argument as in the proof of equality (17), the following holds:

$$\frac{h^2(p_{G_n}, p_{G_0})}{\widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n, G_0)} \precsim \int \frac{\overline{R}_1^2(X,Y)}{p_{G_0}(X,Y) \widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n, G_0)} d(X,Y) \precsim \frac{\mathcal{O}(n^{-4})}{n^{-2\kappa'^{(2)}}} \to 0$$

as $n \to \infty$. Therefore, we achieve the conclusion of equality (29) under Case 1.

Case 2: $\kappa' = (\kappa'^{(1)}, 2, \kappa'^{(3)})$ where $(\kappa'^{(1)}, \kappa'^{(3)}) \prec (\overline{r}, \overline{r}/2)$. Under this setting, we construct $G_n = \sum_{i=1}^{k} \pi_i^n \delta_{(\theta_{1i}^n, \theta_{2i}^n)}$ such that $(\pi_{i+k-k_0}^n, \theta_{1(i+k-k_0)}^n, \theta_{2(i+k-k_0)}^n) = (\pi_i^0, \theta_{1i}^0, \theta_{2i}^0)$ for $2 \leq i \leq k_0$. For $1 \leq j \leq k - k_0 + 1$, we choose $(\theta_{1j}^n)^{(2)} = (\theta_{11}^0)^{(2)}$ and

$$(\theta_{1j}^n)^{(1)} = (\theta_{11}^0)^{(1)} + \frac{a_j^*}{n}, \ \theta_{2j}^n = \theta_{21}^0 + \frac{2b_j^*}{n^2}, \ \pi_j^n = \frac{\pi_1^0 (c_j^*)^2}{\sum_{i=1}^{k-k_0+1} (c_j^*)^2},$$

where $(c_i^*, a_i^*, b_i^*)_{i=1}^{k-k_0+1}$ are the nontrivial solution of the system of polynomial equations (6) when $r = \overline{r} - 1$. With this formulation of G_n, it is clear that

$$\begin{aligned} &p_{G_n}(X,Y) - p_{G_0}(X,Y) \\ &\quad = \sum_{i=1}^{k-k_0+1} \pi_i^n \left(f(Y|h_1(X,\theta_{1i}^n), h_2(X,\theta_{2i}^n)) - f(Y|h_1(X,\theta_{11}^0), h_2(X,\theta_{21}^0)) \right) \overline{f}(X). \end{aligned}$$

By means of a Taylor expansion up to the $(\overline{r}-1)$th order around $\left((\theta_{11}^0)^{(1)}, \theta_{21}^0\right)$, i.e., along the direction of the first component of θ_{11}^0 and θ_{21}^0, the following equation holds:

$$\begin{aligned} &[f(Y|h_1(X,\theta_{1i}^n), h_2(X,\theta_{2i}^n)) - f(Y|h_1(X,\theta_{11}^0), h_2(X,\theta_{21}^0))]\overline{f}(X) \\ &= \sum_{1 \leq |\alpha| \leq \overline{r}-1} \frac{1}{\alpha!} \left\{ (\Delta\theta_{1i}^n)^{(1)} \right\}^{\alpha_1} (\Delta\theta_{2i}^n)^{\alpha_2} \frac{\partial^{|\alpha|} f}{\partial (\theta_1^{(1)})^{\alpha_1} \partial \theta_2^{\alpha_2}} \left(Y|h_1(X,\theta_{11}^0), h_2(X,\theta_{21}^0)\right) \overline{f}(X) + \overline{R}_{2i}(X,Y) \\ &= \sum_{1 \leq |\alpha| \leq \overline{r}-1} \frac{1}{\alpha!} \left\{ (\Delta\theta_{1i}^n)^{(1)} \right\}^{\alpha_1} (\Delta\theta_{2i}^n)^{\alpha_2} \frac{\partial^{\alpha_1+2\alpha_2} f}{\partial h_1^{\alpha_1+2\alpha_2}} (Y|h_1(X,\theta_{1i}^0), h_2(X,\theta_{2i}^0)) \overline{f}(X) + \overline{R}_{2i}(X,Y), \end{aligned}$$

where $\alpha = (\alpha_1, \alpha_2)$ in the sum and $\overline{R}_{2i}(X,Y)$ is a remainder. Equipped with this equation, we can rewrite $p_{G_n}(X,Y) - p_{G_0}(X,Y)$ as

$$\begin{aligned} p_{G_n}(X,Y) - p_{G_0}(X,Y) &= \sum_{i=1}^{k-k_0+1} \pi_i^n \sum_{1 \leq |\alpha| \leq \overline{r}-1} \frac{1}{\alpha!} \left\{ (\Delta\theta_{1i}^n)^{(1)} \right\}^{\alpha_1} (\Delta\theta_{2i}^n)^{\alpha_2} \\ &\quad \times \frac{\partial^{\alpha_1+2\alpha_2} f}{\partial h_1^{\alpha_1+2\alpha_2}} (Y|h_1(X,\theta_{1i}^0), h_2(X,\theta_{2i}^0)) \overline{f}(X) + \overline{R}_2(X,Y) \\ &= \sum_{l=1}^{2(\overline{r}-1)} \left[\sum_{\substack{\alpha_1+2\alpha_2=l \\ \alpha_1+\alpha_2 \leq \overline{r}-1}} \frac{1}{\alpha!} \sum_{i=1}^{k-k_0+1} \pi_i^n \left\{ (\Delta\theta_{1i}^n)^{(1)} \right\}^{\alpha_1} (\Delta\theta_{2i}^n)^{\alpha_2} \right] \\ &\quad \times \frac{\partial^l f}{\partial h_1^l} (Y|h_1(X,\theta_{1i}^0), h_2(X,\theta_{2i}^0)) \overline{f}(X) + \overline{R}_2(X,Y), \end{aligned}$$

where $\overline{R}_2(X,Y) = \sum_{i=1}^{k-k_0+1} \pi_i^n \overline{R}_{2i}(X,Y)$ and the range of α in the second equality satisfies $\alpha_1 + 2\alpha_2 = l$ and $\alpha_1 + \alpha_2 \leq \overline{r} - 1$. From the formulations of π_i^n, θ_{1i}^n, and θ_{2i}^n as $1 \leq i \leq k - k_0 + 1$, we can check that

$$\sum_{\substack{\alpha_1+2\alpha_2=l \\ \alpha_1+\alpha_2 \leq \overline{r}-1}} \frac{1}{\alpha!} \sum_{i=1}^{k-k_0+1} \pi_i^n \left\{ (\Delta\theta_{1i}^n)^{(1)} \right\}^{\alpha_1} (\Delta\theta_{2i}^n)^{\alpha_2} = 0,$$

when $1 \leq l \leq \overline{r}-1$. Additionally, we also have

$$L_l = \sum_{\substack{\alpha_1+2\alpha_2=l \\ \alpha_1+\alpha_2 \leq \overline{r}-1}} \frac{1}{\alpha!} \sum_{i=1}^{k-k_0+1} \pi_i^n \Big\{ (\Delta\theta_{1i}^n)^{(1)} \Big\}^{\alpha_1} (\Delta\theta_{2i}^n)^{\alpha_2} = \mathcal{O}(n^{-\overline{r}}),$$

when $\overline{r} \leq l \leq 2(\overline{r}-1)$. Furthermore, the explicit form of $\overline{R}_2(X,Y)$ is as follows:

$$\begin{aligned}\overline{R}_2(X,Y) = &\sum_{i=1}^{k-k_0+1} \pi_i^n \sum_{|\alpha|=\overline{r}} \frac{\overline{r}}{\alpha!} \Big\{ (\Delta\theta_{1i}^n)^{(1)} \Big\}^{\alpha_1} (\Delta\theta_{2i}^n)^{\alpha_2} \\ &\times \int_0^1 (1-t)^{\overline{r}-1} \frac{\partial^{\overline{r}} f}{\partial(\theta_1^{(1)})^{\alpha_1} \partial\theta_2^{\alpha_2}} \left(Y|h_1(X,\theta_{11}^0 + t\Delta\theta_{1i}^n), h_2(X,\theta_{21}^0 + t\Delta\theta_{2i}^n)\right) \overline{f}(X) dt.\end{aligned}$$

It is not hard to check that $\sum_{i=1}^{k-k_0+1} \pi_i^n \Big\{ (\Delta\theta_{1i}^n)^{(1)} \Big\}^{\alpha_1} (\Delta\theta_{2i}^n)^{\alpha_2} = \mathcal{O}(n^{-\overline{r}})$ and

$$\sup_{t\in[0,1]} \int \frac{\left(\dfrac{\partial^{\overline{r}} f}{\partial(\theta_1^{(1)})^{\alpha_1} \partial\theta_2^{\alpha_2}} \left(Y|h_1(X,\theta_{11}^0 + t\Delta\theta_{1i}^n), h_2(X,\theta_{21}^0 + t\Delta\theta_{2i}^n)\right) \right)^2}{f(Y|h_1(X,\theta_{11}^0), h_2(X,\theta_{21}^0)} d(X,Y) < \infty,$$

for any $|\alpha| = r$. By the Cauchy-Schwartz inequality, the following inequality holds:

$$\begin{aligned}\frac{h^2(p_{G_n}, p_{G_0})}{\widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n, G_0)} \precsim &\sum_{l=\overline{r}}^{2(\overline{r}-1)} \frac{L_l^2}{\widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n, G_0)} \int \frac{\left(\frac{\partial^l f}{\partial h_1^l}(Y|h_1(X,\theta_{1i}^0), h_2(X,\theta_{2i}^0))\overline{f}(X) \right)^2}{p_{G_0}(X,Y)} d(X,Y) \\ &+ \frac{\overline{R}_2^2(X,Y)}{\widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n, G_0)}.\end{aligned}$$

From a property of the location-scale Gaussian distribution, we have

$$\int \frac{\left(\frac{\partial^l f}{\partial h_1^l}(Y|h_1(X,\theta_{1i}^0), h_2(X,\theta_{2i}^0))\overline{f}(X) \right)^2}{p_{G_0}(X,Y)} d(X,Y) < \infty,$$

for any $\overline{r} \leq l \leq 2(\overline{r}-1)$. Furthermore, by means of a similar argument as in the proof of equality (17), we can argue that

$$\frac{\overline{R}_2^2(X,Y)}{\widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n, G_0)} \precsim \frac{\mathcal{O}(n^{-2\overline{r}})}{n^{-2\min\{\kappa'^{(1)}, \kappa'^{(3)}\}}}.$$

Putting these results together, we have

$$\frac{h^2(p_{G_n}, p_{G_0})}{\widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n, G_0)} \precsim \frac{\mathcal{O}(n^{-2\overline{r}})}{n^{-2\min\{\kappa'^{(1)}, \kappa'^{(3)}\}}} \to 0, \tag{35}$$

as $n \to \infty$. Therefore, we obtain the conclusion of equality (29) under Case 2.

6 Discussion

We have provided a systematic theoretical understanding of the convergence rates of parameter estimation under over-specified Gaussian mixtures of experts based on an analysis of an underlying algebraic structure. In particular, we have introduced a new theoretical tool, which we refer to as algebraic independence, and we have established a connection between this algebraic structure and a certain family of PDEs. This connection allows us to determine convergence rates of the MLE under various choices of expert functions h_1 and h_2.

There are several directions for future research. First, the current convergence rates of the MLE are established under the assumptions that the parameter spaces are bounded; it would be important to remove this assumption for wider practical applicability. Second, the results of the paper demonstrate that the convergence rates of MLE are only very slow when the expert functions are algebraically dependent. When we indeed fit the models with algebraically independent expert functions while the true expert functions are algebraically dependent, i.e., we misspecify the expert functions, the convergence rates of MLE become $n^{-1/4}$. However, the MLE will not converge to the true mixing measure. This raises an interesting challenge of how to characterize the difference between the limiting mixing measure and the true mixing measure in terms of the generalized transportation distance. Finally, since the log-likelihood function of over-specified Gaussian mixtures of experts is nonconcave, the MLE does not have a closed form in practice. Therefore, heuristic optimization algorithms, such as Expectation-Maximization (EM) algorithm, are generally used to approximate MLE. It is of practical importance to investigate the computational errors arising from the updates of these algorithms on the convergence rates of MLE.

7 Acknowledgements

This work was supported in part by the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764.

References

[1] J. Chen. Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23(1):221–233, 1995. (Cited on pages 1, 8, and 17.)

[2] G. Compiani and Y. Kitamura. Using mixtures in econometric models: a brief review and some new results. *Econometrics Journal*, 19:95–127, 2016. (Cited on page 1.)

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

[4] S. Ghosal and A. van der Vaart. Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *Annals of Statistics*, 29:1233–1263, 2001. (Cited on page 53.)

[5] P. Heinrich and J. Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *Annals of Statistics*, 46, 2018. (Cited on pages 1 and 17.)

[6] N. Ho and X. Nguyen. Singularity structures and impacts on parameter estimation in finite mixtures of distributions. *arXiv preprint arXiv:1607.01251*, 2016. (Cited on pages 1, 2, 9, 17, 29, and 52.)

[7] M. Huang, R. Li, and S. Wang. Nonparametric mixture of regression models. *Journal of the American Statistical Association*, 108:929–941, 2013. (Cited on page 1.)

[8] M. Huang and W. Yao. Mixture of regression models with varying mixing proportions: A semiparametric approach. *Journal of the American Statistical Association*, 107:711–724, 2012. (Cited on page 1.)

[9] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3, 1991. (Cited on pages 1 and 4.)

[10] W. Jiang and M. A. Tanner. Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *Annals of Statistics*, 27:987–1011, 1999. (Cited on page 2.)

[11] W. Jiang and M. A. Tanner. Hierarchical mixtures-of-experts for generalized linear models: some results on denseness and consistency. In *AISTATS*, 1999. (Cited on page 2.)

[12] W. Jiang and M. A. Tanner. On the approximation rate of hierarchical mixtures-of-experts for generalized linear models. *Neural computation*, 11:1183–1198, 1999. (Cited on page 2.)

[13] W. Jiang and M. A. Tanner. On the identifiability of mixtures-of-experts. *Neural Networks*, 9:1253–1258, 1999. (Cited on page 2.)

[14] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994. (Cited on pages 1 and 4.)

[15] M. I. Jordan and L. Xu. Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks*, 8, 1995. (Cited on pages 1 and 4.)

[16] A. Khalili and J. Chen. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102:1025–1038, 2007. (Cited on page 2.)

[17] B. Lindsay. *Mixture models: Theory, geometry and applications*. In NSF-CBMS Regional Conference Series in Probability and Statistics. IMS, Hayward, CA., 1995. (Cited on page 1.)

[18] A. Makkuva, P. Viswanath, S. Kannan, and S. Oh. Breaking the gridlock in mixture-of-experts: Consistent and efficient algorithms. In *ICML*, 2019. (Cited on page 1.)

[19] A. Makkuva, P. Viswanath, S. Kannan, and S. Oh. Learning in gated neural networks. *arXiv preprint arXiv:1906.02777*, 2019. (Cited on page 1.)

[20] X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 4(1):370–400, 2013. (Cited on pages 1 and 8.)

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

[22] C. E. Rasmussen and Z. Ghahramani. Infinite mixtures of Gaussian process experts. In *NIPS 14*, 2002. (Cited on page 1.)

[23] J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B*, 73(5):689–710, 2011. (Cited on page 1.)

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

[25] H. Teicher. On the mixture of distributions. *Annals of Statistics*, 31:55–73, 1960. (Cited on page 1.)

[26] H. Teicher. Identifiability of mixtures. *Annals of Statistics*, 32:244–248, 1961. (Cited on page 1.)

[27] H. Teicher. Identifiability of finite mixtures. *Annals of Statistics*, 34:1265–1269, 1963. (Cited on page 1.)

[28] S. van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000. (Cited on page 52.)

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

A Appendix

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

A.1 Proof of Theorem 3

Similar to previous proofs in Section 5, it is sufficient to demonstrate the following results:

$$\lim_{\epsilon\to 0}\inf_{G\in\mathcal{O}_{k,\bar{c}_0}(\Omega):\widetilde{W}_\kappa(G,G_0)\leq\epsilon} V(p_G,p_{G_0})/\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G,G_0)>0, \tag{36}$$

$$\inf_{G\in\mathcal{O}_k(\Omega)} h(p_G,p_{G_0})/\widetilde{W}_{\kappa'}^{\|\kappa'\|_\infty}(G,G_0)=0, \tag{37}$$

for any $\kappa'\prec\kappa$ where $\kappa=(2,\overline{r},\lceil\overline{r}/2\rceil)$. Without loss of generality, we assume that $\overline{r}$ is an even number throughout this proof, i.e., $\kappa=(2,\overline{r},\overline{r}/2)$.

A.1.1 Proof for inequality (36)

Assume the inequality (36) does not hold. It indicates that there exists a sequence $G_n\in\mathcal{O}_{k,\bar{c}_0}(\Omega)$ such that $V(p_G,p_{G_0})/\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G_n,G_0)\to 0$ and $\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G_n,G_0)\to 0$. To simplify the presentation, we reuse the notation of G_n as in (30) in the proof of Theorem 2 in Section 5.2.1. Since $\kappa=(2,\overline{r},\overline{r}/2)$, we have

$$\begin{aligned}\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G_n,G_0)\precsim\sum_{i=1}^{k_0}\sum_{j=1}^{s_i}p_{ij}^n\left(\left|(\Delta\theta_{1ij}^n)^{(1)}\right|^2+\left|(\Delta\theta_{1ij}^n)^{(2)}\right|^{\overline{r}}+|\Delta\theta_{2ij}^n|^{\overline{r}/2}\right)\\ +\sum_{i=1}^{k_0}|\sum_{j=1}^{s_i}p_{ij}^n-\pi_i^0|:=D_\kappa(G_n,G_0).\end{aligned}$$

Similar to the proof of Theorem 2, by means of Taylor expansion up to the $\overline{r}$ order, we can represent

$$p_{G_n}(X,Y)-p_{G_0}(X,Y)=A_n+B_n+R(X,Y),$$

where A_n, B_n, and $R(X,Y)$ are identifical to those in (31) such that $R(X,Y)/D_\kappa(G_n,G_0)\to 0$ as $n\to\infty$. Given the formulation of expert functions h_1,h_2, we have the following key equation:

$$\frac{\partial^{|\alpha|}f}{\partial(\theta_1^{(1)})^{\alpha_1}\partial(\theta_1^{(2)})^{\alpha_2}\partial\theta_2^{\alpha_3}}(Y|h_1(X|\theta_1),h_2(X|\theta_2))=\frac{X^{\alpha_2+2\alpha_3}}{2^{\alpha_3}}\frac{\partial^{\alpha_1+\alpha_2+2\alpha_3}f}{\partial h_1^{\alpha_1+\alpha_2+2\alpha_3}}(Y|h_1(X|\theta_1),h_2(X|\theta_2)),$$

for any $\alpha_1,\alpha_2,\alpha_3\in\mathbb{N}$. Equipped with the above equation, A_n can be rewritten as

$$\begin{aligned}A_n=\sum_{i=1}^{k_0}\sum_{j=1}^{s_i}\pi_{ij}^n\sum_{\alpha_1=0}^{\overline{r}}\sum_{l=0}^{2(\overline{r}-\alpha_1)}\sum_{\alpha_1,\alpha_3}\frac{1}{2^{\alpha_3}\alpha!}\left\{(\Delta\theta_{1ij}^n)^{(1)}\right\}^{\alpha_1}\left\{(\Delta\theta_{1ij}^n)^{(2)}\right\}^{\alpha_2}(\Delta\theta_{2ij}^n)^{\alpha_3}\\ \times X^l\frac{\partial^{\alpha_1+l}f}{\partial h_1^{\alpha_1+l}}(Y|h_1(X|\theta_{1i}^0),h_2(X|\theta_{2i}^0))\overline{f}(X),\end{aligned} \tag{38}$$

where $\alpha_2, \alpha_3 \in \mathbb{N}$ in the above sum satisfies $\alpha_2 + 2\alpha_3 = l$ and $1 - \alpha_1 \leq \alpha_2 + \alpha_3 \leq \overline{r} - \alpha_1$. If we define

$$\mathcal{F} = \Big\{ X^l \frac{\partial^{\alpha_1 + l} f}{\partial h_1^{\alpha_1 + l}}(Y|h_1(X|\theta_{1i}^0), h_2(X|\theta_{2i}^0))\overline{f}(X) : \ 0 \leq \alpha_1 \leq \overline{r}, \ 0 \leq l \leq 2(\overline{r} - \alpha_1), \ 1 \leq i \leq k_0 \Big\},$$

then the elements of $\mathcal{F}$ are linearly independent with respect to X and Y. The proof argument of this claim is similar to that in (34) in Section 5.2.1. Therefore, we can treat $A_n/D_\kappa(G_n, G_0)$, $B_n/D_\kappa(G_n, G_0)$ as a linear combination of elements of $\mathcal{F}$.

Similar to the proof of Theorem 2 in Section 5.2.1, we denote $F_{\alpha_1,l}(\theta_{1i}^0, \theta_{2i}^0)$ as the coefficient of $X^l \frac{\partial^{\alpha_1 + l} f}{\partial h_1^{\alpha_1 + l}}(Y|h_1(X|\theta_{1i}^0), h_2(X|\theta_{2i}^0))\overline{f}(X)$ in A_n and B_n for any $0 \leq \alpha_1 \leq \overline{r}$, $0 \leq l \leq 2(\overline{r} - \alpha_1)$, and $1 \leq i \leq k_0$. Then, the coefficients of $X^l \frac{\partial^{\alpha_1 + l} f}{\partial h_1^{\alpha_1 + l}}(Y|h_1(X|\theta_{1i}^0), h_2(X|\theta_{2i}^0))\overline{f}(X)$ in $A_n/D_\kappa(G_n, G_0)$ and $B_n/D_\kappa(G_n, G_0)$ will be $F_{\alpha_1,l}(\theta_{1i}^0, \theta_{2i}^0)/D_\kappa(G_n, G_0)$.

Assume that all of these coefficients go to 0 as $n \to \infty$. By taking the summation of $|F_{0,0}(\theta_{1i}^0, \theta_{2i}^0)/D_\kappa(G_n, G_0)|$ for all $1 \leq i \leq k_0$, we obtain that

$$\Big(\sum_{i=1}^{k_0} |\sum_{j=1}^{s_i} p_{ij}^n - \pi_i^0| \Big) / D_\kappa(G_n, G_0) \to 0.$$

Additionally, according to equation (38), we can verify that

$$F_{2,0}(\theta_{1i}^0, \theta_{2i}^0)/D_\kappa(G_n, G_0) = \Big(\sum_{j=1}^{s_i} p_{ij}^n \Big| (\Delta\theta_{1ij}^n)^{(1)} \Big|^2 \Big) / D_\kappa(G_n, G_0) \to 0$$

for all $1 \leq i \leq k_0$. From the formulation of $D_\kappa(G_n, G_0)$, the above limits lead to

$$\Big\{ \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n \Big(\Big| (\Delta\theta_{1ij}^n)^{(2)} \Big|^{\overline{r}} + |\Delta\theta_{2ij}^n|^{\overline{r}/2} \Big) \Big\} / D_\kappa(G_n, G_0) \to 1 \text{ as } n \to \infty.$$

Thus, we can find an index $i^* \in \{1, \ldots, k_0\}$ such that

$$J = \Big\{ \sum_{j=1}^{s_i} p_{i^*j}^n \Big(\Big| (\Delta\theta_{1i^*j}^n)^{(2)} \Big|^{\overline{r}} + |\Delta\theta_{2i^*j}^n|^{\overline{r}/2} \Big) \Big\} / D_\kappa(G_n, G_0) \not\to 0$$

as $n \to \infty$. Without loss of generality, we assume that $i^* = 1$. Now, since we have $F_{\alpha_1,l}(\theta_{1i}^0, \theta_{2i}^0)/D_\kappa(G_n, G_0) \to 0$ for all values of α_1, l, i, we obtain that

$$M_{\alpha_1,l}(\theta_{1i}^0, \theta_{2i}^0) = \frac{F_{\alpha_1,l}(\theta_{1i}^0, \theta_{2i}^0)}{\sum_{j=1}^{s_i} p_{1j}^n \Big(\Big| (\Delta\theta_{11j}^n)^{(2)} \Big|^{\overline{r}} + |\Delta\theta_{21j}^n|^{\overline{r}/2} \Big)} = \frac{1}{J} \frac{F_{\alpha_1,l}(\theta_{1i}^0, \theta_{2i}^0)}{D_\kappa(G_n, G_0)} \to 0,$$

for any $0 \leq \alpha_1 \leq \overline{r}$, $0 \leq l \leq 2(\overline{r} - \alpha_1)$, and $1 \leq i \leq k_0$. From the representation of A_n in (38), we can verify that

$$M_{0,l}(\theta_{11}^0, \theta_{21}^0) = \frac{\sum_{j=1}^{s_1} p_{1j}^n \sum_{\substack{\alpha_2 + 2\alpha_3 = l \\ \alpha_2 + \alpha_3 \leq \overline{r}}} \frac{\Big\{ (\Delta\theta_{11j}^n)^{(2)} \Big\}^{\alpha_2} (\Delta\theta_{21j}^n)^{\alpha_3}}{2^{\alpha_3} \alpha_2! \alpha_3!}}{\sum_{j=1}^{s_i} p_{1j}^n \Big(\Big| (\Delta\theta_{11j}^n)^{(2)} \Big|^{\overline{r}} + |\Delta\theta_{21j}^n|^{\overline{r}/2} \Big)} \to 0.$$

for $0 \leq l \leq 2\overline{r}$. Using the same argument as that in Step 3 of the proof of Theorem 2 in Section 5.2.1, the above system of polynomial limits does not hold. As a consequence, not all the coefficients in the linear combinations of $A_n/D_\kappa(G_n, G_0)$ and $B_n/D_\kappa(G_n, G_0)$ go to 0 as $n \to \infty$. From here, using the Fatou's argument in Step 4 of the proof of Theorem 2 and the fact that the elements of $\mathcal{F}$ are linearly independent with respect to X and Y, we achieve the conclusion of (36).

A.1.2 Proof for equality (37)

To alleviate the presentation, we will only provide a proof sketch of equality (37). We also divide the proof into two settings of $\kappa' \prec \kappa = (2, \overline{r}, \overline{r}/2)$.

Case 1: $\kappa' = \left(\kappa'^{(1)}, \kappa'^{(2)}, \kappa'^{(3)}\right)$ when $\kappa'^{(1)} < 2$. Under this setting, we construct $G_n = \sum_{i=1}^{k_0+1} \pi_i^n \delta_{(\theta_{1i}^n, \theta_{2i}^n)}$ such that $(\pi_i^n, \theta_{1i}^n, \theta_{2i}^n) \equiv (\pi_{i-1}^0, \theta_{1(i-1)}^0, \theta_{2(i-1)}^0)$ for $3 \leq i \leq k_0 + 1$. Additionally, $\pi_1^n = \pi_2^n = \pi_1^0/2$, $\left((\theta_{1i}^n)^{(2)}, \theta_{2i}^n\right) = \left((\theta_{11}^0)^{(2)}, \theta_{21}^0\right)$ for $1 \leq i \leq 2$, and $(\theta_{11}^n)^{(1)} = (\theta_{11}^0)^{(1)} - 1/n$, $(\theta_{12}^n)^{(1)} = (\theta_{11}^0)^{(1)} + 1/n$. From this construction of G_n, we can verify that $\widetilde{W}_{\kappa'}^{\|\kappa'\|_\infty}(G_n, G_0) \asymp n^{-\kappa'^{(1)}}$. Given that formulation of G_n, when we perform Taylor expansion up to the first order around $(\theta_{11}^0)^{(1)}$, the following equation holds

$$p_{G_n}(X, Y) - p_{G_0}(X, Y) = \overline{R}_1(X, Y),$$

where $\overline{R}_1(X, Y)$ is Taylor remainder such that

$$\begin{aligned}
&\overline{R}_1(X, Y) \\
&= \sum_{i=1}^{2} \pi_i^n \left\{ (\Delta\theta_{1i}^n)^{(1)} \right\}^2 \int_0^1 (1-t) \frac{\partial^2 f}{\partial (\theta_1^{(1)})^2} \left(Y | h_1(X, \theta_{11}^0 + t\Delta\theta_{1i}^n), h_2(X, \theta_{21}^0 + t\Delta\theta_{2i}^n)\right) \overline{f}(X) dt.
\end{aligned}$$

Using the same argument as that in Case 1 in the proof of equality (29) in Section 5.2.2, the following holds

$$\frac{h^2(p_{G_n}, p_{G_0})}{\widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n, G_0)} \precsim \int \frac{\overline{R}_1^2(X, Y)}{p_{G_0}(X, Y)\widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n, G_0)} d(X, Y) \precsim \frac{\mathcal{O}(n^{-4})}{n^{-2\kappa'^{(1)}}} \to 0$$

as $n \to \infty$. Therefore, we achieve the conclusion of equality (37) under Case 1.

Case 2: $\kappa' = \left(2, \kappa'^{(2)}, \kappa'^{(3)}\right)$ when $(\kappa_1'^{(2)}, \kappa'^{(3)}) \prec (\overline{r}, \overline{r}/2)$. Under this setting of κ', we construct $G_n = \sum_{i=1}^{k} \pi_i^n \delta_{(\theta_{1i}^n, \theta_{2i}^n)}$ such that $(\pi_{i+k-k_0}^n, \theta_{1(i+k-k_0)}^n, \theta_{2(i+k-k_0)}^n) = (\pi_i^0, \theta_{1i}^0, \theta_{2i}^0)$ for $2 \leq i \leq k_0$. For $1 \leq j \leq k - k_0 + 1$, we choose $(\theta_{1j}^n)^{(1)} = (\theta_{11}^0)^{(1)}$ and

$$(\theta_{1j}^n)^{(2)} = (\theta_{11}^0)^{(2)} + \frac{a_j^*}{n}, \ \theta_{2j}^n = \theta_{21}^0 + \frac{2b_j^*}{n^2}, \ \pi_j^n = \frac{\pi_1^0 (c_j^*)^2}{\sum_{i=1}^{k-k_0+1} (c_j^*)^2},$$

where $(c_i^*, a_i^*, b_i^*)_{i=1}^{k-k_0+1}$ are the non-trivial solution of system of polynomial equations (6) when $r = \overline{r} - 1$. From here, by performing Taylor expansion around $((\theta_{11}^0)^{(2)}, \theta_{21}^0)$, i.e., along

the direction of the second component of θ_{11}^0 and θ_{21}^0 and arguing similarly as Case 2 in the proof of equality (29) in Section 5.2.2, we obtain that

$$p_{G_n}(X,Y) - p_{G_0}(X,Y) = \sum_{l=\overline{r}}^{2\overline{r}-2} \mathcal{O}(n^{-\overline{r}}) \frac{\partial^l f}{\partial h_1^l}(Y|h_1(X,\theta_{1i}^0), h_2(X,\theta_{2i}^0))\overline{f}(X) + \overline{R}_2(X,Y),$$

where $\overline{R}_2(X,Y)$ is Taylor remainder such that the following limit holds

$$\int \frac{\overline{R}_2^2(X,Y)}{p_{G_0}(X,Y)\widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n,G_0)} d(X,Y) \precsim \frac{\mathcal{O}(n^{-2\overline{r}})}{n^{-2\min\{\kappa'^{(2)},\kappa'^{(2)}\}}} \to 0.$$

Therefore, we achieve that

$$h^2(p_{G_n}, p_{G_0}) \Big/ \widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n,G_0) \to 0$$

as $n \to \infty$. As a consequence, we reach the conclusion of equality (37) under Case 2.

A.2 Proof of Theorem 4

Similar to the previous proofs, it is sufficient to demonstrate the following results:

$$\lim_{\epsilon\to 0} \inf_{G\in\mathcal{O}_{k,\bar{c}_0}(\Omega):\widetilde{W}_\kappa(G,G_0)\le\epsilon} V(p_G, p_{G_0})/\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G,G_0) > 0, \tag{39}$$

$$\inf_{G\in\mathcal{O}_k(\Omega)} h(p_G, p_{G_0})/\widetilde{W}_{\kappa'}^{\|\kappa'\|_\infty}(G,G_0) = 0, \tag{40}$$

for any $\kappa' \prec \kappa$ where $\kappa = (\overline{r}, \overline{r}, \lceil \overline{r}/2 \rceil, \lceil \overline{r}/2 \rceil)$. Without loss of generality, we assume that $\overline{r}$ is even, i.e., $\kappa = (\overline{r}, \overline{r}, \overline{r}/2, \overline{r}/2)$.

A.2.1 Proof for inequality (39)

Assume that the conclusion of (39) does not hold. By using the same notations of G_n as in the proof of Theorem 2, we can find a sequence G_n that has representation (30) such that $V(p_{G_n}, p_{G_0})/\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G_n,G_0) \to 0$ and $\widetilde{W}_\kappa(G_n,G_0) \to 0$. Here, since θ_{2ij}^n and $\theta_{2i}^{(0)}$ have 2 dimensions, we denote $\Delta\theta_{2ij}^n = ((\Delta\theta_{2ij}^n)^{(1)}, (\Delta\theta_{2ij}^n)^{(2)})$ for all $1 \le i \le k_0$ and $1 \le j \le s_i$ throughout this proof. According to Lemma 5, we have

$$\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G_n,G_0) \precsim \sum_{i=1}^{k_0}\sum_{j=1}^{s_i} p_{ij}^n \left(\left|(\Delta\theta_{1ij}^n)^{(1)}\right|^{\overline{r}} + \left|(\Delta\theta_{1ij}^n)^{(2)}\right|^{\overline{r}} + \left|(\Delta\theta_{2ij}^n)^{(1)}\right|^{\overline{r}/2} + \left|(\Delta\theta_{2ij}^n)^{(2)}\right|^{\overline{r}/2} \right)$$
$$+ \sum_{i=1}^{k_0} \Big| \sum_{j=1}^{s_i} p_{ij}^n - \pi_i^0 \Big| := D_\kappa(G_n,G_0).$$

Invoking Taylor expansion up to the order $\overline{r}$, we obtain that

$$\begin{aligned}
p_{G_n}(X,Y) - p_{G_0}(X,Y) &= \sum_{i=1}^{k_0}\sum_{j=1}^{s_i} p_{ij}^n \sum_{1\leq|\alpha|\leq\overline{r}} \frac{1}{\alpha!}\Big\{(\Delta\theta_{1ij}^n)^{(1)}\Big\}^{\alpha_1}\Big\{(\Delta\theta_{1ij}^n)^{(2)}\Big\}^{\alpha_2}\Big\{(\Delta\theta_{2ij}^n)^{(1)}\Big\}^{\alpha_3} \\
&\times\Big\{(\Delta\theta_{2ij}^n)^{(2)}\Big\}^{\alpha_4} \frac{\partial^{|\alpha|} f}{\partial(\theta_1^{(1)})^{\alpha_1}\partial(\theta_1^{(2)})^{\alpha_2}\partial(\theta_2^{(1)})^{\alpha_3}\partial(\theta_2^{(2)})^{\alpha_4}}(Y|h_1(X|\theta_{1i}^0), h_2(X|\theta_{2i}^0))\overline{f}(X) \\
&+ \sum_{i=1}^{k_0}\Big(\sum_{j=1}^{s_i} p_{ij}^n - p_i^0\Big) f(Y|h_1(X|\theta_{1i}^0, h_2(X|\theta_{2i}^0))\overline{f}(X) + R(X,Y) \\
&:= A_n + B_n + R(X,Y),
\end{aligned}$$

where $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ and $R(X,Y)$ is a Taylor remainder such that $R(X,Y)/D_\kappa(G_n, G_0) \to 0$ as $n \to \infty$ for all (X,Y). The formulation of expert functions h_1, h_2 and the PDE structure of Gaussian kernel lead to

$$\begin{aligned}
&\frac{\partial^{|\alpha|} f}{\partial(\theta_1^{(1)})^{\alpha_1}\partial(\theta_1^{(2)})^{\alpha_2}\partial(\theta_2^{(1)})^{\alpha_3}\partial(\theta_2^{(2)})^{\alpha_4}}(Y|h_1(X,\theta_1), h_2(X,\theta_2)) \\
&\qquad = \frac{X^{\alpha_2+2\alpha_4}}{2^{\alpha_3+\alpha_4}}\frac{\partial^{\alpha_1+\alpha_2+2\alpha_3+2\alpha_4} f}{\partial h_1^{\alpha_1+\alpha_2+2\alpha_3+2\alpha_4}}(Y|h_1(X,\theta_1), h_2(X,\theta_2)),
\end{aligned}$$

for any $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in \mathbb{N}$, $\theta_1 \in \Omega_1$, and $\theta_2 \in \Omega_2$. With the above equation, we can rewrite A_n as follows

$$\begin{aligned}
A_n &= \sum_{i=1}^{k_0}\sum_{j=1}^{s_i} p_{ij}^n \sum_{1\leq|\alpha|\leq\overline{r}} \frac{1}{\alpha!}\Big\{(\Delta\theta_{1ij}^n)^{(1)}\Big\}^{\alpha_1}\Big\{(\Delta\theta_{1ij}^n)^{(2)}\Big\}^{\alpha_2}\Big\{(\Delta\theta_{2ij}^n)^{(1)}\Big\}^{\alpha_3}\Big\{(\Delta\theta_{2ij}^n)^{(2)}\Big\}^{\alpha_4} \\
&\times\frac{X^{\alpha_2+2\alpha_4}}{2^{\alpha_3}}\frac{\partial^{\alpha_1+\alpha_2+2\alpha_3+2\alpha_4} f}{\partial h_1^{\alpha_1+\alpha_2+2\alpha_3+2\alpha_4}}(Y|h_1(X|\theta_{1i}^0), h_2(X|\theta_{2i}^0))\overline{f}(X) \\
&= \sum_{i=1}^{k_0}\sum_{j=1}^{s_i} p_{ij}^n \sum_{1\leq l_1+l_2\leq 2\overline{r}}\Bigg(\sum_{\alpha_1,\alpha_2,\alpha_3,\alpha_4} \frac{1}{2^{\alpha_3+\alpha_4}\alpha!}\Big\{(\Delta\theta_{1ij}^n)^{(1)}\Big\}^{\alpha_1}\Big\{(\Delta\theta_{1ij}^n)^{(2)}\Big\}^{\alpha_2}\Big\{(\Delta\theta_{2ij}^n)^{(1)}\Big\}^{\alpha_3} \\
&\times\Big\{(\Delta\theta_{2ij}^n)^{(2)}\Big\}^{\alpha_4}\Bigg) X^{l_2}\frac{\partial^{l_1+l_2} f}{\partial h_1^{l_1+l_2}}(Y|h_1(X|\theta_{1i}^0), h_2(X|\theta_{2i}^0))\overline{f}(X), \qquad (41)
\end{aligned}$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in \mathbb{N}$ in the sum of second equation satisfies $\alpha_1 + 2\alpha_3 = l_1$, $\alpha_2 + 2\alpha_4 = l_2$, and $1 \leq \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 \leq \overline{r}$.

As demonstrated in the earlier proofs, we can treat $A_n/D_\kappa(G_n, G_0)$ and $B_n/D_\kappa(G_n, G_0)$ as a linear combination of $X^{l_2}\dfrac{\partial^{l_1+l_2} f}{\partial h_1^{l_1+l_2}}(Y|h_1(X|\theta_{1i}^0), h_2(X|\theta_{2i}^0))\overline{f}(X)$ for $0 \leq l_1 + l_2 \leq 2\overline{r}$ and $1 \leq i \leq k_0$, which are linearly independent with respect to X and Y. For the simplicity of presentation, we denote $E_{l_1,l_2}(\theta_{1i}^0, \theta_{2i}^0)$ as the coefficient of $X^{l_2}\dfrac{\partial^{l_1+l_2} f}{\partial h_1^{l_1+l_2}}(Y|h_1(X|\theta_{1i}^0), h_2(X|\theta_{2i}^0))\overline{f}(X)$

in A_n, B_n. From the equation (41), we can check that

$$E_{0,l}(\theta^0_{1i},\theta^0_{2i}) = \Bigg(\sum_{j=1}^{s_i} p^n_{ij} \sum_{\substack{\alpha_2+2\alpha_4=l\\ \alpha_2+\alpha_4\leq \overline{r}}} \Big\{(\Delta\theta^n_{1ij})^{(2)}\Big\}^{\alpha_2}\Big\{(\Delta\theta^n_{2ij})^{(2)}\Big\}^{\alpha_4}\Bigg)/(2^{\alpha_4}\alpha_2!\alpha_4!),$$

$$E_{l,0}(\theta^0_{1i},\theta^0_{2i}) = \Bigg(\sum_{j=1}^{s_i} p^n_{ij} \sum_{\substack{\alpha_1+2\alpha_3=l\\ \alpha_1+\alpha_3\leq \overline{r}}} \Big\{(\Delta\theta^n_{1ij})^{(1)}\Big\}^{\alpha_1}\Big\{(\Delta\theta^n_{2ij})^{(1)}\Big\}^{\alpha_3}\Bigg)/(2^{\alpha_3}\alpha_1!\alpha_3!),$$

for any $1 \leq l \leq 2\overline{r}$.

Assume that all of the coefficients of $A_n/D_\kappa(G_n,G_0)$ and $B_n/D_\kappa(G_n,G_0)$ go to 0 as $n \to \infty$. The summation of $|E_{0,0}(\theta^0_{1i},\theta^0_{2i})/D_\kappa(G_n,G_0)|$ for all $1 \leq i \leq k_0$ leads to

$$\Bigg(\sum_{i=1}^{k_0}|\sum_{j=1}^{s_i} p^n_{ij} - \pi^0_i|\Bigg)/D_\kappa(G_n,G_0) \to 0.$$

From the formulation of $D_\kappa(G_n,G_0)$, the above limit implies that

$$\Bigg\{\sum_{i=1}^{k_0}\sum_{j=1}^{s_i} p^n_{ij}\Bigg(\Big|(\Delta\theta^n_{1ij})^{(1)}\Big|^{\overline{r}} + \Big|(\Delta\theta^n_{1ij})^{(2)}\Big|^{\overline{r}} + \Big|(\Delta\theta^n_{2ij})^{(1)}\Big|^{\overline{r}/2} + \Big|(\Delta\theta^n_{2ij})^{(2)}\Big|^{\overline{r}/2}\Bigg)\Bigg\}/D_\kappa(G_n,G_0) \to 1.$$

Therefore, we can find an index $i^* \in \{1,\ldots,k_0\}$ such that

$$\Bigg\{\sum_{j=1}^{s_{i^*}} p^n_{i^*j}\Bigg(\Big|(\Delta\theta^n_{1i^*j})^{(1)}\Big|^{\overline{r}} + \Big|(\Delta\theta^n_{1i^*j})^{(2)}\Big|^{\overline{r}} + \Big|(\Delta\theta^n_{2i^*j})^{(1)}\Big|^{\overline{r}/2} + \Big|(\Delta\theta^n_{2i^*j})^{(2)}\Big|^{\overline{r}/2}\Bigg)\Bigg\}/D_\kappa(G_n,G_0) \not\to 0.$$

The above result leads to two distinct cases.

Case 1: $\Bigg\{\Bigg(\sum\limits_{j=1}^{s_{i^*}} p^n_{i^*j}\Bigg(\Big|(\Delta\theta^n_{1i^*j})^{(1)}\Big|^{\overline{r}} + \Big|(\Delta\theta^n_{2i^*j})^{(1)}\Big|^{\overline{r}/2}\Bigg)\Bigg\}/D_\kappa(G_n,G_0) \not\to 0$. By taking the product between the inverse of the previous ratio and $E_{l,0}(\theta^0_{1i^*},\theta^0_{2i^*})/D_\kappa(G_n,G_0)$, we achieve the following system of limits

$$\frac{\Bigg(\sum\limits_{j=1}^{s_{i^*}} p^n_{i^*j} \sum\limits_{\substack{\alpha_1+2\alpha_3=l\\ \alpha_1+\alpha_3\leq \overline{r}}} \Big\{(\Delta\theta^n_{1i^*j})^{(1)}\Big\}^{\alpha_1}\Big\{(\Delta\theta^n_{2i^*j})^{(1)}\Big\}^{\alpha_3}\Bigg)/(2^{\alpha_3}\alpha_1!\alpha_3!)}{\Bigg(\sum\limits_{j=1}^{s_{i^*}} p^n_{i^*j}\Bigg(\Big|(\Delta\theta^n_{1i^*j})^{(1)}\Big|^{\overline{r}} + \Big|(\Delta\theta^n_{2i^*j})^{(1)}\Big|^{\overline{r}/2}\Bigg)} \to 0,$$

for all $1 \leq l \leq 2\overline{r}$, which does not hold according to the argument of the proof of Theorem 2. Therefore, Case 1 can not hold.

Case 2: $\Bigg\{\Bigg(\sum\limits_{j=1}^{s_{i^*}} p^n_{i^*j}\Bigg(\Big|(\Delta\theta^n_{1i^*j})^{(2)}\Big|^{\overline{r}} + \Big|(\Delta\theta^n_{2i^*j})^{(2)}\Big|^{\overline{r}/2}\Bigg)\Bigg\}/D_\kappa(G_n,G_0) \not\to 0$. By taking the product between the inverse of the previous ratio with $E_{0,l}(\theta^0_{1i^*},\theta^0_{2i^*})/D_\kappa(G_n,G_0)$, we obtain that following system of limits

$$\frac{\Bigg(\sum\limits_{j=1}^{s_{i^*}} p^n_{i^*j} \sum\limits_{\substack{\alpha_2+2\alpha_4=l\\ \alpha_2+\alpha_4\leq \overline{r}}} \Big\{(\Delta\theta^n_{1i^*j})^{(2)}\Big\}^{\alpha_2}\Big\{(\Delta\theta^n_{2i^*j})^{(2)}\Big\}^{\alpha_4}\Bigg)/(2^{\alpha_4}\alpha_2!\alpha_4!)}{\Bigg(\sum\limits_{j=1}^{s_{i^*}} p^n_{i^*j}\Bigg(\Big|(\Delta\theta^n_{1i^*j})^{(2)}\Big|^{\overline{r}} + \Big|(\Delta\theta^n_{2i^*j})^{(2)}\Big|^{\overline{r}/2}\Bigg)} \to 0,$$

for all $1 \leq l \leq 2\overline{r}$, which does not hold. Thus, Case 2 can not happen.

As a consequence, not all the coefficients of $A_n/D_\kappa(G_n, G_0)$ and $B_n/D_\kappa(G_n, G_0)$ go to 0 as $n \to \infty$. From here, using the same argument as the Fatou's argument in Step 4 of the proof of Theorem 2, we achieve the conclusion of inequality (39).

A.3 Proof for equality (40)

To avoid unnecessary repetition, we only sketch the proof for equality (40). Since $\kappa' = (\kappa'^{(1)}, \kappa'^{(2)}, \kappa'^{(3)}, \kappa'^{(4)}) \prec (\overline{r}, \overline{r}, \overline{r}/2, \overline{r}/2)$, one of the two pairs $(\kappa'^{(1)}, \kappa'^{(3)})$, $(\kappa'^{(2)}, \kappa'^{(4)})$ is strictly dominated by $(\overline{r}, \overline{r}/2)$. Without loss of generality, we assume that $(\kappa'^{(1)}, \kappa'^{(3)}) \prec (\overline{r}, \overline{r}/2)$. Under this setting of κ', we construct a sequence of mixing measures $G_n = \sum_{i=1}^{k} \pi_i^n \delta_{(\theta_{1i}^n, \theta_{2i}^n)}$ as follows. We choose $(\pi_{i+k-k_0}^n, \theta_{1(i+k-k_0)}^n, \theta_{2(i+k-k_0)}^n) = (\pi_i^0, \theta_{1i}^0, \theta_{2i}^0)$ for $2 \leq i \leq k_0$. For $1 \leq j \leq k - k_0 + 1$, we choose $\Big((\theta_{1j}^n)^{(2)}, (\theta_{2j}^n)^{(2)}\Big) \equiv \big((\theta_{11}^0)^{(2)}, (\theta_{21}^0)^{(2)}\big)$ and

$$(\theta_{1j}^n)^{(1)} = (\theta_{11}^0)^{(1)} + \frac{a_j^*}{n}, \ (\theta_{2j}^n)^{(1)} = (\theta_{21}^0)^{(1)} + \frac{2b_j^*}{n^2}, \ \pi_j^n = \frac{\pi_1^0 (c_j^*)^2}{\sum_{i=1}^{k-k_0+1} (c_j^*)^2},$$

where $(c_i^*, a_i^*, b_i^*)_{i=1}^{k-k_0+1}$ are the non-trivial solution of system of polynomial equations (6) when $r = \overline{r} - 1$. From here, by performing Taylor expansion around $((\theta_{11}^0)^{(1)}, (\theta_{21}^0)^{(1)})$, i.e., along the direction of the first component of θ_{11}^0 and θ_{21}^0 and arguing similarly as Case 2 in the proof of equality (29) in Section 5.2.2, we obtain that

$$p_{G_n}(X, Y) - p_{G_0}(X, Y) = \sum_{l=\overline{r}}^{2\overline{r}-2} \mathcal{O}(n^{-\overline{r}}) \frac{\partial^l f}{\partial h_1^l}(Y | h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0)) \overline{f}(X) + \overline{R}(X, Y),$$

where $\overline{R}(X, Y)$ is a Taylor remainder such that the following limit holds

$$\int \frac{\overline{R}^2(X, Y)}{p_{G_0}(X, Y) \widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n, G_0)} d(X, Y) \precsim \frac{\mathcal{O}(n^{-2\overline{r}})}{n^{-2\min\{\kappa'^{(1)}, \kappa'^{(3)}\}}} \to 0.$$

As a consequence, we eventually achieve that

$$h^2(p_{G_n}, p_{G_0}) \Big/ \widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n, G_0) \to 0$$

as $n \to \infty$, which leads to the conclusion of equality (40).

A.4 Proof of Theorem 5

(a) Similar to the proof of Theorem 2, to obtain the conclusion of part (a), it is sufficient to demonstrate that

$$\lim_{\epsilon \to 0} \inf_{G \in \mathcal{O}_{k, \overline{c}_0}(\Omega): W_\kappa(G, G_0) \leq \epsilon} V(p_G, p_{G_0}) / \widetilde{W}_\kappa^{\|\kappa\|_\infty}(G, G_0) > 0 \tag{42}$$

where $\kappa := (2, 2, 2)$. Assume that the above result does not hold, which leads to the existence of sequence $G_n = \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n \delta_{(\theta_{1ij}^n, \theta_{2ij}^n)}$ such that $V(p_{G_n}, p_{G_0}) / \widetilde{W}_\kappa^{\|\kappa\|_\infty}(G_n, G_0) \to 0$ and $\widetilde{W}_\kappa(G_n, G_0) \to 0$. Here, $(\theta_{1ij}^n, \theta_{2ij}^n) \to (\theta_{1i}^0, \theta_{2i}^0)$ for all $1 \leq i \leq k_0, 1 \leq j \leq s_i$ and $\sum_{j=1}^{s_i} p_{ij}^n \to \pi_i^0$

for all $1 \le i \le k_0$. In this proof, we denote $\Delta\theta^n_{1ij} := ((\Delta\theta^n_{1ij})^{(1)}, (\Delta\theta^n_{1ij})^{(2)})$ for all $1 \le i \le k_0$ and $1 \le j \le s_i$. According to Lemma 5 in Appendix B, we have

$$\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G_n, G_0) \precsim \sum_{i=1}^{k_0}\sum_{j=1}^{s_i} p_{ij}^n \left(\left|(\Delta\theta^n_{1ij})^{(1)}\right|^{\kappa^{(1)}} + \left|(\Delta\theta^n_{1ij})^{(2)}\right|^{\kappa^{(2)}} + \left|\Delta\theta^n_{2ij}\right|^{\kappa^{(3)}} \right) + \sum_{i=1}^{k_0}\left|\sum_{j=1}^{s_i} p_{ij}^n - \pi_i^0\right| := D_\kappa(G_n, G_0).$$

Since the proof argument for (42) is rather intricate, we divide this argument into several steps.

Step 1 - Structure of Taylor expansion By means of Taylor expansion up to the order $\|\kappa\|_\infty = 2$, we obtain that

$$\begin{aligned} p_{G_n}(X,Y) - p_{G_0}(X,Y) &= \sum_{i=1}^{k_0}\sum_{j=1}^{s_i} p_{ij}^n \sum_{1\le|\alpha|\le\|\kappa\|_\infty} \frac{1}{\alpha!}\left\{(\Delta\theta^n_{1ij})^{(1)}\right\}^{\alpha_1}\left\{(\Delta\theta^n_{1ij})^{(2)}\right\}^{\alpha_2}(\Delta\theta^n_{2ij})^{\alpha_3} \\ &\times \frac{\partial^{|\alpha|} f}{\partial(\theta_1^{(1)})^{\alpha_1}\partial(\theta_1^{(2)})^{\alpha_2}\partial\theta_2^{\alpha_3}}(Y|h_1(X|\theta^0_{1i}), h_2(X|\theta^0_{2i}))\overline{f}(X) \\ &+ \sum_{i=1}^{k_0}\left(\sum_{j=1}^{s_i} p_{ij}^n - p_i^0\right) f(Y|h_1(X|\theta^0_{1i}, h_2(X|\theta^0_{2i}))\overline{f}(X) + R(X,Y) \\ &:= A_n + B_n + R(X,Y), \end{aligned}$$

where $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ and $R(X,Y)$ is a Taylor remainder such that $R(X,Y)/D_\kappa(G_n, G_0) \to 0$ as $n \to \infty$.

With the formation of expert functions h_1 and h_2, we can check that $\dfrac{\partial^{|\alpha|} f}{\partial(\theta_1^{(1)})^{\alpha_1}\partial(\theta_1^{(2)})^{\alpha_2}\partial\theta_2^{\alpha_3}}(Y|h_1(X|\theta^0_{1i}), h_2(X|\theta^0_{2i}))\overline{f}(X)$ are not linearly independent with respect to X and Y. Therefore, as being argued in the proof of Theorem 2, we can not consider A_n as a linear combinations of these derivatives. To see clearly the influence of non-linearity setting I of G_0 on the set of linear independent elements of A_n, we will provide the detail formulations of key partial derivatives of f with respect to θ_1 and θ_2 up to the second order.

Key partial derivatives up to the second order In particular, for any θ_1 and θ_2, by means of direct computation and the PDE equation $\dfrac{\partial^2 f}{\partial h_1^2} = 2\dfrac{\partial f}{\partial h_2^2}$, we can verify that

$$\frac{\partial f}{\partial\theta_1^{(1)}} = 2\left(\theta_1^{(1)} + \theta_1^{(2)}X\right)\frac{\partial f}{\partial h_1}, \quad \frac{\partial f}{\partial\theta_1^{(2)}} = 2X\left(\theta_1^{(1)} + \theta_1^{(2)}X\right)\frac{\partial f}{\partial h_1},$$
$$\frac{\partial^2 f}{\partial(\theta_1^{(1)})^2} = 2\frac{\partial f}{\partial h_1} + 4\left(\theta_1^{(1)} + \theta_1^{(2)}X\right)^2\frac{\partial^2 f}{\partial h_1^2},$$

$$\frac{\partial^2 f}{\partial (\theta_1^{(2)})^2} = 2X^2 \frac{\partial f}{\partial h_1} + 4X^2 \left(\theta_1^{(1)} + \theta_1^{(2)} X\right)^2 \frac{\partial^2 f}{\partial h_1^2},$$
$$\frac{\partial^2 f}{\partial \theta_1^{(1)} \partial \theta_1^{(2)}} = 2X \frac{\partial f}{\partial h_1} + 4X \left(\theta_1^{(1)} + \theta_1^{(2)} X\right)^2 \frac{\partial^2 f}{\partial h_1^2},$$
$$\frac{\partial f}{\partial \theta_2} = \frac{\partial f}{\partial h_2^2} = \frac{1}{2}\frac{\partial^2 f}{\partial h_1^2}, \ \frac{\partial^2 f}{\partial \theta_2^2} = \frac{\partial^2 f}{\partial h_2^4} = \frac{1}{4}\frac{\partial^4 f}{\partial h_1^4}. \tag{43}$$

Here, we suppress the condition on $h_1(X, \theta_1)$ and $h_2(X, \theta_2)$ in the notation to simplify the presentation.

Set of linear independent elements We define

$$\mathcal{F} := \left\{ X^{l_1} \frac{\partial^{l_2} f}{\partial h_1^{l_2}}(Y|h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0))\overline{f}(X) : \ (l_1, l_2) \in \mathcal{B}, \ 1 \leq i \leq k_0 \right\},$$

where $\mathcal{B} = \{(0,0), (0,1)(0,2), (0,3), (0,4), (1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,2), (4,2)\}$. According to the key partial derivatives of f up to the second order given by (43), we can validate that the elements of $\mathcal{F}$ are linearly independent with respect to X and Y. Therefore, we can treat $A_n/D_\kappa(G_n, G_0)$, $B_n/D_\kappa(G_n, G_0)$ as a linear combination of linear independent elements of $\mathcal{F}$.

Step 2 - Non-vanishing coefficients Assume that all the coefficients in the representation of $A_n/D_\kappa(G_n, G_0)$ and $B_n/D_\kappa(G_n, G_0)$ go to 0 as $n \to \infty$. By taking the summation of the absolute value of coefficients in $B_n/D_\kappa(G_n, G_0)$, it implies that

$$\left(\sum_{i=1}^{k_0} |\sum_{j=1}^{s_i} p_{ij}^n - \pi_i^0| \right) / D_\kappa(G_n, G_0) \to 0.$$

Furthermore, from the formulations of key partial derivatives in (43), the vanishing of coefficients of $\frac{\partial^4 f}{\partial h_1^4}(Y|h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0))$ to 0 as $1 \leq i \leq k_0$ leads to

$$\left(\sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n \left|\Delta\theta_{2ij}^n\right|^2 \right) / D_\kappa(G_n, G_0) \to 0.$$

Given the above results, the following holds

$$\left(\sum_{i=1}^{k_0} \left\{ \sum_{j=1}^{s_i} p_{ij}^n \left|\Delta\theta_{2ij}^n\right|^2 + |\sum_{j=1}^{s_i} p_{ij}^n - \pi_i^0| \right\} \right) / D_\kappa(G_n, G_0) \to 0. \tag{44}$$

On the other hand, the hypothesis that the coefficients of $X^{l_1} \frac{\partial f}{\partial h_1}(Y|h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0))$ as $l_1 \in \{0, 2\}$ and $X^{l_2} \frac{\partial^2 f}{\partial h_1^2}(Y|h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0))$ as $l_2 \in \{0, 2\}$ as $1 \leq l_2 \leq 4$ go to 0 as

$1 \leq i \leq k_0$ respectively lead to the following system of polynomial limits:

$$\begin{aligned}
2(\theta_{1i}^0)^{(1)} \left(\sum_{j=1}^{s_i} p_{ij}^n (\Delta\theta_{1ij}^n)^{(1)} \right) / D_\kappa(G_n, G_0) + 2I_{n,i} &\to 0, \\
2(\theta_{1i}^0)^{(2)} \left(\sum_{j=1}^{s_i} p_{ij}^n (\Delta\theta_{1ij}^n)^{(2)} \right) / D_\kappa(G_n, G_0) + 2K_{n,i} &\to 0, \\
2(\theta_{1i}^0)^{(1)} (\theta_{1i}^0)^{(2)} I_{n,i} + \left\{ (\theta_{1i}^0)^{(1)} \right\}^2 J_{n,i} &\to 0, \\
\left\{ (\theta_{1i}^0)^{(2)} \right\}^2 I_{n,i} + 2(\theta_{1i}^0)^{(1)} (\theta_{1i}^0)^{(2)} J_{n,i} + \left\{ (\theta_{1i}^0)^{(1)} \right\}^2 K_{n,i} &\to 0, \\
2(\theta_{1i}^0)^{(1)} (\theta_{1i}^0)^{(2)} K_{n,i} + \left\{ (\theta_{1i}^0)^{(2)} \right\}^2 J_{n,i} &\to 0, \\
\left\{ (\theta_{1i}^0)^{(2)} \right\} K_{n,i} &\to 0,
\end{aligned} \tag{45}$$

for all $1 \leq i \leq k_0$ where the explicit forms of $I_{n,i}$, $J_{n,i}$, and $K_{n,i}$ are as follows:

$$\begin{aligned}
I_{n,i} &:= \left(\sum_{j=1}^{s_i} p_{ij}^n \left| (\Delta\theta_{1ij}^n)^{(1)} \right|^2 \right) / D_\kappa(G_n, G_0), \\
J_{n,i} &:= \left(\sum_{j=1}^{s_i} p_{ij}^n (\Delta\theta_{1ij}^n)^{(1)} (\Delta\theta_{1ij}^n)^{(2)} \right) / D_\kappa(G_n, G_0), \\
K_{n,i} &:= \left(\sum_{j=1}^{s_i} p_{ij}^n \left| (\Delta\theta_{1ij}^n)^{(2)} \right|^2 \right) / D_\kappa(G_n, G_0).
\end{aligned}$$

According to the formulation of non-linearity setting I of G_0, we only have two possible cases to consider with respect to a pair $\left((\theta_{1i}^0)^{(1)}, (\theta_{1i}^0)^{(2)}\right)$:

Case 1: $(\theta_{1i}^0)^{(2)} \neq 0$. Under this case, the final limit in (45) indicates that $K_{n,i} \to 0$ as $n \to \infty$. Plugging this result into the fifth limit in this system, we achieve that $J_{n,i} \to 0$ as $n \to \infty$. Putting the previous results together, the third limit in the system of limits leads to $I_{n,i} \to 0$ as $n \to \infty$.

Case 2: $(\theta_{1i}^0)^{(1)} = (\theta_{1i}^0)^{(2)} = 0$. Under this case, the final four limits in system of polynomial limits (45) always hold. On the other hand, the first two limits of this system leads to $I_{n,i} \to 0$ and $K_{n,i} \to 0$ as $n \to \infty$.

Given the results from Case 1 and Case 2, the following limit holds

$$\frac{\sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n \left(\left| (\Delta\theta_{1ij}^n)^{(1)} \right|^2 + \left| (\Delta\theta_{1ij}^n)^{(2)} \right|^2 \right)}{D_\kappa(G_n, G_0)} \to 0. \tag{46}$$

Putting the results from (44) and (46) together, we obtain that

$$1 = D_\kappa(G_n, G_0) / D_\kappa(G_n, G_0) \to 0,$$

which is a contradiction. Therefore, not all the coefficients of $A_n/D_\kappa(G_n, G_0)$ and $B_n/D_\kappa(G_n, G_0)$ go to 0 as $n \to \infty$. From here, using the same argument as that of using the Fatou's argument in Step 4 of the proof of Theorem 2, we achieve the conclusion of inequality (42) under non-linearity setting I of G_0.

(b) According to Lemma 3, the proof regarding minimax lower bound for estimators under non-linearity setting I of G_0 follows by demonstrating that

$$\lim_{\epsilon \to 0} \inf_{G \in \mathcal{O}_k(\Omega):\ W_{\kappa'}(G, G_0) \leq \epsilon} h(p_G, p_{G_0}) / \widetilde{W}_{\kappa'}^{\|\kappa'\|_\infty}(G, G_0) = 0, \tag{47}$$

for all $\kappa' \prec \kappa = (2, 2, 2)$. In fact, we will construct a similar sequence of mixing measures G_n as that in the proof of (17) in Section 5.1.2. More precisely, we define $G_n = \sum_{i=1}^{k_0+1} \pi_i^n \delta_{(\theta_{1i}^n, \theta_{2i}^n)}$ with $k_0 + 1$ components as follows: $(\pi_i^n, \theta_{1i}^n, \theta_{2i}^n) \equiv (\pi_{i-1}^0, \theta_{1(i-1)}^0, \theta_{2(i-1)}^0)$ for $3 \leq i \leq k_0 + 1$. Additionally, $\pi_1^n = \pi_2^n = 1/2$, $(\theta_{11}^n, \theta_{21}^n) \equiv (\theta_{11}^0 - \mathbf{1}_2/n, \theta_{21}^0 - 1/n)$, and $(\theta_{12}^n, \theta_{22}^n) \equiv (\theta_{11}^0 + \mathbf{1}_2/n, \theta_{21}^0 + 1/n)$. Now, by means of Taylor expansion up to the first order, the detail formulations of first order derivatives in (43), and the choice of G_n, we have

$$\begin{aligned}
p_{G_n}(X, Y) - p_{G_0}(X, Y) &= \sum_{i=1}^{2} \pi_i^n \left(f(Y|h_1(X, \theta_{1i}^n, \theta_{2i}^n) - f(Y|h_1(X, \theta_{11}^0, \theta_{21}^0)) \right) \overline{f}(X) \\
&= \sum_{i=1}^{2} \pi_i^n \sum_{|\alpha|+|\beta|=1} \frac{1}{\alpha!\beta!} \prod_{u=1}^{q_1} \left\{ (\Delta\theta_{1i}^n)^{(u)} \right\}^{\alpha_u} \prod_{u=1}^{q_1} \left\{ (\Delta\theta_{2i}^n)^{(v)} \right\}^{\beta_v} \\
&\times \frac{\partial f}{\partial \theta_1^\alpha \partial \theta_2^\beta} \left(Y|h_1(X, \theta_{11}^0), h_2(X, \theta_{21}^0) \right) \overline{f}(X) + \overline{R}(X, Y) \\
&= \overline{R}(X, Y),
\end{aligned}$$

where $\Delta\theta_{1i}^n = \theta_{1i}^n - \theta_{11}^0$ and $\Delta\theta_{2i}^n = \theta_{2i}^n - \theta_{21}^0$ for $1 \leq i \leq 2$. Using the similar argument as that in the proof of (17) in Section 5.1.2, $\overline{R}(X, Y)$ is a Taylor remainder from the above expansion such that

$$\int \frac{\overline{R}_1^2(X, Y)}{p_{G_0}(X, Y) \widetilde{W}_{\kappa'}^{2\|\kappa'\|_\infty}(G_n, G_0)} d(X, Y) \precsim \frac{\mathcal{O}(n^{-4})}{n^{-2\min\{\kappa'(1), \kappa'(2), \kappa'(3)\}}} \to 0$$

as $n \to \infty$. As a consequence, we achieve the conclusion of equality (47).

A.5 Proof of Theorem 6

To achieve the conclusion of the theorem, it is sufficient to demonstrate that

$$\lim_{\epsilon \to 0} \inf_{G \in \mathcal{O}_{k, \bar{c}_0(\Omega)}:\ \widetilde{W}_{\widetilde{\kappa}_{\sin}}(G, G_0) \leq \epsilon} V(p_G, p_{G_0}) / \widetilde{W}_{\widetilde{\kappa}_{\sin}}^{\widetilde{\mathrm{r}}_{\sin}}(G, G_0) > 0$$

where $\widetilde{\mathrm{r}}_{\sin} = \widetilde{\mathrm{r}}\left((\theta_{1i_{\max}}^0)^{(1)}, k - k_0 + 1\right)$ and $\widetilde{\kappa}_{\sin} = (\widetilde{\mathrm{r}}_{\sin}, 2, \lceil \widetilde{\mathrm{r}}_{\sin}/2 \rceil)$. Assume that the above result does not hold. It implies that we can find sequence G_n such that

$$V\left(p_{G_n}, p_{G_0}\right) / \widetilde{W}_{\widetilde{\kappa}_{\sin}}^{\widetilde{\mathrm{r}}_{\sin}}(G_n, G_0) \to 0,$$

and $\widetilde{W}_{\widetilde{\kappa}_{\sin}}(G_n, G_0) \to 0$. To avoid unnecessary repetition, we utilize the same notation of G_n as part (a) in the proof of Theorem 5 in Appendix A.4.

Step 1 - Structure of Taylor expansion Similar to the proof of Theorem 5, we have the following representation when we perform Taylor expansion up to the order $\widetilde{\mathrm{r}}_{\sin}$:

$$p_{G_n}(X,Y) - p_{G_0}(X,Y) := A_n + B_n + R(X,Y),$$

where $R(X,Y)$ is a Taylor remainder such that $R(X,Y)/D_{\widetilde{\kappa}_{\sin}}(G_n,G_0) \to 0$ as $n \to \infty$. The forms of B_n and $D_{\widetilde{\kappa}_{\sin}}(G_n,G_0)$ are similar to that in Step 1 of Theorem 5 except that we use $\widetilde{\kappa}_{\sin}$ instead of κ. Furthermore, A_n has the following form:

$$A_n := \sum_{i=1}^{k_0} A_n(i) = \sum_{i \in \mathcal{A}} A_n(i) + \sum_{i \in \mathcal{A}^c} A_n(i),$$

where $\mathcal{A} := \{i \in [k_0] : \ (\theta_{1i}^0)^{(1)} \neq 0 \text{ and } (\theta_{1i}^0)^{(2)} = 0\}$ and

$$A_n(i) := \sum_{j=1}^{s_i} p_{ij}^n \sum_{1 \leq |\alpha| \leq \widetilde{\mathrm{r}}_{\sin}} \frac{1}{\alpha!} \left\{(\Delta\theta_{1ij}^n)^{(1)}\right\}^{\alpha_1} \left\{(\Delta\theta_{1ij}^n)^{(2)}\right\}^{\alpha_2} (\Delta\theta_{2ij}^n)^{\alpha_3}$$
$$\times \frac{\partial^{|\alpha|} f}{\partial (\theta_1^{(1)})^{\alpha_1} \partial (\theta_1^{(2)})^{\alpha_2} \partial \theta_2^{\alpha_3}} \left(Y | h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0)\right) \overline{f}(X)$$

for $i \in [k_0]$. Under the non-linearity setting II of G_0, there exists an index i such that $(\theta_{1i}^0)^{(1)} \neq 0$ and $(\theta_{1i}^0)^{(2)} = 0$. Therefore, we have $|\mathcal{A}| \geq 1$. To analyze the structure of $A_n(i)$, we consider two settings of index i: $i \in \mathcal{A}$ and $i \in \mathcal{A}^c$.

Index $i \in \mathcal{A}$: For any $i \in \mathcal{A}$, the collection of full partial derivatives $\dfrac{\partial^{|\alpha|} f}{\partial (\theta_1^{(1)})^{\alpha_1} \partial (\theta_1^{(2)})^{\alpha_2} \partial \theta_2^{\alpha_3}}$ $(Y | h_1(X|\theta_{1i}^0), h_2(X|\theta_{2i}^0))\overline{f}(X)$ up to order $\widetilde{\mathrm{r}}_{\sin} \geq 2$ is not linearly independent with respect to X and Y. Therefore, we cannot treat $A_n(i)$ as a linear combination of these derivatives as long as $i \in \mathcal{A}$. Our strategy is to reduce this collection of full partial derivatives into a collection of linearly independent terms of the forms $X^{l_1} \dfrac{\partial^{l_2} f}{\partial h_1^{l_2}}(Y | h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0))\overline{f}(X)$ as those in the previous proofs for some (l_1, l_2). Given that idea, we define

$$\mathcal{F}(i) = \left\{ X^{l_1} \frac{\partial^{l_2} f}{\partial h_1^{l_2}}(Y | h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0))\overline{f}(X) : (l_1, l_2) \in \mathcal{B}(i) \right\},$$

the set of all linear independent terms deriving from computing the partial derivatives of f up to order $\widetilde{\mathrm{r}}_{\sin}$ with respect to θ_1 and θ_2. In general, the exact form of $\mathcal{B}(i)$ is very difficult to obtain. For the purpose of this proof, we only need to focus on a subset of $\mathcal{B}(i)$ in which we have a closed form. In particular, we denote a set $\mathcal{B}_{\text{sub}}$ as follows:

$$\mathcal{B}_{\text{sub}} := \{(2,0)\} \cup \{(0, l_2) : \ 1 \leq l_2 \leq 2\widetilde{\mathrm{r}}_{\sin}\}.$$

We claim that $\mathcal{B}_{\text{sub}}$ is a subset of $\mathcal{B}(i)$ for any $i \in \mathcal{A}$. We prove this claim at the end of this proof. From now on, we assume that this claim is given.

Index $i \in \mathcal{A}^c$: For any $i \in \mathcal{A}^c$, we also have the linear dependence of the set of full partial derivatives $\frac{\partial^{|\alpha|} f}{\partial (\theta_1^{(1)})^{\alpha_1} \partial (\theta_1^{(2)})^{\alpha_2} \partial \theta_2^{\alpha_3}}(Y|h_1(X|\theta_{1i}^0), h_2(X|\theta_{2i}^0))\overline{f}(X)$ up to order $\widetilde{\mathrm{r}}_{\sin} \geq 2$. Similar to the strategy of case $i \in \mathcal{A}$, we also reduce the previous set into a collection of linearly independent terms of the forms $X^{l_1}\frac{\partial^{l_2} f}{\partial h_1^{l_2}}(Y|h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0))\overline{f}(X)$, which can be defined as:

$$\overline{\mathcal{F}}(i) = \left\{ X^{l_1}\frac{\partial^{l_2} f}{\partial h_1^{l_2}}(Y|h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0))\overline{f}(X) : (l_1, l_2) \in \overline{\mathcal{B}}(i) \right\}.$$

However, the structure of $\overline{\mathcal{B}}(i)$ is also very complicated. For the purpose of this proof, we only consider its subset $\overline{\mathcal{B}}_{\text{sub}}$, which has the following form:

$$\overline{\mathcal{B}}_{\text{sub}} := \{(0,1), (0,4), (1,2), (2,1), (2,2), (2,3), (3,2), (4,2)\}.$$

The proof for the claim that $\overline{\mathcal{B}}_{\text{sub}} \subset \overline{\mathcal{B}}(i)$ for any $i \in \mathcal{A}^c$ is similar to that from claim $\mathcal{B}_{\text{sub}} \subset \mathcal{B}(i)$ as $i \in \mathcal{A}$; therefore, it is omitted. From now on, we also assume that the above claim is true.

Given the formulations of $\mathcal{F}(i)$ and $\overline{\mathcal{F}}(i)$, we can treat $A_n(i)/D_{\widetilde{\kappa}_{\sin}}(G_n, G_0)$ as the linear combinations of elements from $\mathcal{F}(i)$ divided by $D_{\widetilde{\kappa}_{\sin}}(G_n, G_0)$ for $i \in \mathcal{A}$ and from $\overline{\mathcal{F}}(i)$ divided by $D_{\widetilde{\kappa}_{\sin}}(G_n, G_0)$ for $i \in \mathcal{A}^c$.

Non-vanishing coefficients Similar to the previous proofs, we assume that all the coefficients in the representation of $A_n(i)/D_{\widetilde{\kappa}_{\sin}}(G_n, G_0)$ and $B_n/D_{\widetilde{\kappa}_{\sin}}(G_n, G_0)$ go to 0 as $n \to \infty$ for all $i \in [k_0]$. From the definitions of $\mathcal{B}_{\text{sub}}$ and $\overline{\mathcal{B}}_{\text{sub}}$, we have the coefficients associated with $X^{l_1}\frac{\partial^{l_2} f}{\partial h_1^{l_2}}(Y|h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0))\overline{f}(X)$ go to 0 when $(l_1, l_2) \in \mathcal{B}_{\text{sub}}$ for $i \in \mathcal{A}$ or $(l_1, l_2) \in \overline{\mathcal{B}}_{\text{sub}}$ for $i \in \mathcal{A}^c$.

For the simplicity of the presentation, we denote $E_{(l_1,l_2)}(i)$ the coefficients of the element $X^{l_1}\frac{\partial^{l_2} f}{\partial h_1^{l_2}}(Y|h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0))\overline{f}(X)$ when $(l_1, l_2) \in \mathcal{B}_{\text{sub}}$ and $i \in \mathcal{A}$. Similarly, $\overline{E}_{(l_1,l_2)}(i)$ are the coefficients of $X^{l_1}\frac{\partial^{l_2} f}{\partial h_1^{l_2}}(Y|h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0))\overline{f}(X)$ when $(l_1, l_2) \in \overline{\mathcal{B}}_{\text{sub}}$ and $i \in \mathcal{A}^c$.

For $(l_1, l_2) = (0, l)$ as $1 \leq l \leq 2\widetilde{\mathrm{r}}_{\sin}$, the exact formulation of $E_{(l_1,l_2)}(i)$ can be derived from determining the coefficient of $X^{l_1}\frac{\partial^{l_2} f}{\partial h_1^{l_2}}(Y|h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0))\overline{f}(X)$ in the following term:

$$\sum_{j=1}^{s_i} p_{ij}^n \sum_{1 \leq |\gamma| \leq \widetilde{\mathrm{r}}_{\sin}} \frac{1}{\gamma!}\left\{(\Delta\theta_{1ij}^n)^{(1)}\right\}^{\gamma_1}(\Delta\theta_{2ij}^n)^{\gamma_2}\frac{\partial^{|\gamma|} f}{\partial (\theta_1^{(1)})^{\gamma_1}\partial\theta_2^{\gamma_2}}\left(Y|h_1(X, \theta_{1i}^0), h_2(X, \theta_{2i}^0)\right)\overline{f}(X),$$

for $\gamma = (\gamma_1, \gamma_2)$. Equipped with the result of Lemma 1, we can verify that

$$E_{(0,l)}(i) = \left[\sum_{\gamma_1,\gamma_2,\tau} \frac{P_\tau^{(\gamma_1)}\left((\theta_{1i}^0)^{(1)}\right)}{2^{\gamma_2}}\left(\sum_{j=1}^{s_i} p_{ij}^n \frac{\left\{(\Delta\theta_{1ij}^n)^{(1)}\right\}^{\gamma_1}(\Delta\theta_{2ij}^n)^{\gamma_2}}{\gamma_1!\gamma_2!}\right)\right] \Big/ D_{\widetilde{\kappa}_{\sin}}(G_n, G_0), \quad (48)$$

where the summation with respect to γ_1, γ_2, τ in the numerator satisfies $\gamma_1/2 + \tau + 2\gamma_2 = l$, $\tau \leq \gamma_1/2$ when γ_1 is an even number while $(\gamma_1+1)/2 + \tau + 2\gamma_2 = l$, $\tau \leq (\gamma_1-1)/2$ when γ_1 is an odd number. Furthermore, $\gamma_1 + \gamma_2 \leq \widetilde{\mathrm{r}}_{\sin}$.

By taking the summation of the absolute value of coefficients in $B_n/D_\kappa(G_n, G_0)$, it implies that

$$\left(\sum_{i=1}^{k_0} |\sum_{j=1}^{s_i} p_{ij}^n - \pi_i^0| \right) / D_{\widetilde{\kappa}_{\sin}}(G_n, G_0) \to 0.$$

From the definition of $D_{\widetilde{\kappa}_{\sin}}(G_n, G_0)$, it leads to

$$\left[\sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n \left(\left| (\Delta\theta_{1ij}^n)^{(1)} \right|^{\widetilde{\mathrm{r}}_{\sin}} + \left| (\Delta\theta_{1ij}^n)^{(2)} \right|^2 + \left| \Delta\theta_{2ij}^n \right|^{\lceil \widetilde{\mathrm{r}}_{\sin}/2 \rceil} \right) \right] \Big/ D_{\widetilde{\kappa}_{\sin}}(G_n, G_0) \to 1.$$

Therefore, there exists an index $i^* \in [k_0]$ such that

$$\left[\sum_{j=1}^{s_{i^*}} p_{i^*j}^n \left(\left| (\Delta\theta_{1i^*j}^n)^{(1)} \right|^{\widetilde{\mathrm{r}}_{\sin}} + \left| (\Delta\theta_{1i^*j}^n)^{(2)} \right|^2 + \left| \Delta\theta_{2i^*j}^n \right|^{\lceil \widetilde{\mathrm{r}}_{\sin}/2 \rceil} \right) \right] \Big/ D_{\widetilde{\kappa}_{\sin}}(G_n, G_0) \not\to 0.$$

We denote

$$\overline{D}_{\widetilde{\kappa}_{\sin}}(G_n, G_0) := \sum_{j=1}^{s_{i^*}} p_{i^*j}^n \left(\left| (\Delta\theta_{1i^*j}^n)^{(1)} \right|^{\widetilde{\mathrm{r}}_{\sin}} + \left| (\Delta\theta_{1i^*j}^n)^{(2)} \right|^2 + \left| \Delta\theta_{2i^*j}^n \right|^{\lceil \widetilde{\mathrm{r}}_{\sin}/2 \rceil} \right).$$

As $E_{(l_1,l_2)}(i) \to 0$ and $\overline{E}_{(l_1',l_2')}(j) \to 0$ for $i \in \mathcal{A}$, $j \in \mathcal{A}^c$, $(l_1, l_2) \in \mathcal{B}_{\mathrm{sub}}$, and $(l_1', l_2') \in \overline{\mathcal{B}}_{\mathrm{sub}}$, the following holds:

$$K_{(l_1,l_2)}(i) := \frac{D_{\widetilde{\kappa}_{\sin}}(G_n, G_0)}{\overline{D}_{\widetilde{\kappa}_{\sin}}(G_n, G_0)} E_{(l_1,l_2)}(i) \to 0,$$
$$\overline{K}_{(l_1',l_2')}(j) := \frac{D_{\widetilde{\kappa}_{\sin}}(G_n, G_0)}{\overline{D}_{\widetilde{\kappa}_{\sin}}(G_n, G_0)} \overline{E}_{(l_1',l_2')}(j) \to 0,$$

for all $i \in \mathcal{A}$, $j \in \mathcal{A}^c$, $(l_1, l_2) \in \mathcal{B}_{\mathrm{sub}}$, and $(l_1', l_2') \in \overline{\mathcal{B}}_{\mathrm{sub}}$. Now, we consider two possible settings of i^*.

Setting 1 - $i^* \in \mathcal{A}$: By direct computation, the vanishing of $K_{(2,0)}(i^*)$ to 0 is equivalent to

$$\left(\sum_{j=1}^{s_{i^*}} p_{i^*j}^n \left| (\Delta\theta_{1i^*j}^n)^{(2)} \right|^2 \right) \Big/ \overline{D}_{\widetilde{\kappa}_{\sin}} \to 0.$$

From the definition of $\widetilde{\kappa}_{\sin}$, the above result leads to

$$L_n := \sum_{j=1}^{s_{i^*}} p_{i^*j}^n \left(\left| (\Delta\theta_{1i^*j}^n)^{(1)} \right|^{\widetilde{\mathrm{r}}_{\sin}} + \left| \Delta\theta_{2i^*j}^n \right|^{\lceil \widetilde{\mathrm{r}}_{\sin}/2 \rceil} \right) \Big/ \overline{D}_{\widetilde{\kappa}_{\sin}} \to 1.$$

Equipped with the formulation of $E_{(0,l)}(i^*)$ in (48) for any $1 \leq l \leq 2\widetilde{r}_{\sin}$, the following system of limits holds:

$$\frac{1}{L_n}K_{(0,l)}(i^*) = \frac{\sum\limits_{\gamma_1,\gamma_2,\tau} \frac{P_\tau^{(\gamma_1)}\left((\theta^0_{1i^*})^{(1)}\right)}{2^{\gamma_2}} \left(\sum\limits_{j=1}^{s_i} p^n_{i^*j} \frac{\left\{(\Delta\theta^n_{1i^*j})^{(1)}\right\}^{\gamma_1} (\Delta\theta^n_{2i^*j})^{\gamma_2}}{\gamma_1!\gamma_2!} \right)}{\sum\limits_{j=1}^{s_{i^*}} p^n_{i^*j} \left(\left|(\Delta\theta^n_{1i^*j})^{(1)}\right|^{\widetilde{\mathrm{r}}_{\sin}} + \left|\Delta\theta^n_{2i^*j}\right|^{\lceil \widetilde{\mathrm{r}}_{\sin}/2 \rceil} \right)} \to 0, \tag{49}$$

where the summation with respect to γ_1, γ_2, τ in the numerator satisfies $\gamma_1/2 + \tau + 2\gamma_2 = l$, $\tau \leq \gamma_1/2$ when γ_1 is an even number while $(\gamma_1+1)/2 + \tau + 2\gamma_2 = l$, $\tau \leq (\gamma_1-1)/2$ when γ_1 is an odd number. Additionally, $\gamma_1 + \gamma_2 \leq \widetilde{\mathrm{r}}_{\sin}$.

Recall that $\widetilde{\mathrm{r}}_{\sin} = \widetilde{\mathrm{r}}((\theta^0_{1i_{\max}})^{(1)}, k-k_0+1)$ where $i_{\max} = \arg\max_{i\in\mathcal{A}} \widetilde{\mathrm{r}}((\theta^0_{1i})^{(1)}, k-k_0+1)$. Therefore, $\widetilde{\mathrm{r}}_{\sin} \geq \widetilde{\mathrm{r}}((\theta^0_{1i^*})^{(1)}, k-k_0+1) \geq \widetilde{\mathrm{r}}((\theta^0_{1i^*})^{(1)}, s_{i^*})$ as $s_{i^*} \leq k-k_0+1$. From the definition of $\widetilde{\mathrm{r}}((\theta^0_{1i^*})^{(1)}, s_{i^*})$ in Definition 4, the system of polynomial limit (49) does not hold given the values of $\widetilde{\mathrm{r}}((\theta^0_{1i^*})^{(1)}, s_{i^*})$. Therefore, it does not happen under $\widetilde{\mathrm{r}}_{\sin}$. As a consequence, setting 1 that $i^* \in \mathcal{A}$ will not hold.

Setting 2 - $i^* \in \mathcal{A}^c$: Since $\widetilde{\mathrm{r}}_{\sin} \geq 3$, it is clear that $(2,2,2) \prec (\widetilde{\mathrm{r}}_{\sin}, 2, \lceil \widetilde{\mathrm{r}}_{\sin}/2 \rceil)$. It implies that

$$\overline{D}_{\widetilde{\kappa}_{\sin}}(G_n, G_0) \lesssim \widetilde{D}(G_n, G_0) := \sum_{j=1}^{s_{i^*}} p^n_{i^*j} \left(\left|(\Delta\theta^n_{1i^*j})^{(1)}\right|^2 + \left|(\Delta\theta^n_{1i^*j})^{(2)}\right|^2 + \left|\Delta\theta^n_{2i^*j}\right|^2 \right).$$

Since $\overline{E}_{(l_1,l_2)}(i^*) \to 0$ for all $(l_1,l_2) \in \overline{\mathcal{B}}_{\mathrm{sub}}$, it leads to

$$\overline{F}_{(l_1,l_2)}(i^*) := \frac{\overline{D}_{\widetilde{\kappa}_{\sin}}(G_n, G_0)}{\widetilde{D}(G_n, G_0)} \overline{E}_{(l_1,l_2)}(i^*) \to 0,$$

for all $(l_1,l_2) \in \overline{\mathcal{B}}_{\mathrm{sub}}$. We can check that the vanishing of $\overline{F}_{(0,4)}(i^*)$ to 0 leads to

$$\left(\sum_{j=1}^{s_{i^*}} p^n_{i^*j} \left|(\Delta\theta^n_{1i^*j})^{(2)}\right|^2 \right) \Big/ \widetilde{D}(G_n, G_0) \to 0. \tag{50}$$

Furthermore, the vanishings of $\overline{F}_{(l_1,l_2)}(i^*)$ to 0 for $(l_1,l_2) \in \{(0,1),(1,2),(2,1),(2,2),(3,2),(4,2)\}$ lead to the system of polynomial limits similar to (45) where the index i in this system is replaced by i^* and the distance $D_\kappa(G_n, G_0)$ is replaced by $\widetilde{D}(G_n, G_0)$. Due to the fact that $i^* \in \mathcal{A}^c$, following the argument after (45), we obtain that

$$\sum_{j=1}^{s_{i^*}} p^n_{i^*j} \left(\left|(\Delta\theta^n_{1i^*j})^{(1)}\right|^2 + \left|\Delta\theta^n_{2i^*j}\right|^2 \right) \Big/ \widetilde{D}(G_n, G_0) \to 0. \tag{51}$$

Invoking the results from (50) and (51) leads to

$$1 = \sum_{j=1}^{s_{i^*}} p^n_{i^*j} \left(\left|(\Delta\theta^n_{1i^*j})^{(1)}\right|^2 + \left|(\Delta\theta^n_{1i^*j})^{(2)}\right|^2 + \left|\Delta\theta^n_{2i^*j}\right|^2 \right) \Big/ \widetilde{D}(G_n, G_0) \to 0,$$

which is a contradiction. Therefore, setting 2 that $i^* \in \mathcal{A}^c$ will not hold.

As a consequence, not all the coefficients of $A_n(i)/D_{\widetilde{\kappa}_{\sin}}(G_n, G_0)$ and $B_n/D_{\widetilde{\kappa}_{\sin}}(G_n, G_0)$ go to 0 as $n \to \infty$ for all $i \in [k_0]$. From here, by means of the Fatou's argument as that of the previous proofs, we achieve the conclusion regarding the convergence rate of MLE under non-linearity setting II of G_0.

Proof of claim $\mathcal{B}_{\text{sub}} \subset \mathcal{B}(i)$ for any $i \in \mathcal{A}$: First of all, we demonstrate that the elements $X^{l_1}\dfrac{\partial^{l_2} f}{\partial h_1^{l_2}}(Y|h_1(X,\theta_{1i}^0), h_2(X,\theta_{2i}^0))\overline{f}(X)$ where $(l_1, l_2) \in \mathcal{B}_{\text{sub}}$ are originated from some partial derivatives of f with respect to θ_1 and θ_2. In fact, by means of Lemma 1, the pairs of indices $(l_1, l_2) = (0, l) \in \mathcal{B}_{\text{sub}}$ for $1 \leq l \leq 2\widetilde{\mathrm{r}}_{\sin}$ correspond to the elements coming from the partial derivatives $\dfrac{\partial^{|\gamma|} f}{\partial(\theta_1^{(1)})^{\gamma_1}\partial\theta_2^{\gamma_2}}(Y|h_1(X,\theta_1^0), h_2(X,\theta_2^0))$ for $1 \leq |\gamma| \leq \widetilde{\mathrm{r}}_{\sin}$. Additionally, the pair $(2, 0) \in \mathcal{B}_{\text{sub}}$ is associated with element from the derivation of $\dfrac{\partial^2 f}{\partial(\theta_1^{(2)})^2}(Y|h_1(X,\theta_1^0), h_2(X,\theta_2^0))$.

Furthermore, it is not hard to verify that the collection of $X^{l_1}\dfrac{\partial^{l_2} f}{\partial h_1^{l_2}}(Y|h_1(X,\theta_{1i}^0), h_2(X,\theta_{2i}^0))$ $\overline{f}(X)$ for $(l_1, l_2) \in \mathcal{B}_{\text{sub}}$ is linearly independent with respect to X and Y. Therefore, we achieve the conclusion that $\mathcal{B}_{\text{sub}} \subset \mathcal{B}(i)$.

B Auxiliary results

In this appendix, we provide two lemmas for the whole results in the paper. To streamline the discussion, we recall that $G_0 = \sum_{i=1}^{k_0} \pi_i^0 \delta_{(\theta_{1i}^0, \theta_{2i}^0)}$ is the true mixing measure with exactly k_0 components such that $\theta_{ji}^0 \in \Omega_j$ for all $1 \leq j \leq 2$ and $1 \leq i \leq k_0$ where $\Omega_j \subset \mathbb{R}^{q_j}$ are compact sets for some given $q_j \geq 1$ as $1 \leq j \leq 2$. Furthermore, $\Omega = \Omega_1 \times \Omega_2$.

Lemma 4. *Assume that $\kappa \in \mathbb{N}^{q_1+q_2}$ is a given vector order of generalized transportation distance and $k > k_0$. For any sequence $G_n \in \mathcal{O}_k(\Omega)$ such that $\widetilde{W}_\kappa(G_n, G_0) \to 0$ as $n \to \infty$, we can find a subsequence of G_n (by which we replace by the whole sequence G_n for the simplicity of presentation) that has the following properties:*

(a) (Fixed number of components) G_n has exactly $\overline{k}$ number of components where $k_0 + 1 \leq \overline{k} \leq k$.

(b) (Universal representation) G_n can be represented as:

$$G_n = \sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} p_{ij}^n \delta_{(\theta_{1ij}^n, \theta_{2ij}^n)},$$

where $\bar{l} \geq 0$ is some non-negative integer number and $s_i \geq 1$ for $1 \leq i \leq k_0 + \bar{l}$ such that $\sum_{i=1}^{k_0+\bar{l}} s_i = \overline{k}$. Furthermore, $(\theta_{1ij}^n, \theta_{2ij}^n) \to (\theta_{1i}^0, \theta_{2i}^0)$ and $\sum_{j=1}^{s_i} p_{ij}^n \to \pi_i^0$ for all $1 \leq i \leq k_0 + \bar{l}$. Here, $\pi_i^0 = 0$ as $k_0 + 1 \leq i \leq \overline{k}$ while $(\theta_{1i}^0, \theta_{2i}^0)$ are extra limit points from the convergence of components of G_n as $k_0 + 1 \leq i \leq \overline{k}$.

Lemma 5. *Given the assumptions with G_0 and G_n as those in Lemma 4, we denote $\eta_i^0 = (\theta_{1i}^0, \theta_{2i}^0)$ and $\eta_{ij}^n = \left(\theta_{1ij}^n, \theta_{2ij}^n\right)$ for $1 \leq i \leq k_0$ and $1 \leq j \leq s_i$. For any $\kappa \in \mathbb{N}^{q_1+q_2}$, we define the following distance:*

$$D_\kappa(G_n, G_0) := \sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} p_{ij}^n d_\kappa^{\|\kappa\|_\infty}\left(\eta_{ij}^n, \eta_i^0\right) + \sum_{i=1}^{k_0+\bar{l}} \left| \sum_{j=1}^{s_i} p_{ij}^n - \pi_i^0 \right|$$

where the pseudo-metric $d_\kappa(.,.)$ is defined as in Section 1.2. Then, the following holds:

$$\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G_n, G_0) \precsim D_\kappa(G_n, G_0).$$

The proofs of the above lemmas are similar to those in [6]; therefore, they are omitted.

C Convergence rate of density estimation

In this appendix, we provide a proof for convergence rate of density estimation of over-specified GMCF in Proposition 3. Our proof technique follows standard result on density estimation for M-estimators in [28]. To ease the presentation, we adapt several notion from the empirical process theory into the setting of over-specified GMCF.

C.1 Key notation and results

We denote $\mathcal{P}_k(\Omega) := \{p_G(X,Y) : \ G \in \mathcal{O}_k(\Omega)\}$. Additionally, we define $N\left(\epsilon, \mathcal{P}_k(\Omega), \|.\|_\infty\right)$ as the covering number of metric space $(\mathcal{P}_k(\Omega), \|.\|_\infty)$ and $H_B(\epsilon, \mathcal{P}_k(\Omega), h)$ as the bracketing entropy of $\mathcal{P}_k(\Omega)$ under Hellinger distance h. We start with the following result regarding the upper bounds of these terms.

Lemma 6. *Suppose that Ω_1 and Ω_2 are respectively two bounded subsets of $\mathbb{R}^{q_1}$ and $\mathbb{R}^{q_2}$. Then, for any $0 < \epsilon < 1/2$, the following results hold*

$$\log N\left(\epsilon, \mathcal{P}_k(\Omega), \|.\|_\infty\right) \precsim \log(1/\epsilon), \tag{52}$$

$$H_B(\epsilon, \mathcal{P}_k(\Omega), h) \precsim \log(1/\epsilon). \tag{53}$$

The detail proof of Lemma 6 is deferred to Appendix C.3. To utilize the above bounds with covering number and bracketing entropy of $\mathcal{P}_k(\Omega)$, we will resort to Theorem 7.4 of [28] for density estimation with MLE. In particular, we denote the following key notation:

$$\overline{\mathcal{P}}_k(\Omega) := \{p_{(G+G_0)/2}(X,Y) : G \in \mathcal{O}_k(\Omega)\}, \ \overline{\mathcal{P}}_k^{1/2}(\Omega) := \{p_{(G+G_0)/2}^{1/2}(X,Y) : G \in \mathcal{O}_k(\Omega)\}.$$

For any $\delta > 0$, we define the Hellinger ball centered around $p_{G_0}(X,Y)$ and intersected with $\overline{\mathcal{P}}_k^{1/2}(\Omega)$ as follows:

$$\overline{\mathcal{P}}_k^{1/2}(\Omega, \delta) := \{f^{1/2} \in \overline{\mathcal{P}}_k^{1/2}(\Omega) : h(f, p_{G_0}) \leq \delta\}.$$

Furthermore, the size of this set can be captured by the following integral:

$$\mathcal{J}_B\left(\delta, \overline{\mathcal{P}}_k^{1/2}(\Omega, \delta)\right) := \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}\left(u, \overline{\mathcal{P}}_k^{1/2}(\Omega, u), \|.\|_2\right) du \vee \delta.$$

Equipped with the above notation, the results from Theorem 7.4 of [28] regarding convergence rates of density estimation from MLE can be formulated as follows.

Theorem 7. *Take* $\Psi(\delta) \geq \mathcal{J}_B\left(\delta, \overline{\mathcal{P}}_k^{1/2}(\Omega, \delta)\right)$ *in such a way that* $\Psi(\delta)/\delta^2$ *is a non-decreasing of* δ*. Then, for a universal constant* c *and for*

$$\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n),$$

we have for all $\delta \geq \delta_n$ *that*

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

C.2 Proof for Proposition 3

Given Theorem 7, we are ready to finish the proof of Proposition 3. In fact, we have

$$H_B\left(u, \overline{\mathcal{P}}_k^{1/2}(\Omega, u), \|.\|_2\right) \leq H_B\left(u, \mathcal{P}(\Omega, u), h\right), \tag{54}$$

for any $u > 0$. The above inequality leads to

$$\begin{aligned}\mathcal{J}_B\left(\delta, \overline{\mathcal{P}}_k^{1/2}(\Omega, \delta)\right) &\leq \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}\left(u, \mathcal{P}_k(\Omega, u), h\right) du \vee \delta \\ &\precsim \int_{\delta^2/2^{13}}^{\delta} \log(1/u) du \vee \delta,\end{aligned}$$

where the second inequality is due to (53) in Lemma 6. Therefore, we can choose $\Psi(\delta) = \delta\left(\log(1/\delta)^{1/2}\right.$ such that $\Psi(\delta) \geq \mathcal{J}_B\left(\delta, \overline{\mathcal{P}}_k^{1/2}(\Omega, \delta)\right)$. From here, with $\delta_n = \mathcal{O}\left([\log n/n]^{1/2}\right)$, the result of Theorem 7 indicates that

$$\mathbb{P}(h(p_{\widehat{G}_n}, p_{G_0}) > C(\log n/n)^{1/2}) \precsim \exp(-c\log n)$$

for some universal positive constants C and c that depend only on Ω. As a consequence, we reach the conclusion of Proposition 3.

C.3 Proof for Lemma 6

The proof of the lemma follows the argument of Theorem 3.1 in [4]. To facilitate the proof argument, our proof is divided into two parts.

Proof for covering number bound (52) For any set $\mathcal{E}$, we denote $\mathcal{E}_\epsilon$ an ϵ-net of $\mathcal{E}$ if each element of $\mathcal{E}$ is within ϵ distance from some elements of $\mathcal{E}_\epsilon$. Since Ω_1 and Ω_2 are two bounded subsets of $\mathbb{R}^{q_1}$ and $\mathbb{R}^{q_2}$ respectively, there exist corresponding ϵ-nets $\overline{\Omega}_1(\epsilon)$ and $\overline{\Omega}_2(\epsilon)$ of these sets with M_1 and M_2 elements. We can validate that

$$M_1 \leq c_1(q_1, k, \Omega_1)\left(\frac{1}{\epsilon}\right)^{q_1 k}, \; M_2 \leq c_2(q_2, k, \Omega_2)\left(\frac{1}{\epsilon}\right)^{q_2 k},$$

where $c_i(q_i, k, \Omega_i)$ are universal constants depending only on q_i, k, Ω_i for $1 \leq i \leq 2$. Furthermore, we denote $\Delta(\epsilon)$ an ϵ-net for k-dimensional simplex. It is known that the cardinality of $\Delta(\epsilon)$ is upper bounded by $(5/\epsilon)^k$. We denote

$$\mathcal{S} := \{p_G \in \mathcal{P}_k(\Omega) : \text{ weights and components of } G \text{ are on } \Delta(\epsilon) \times \overline{\Omega}_1(\epsilon) \times \overline{\Omega}_2(\epsilon)\}.$$

For each $p_G \in \mathcal{P}_k(\Omega)$ where $G = \sum_{i=1}^{k'} \pi_i \delta_{(\theta_{1i},\theta_{2i})}$ such that $k' \leq k$, we denote $\overline{G} = \sum_{i=1}^{k'} \pi_i \delta_{(\theta^*_{1i},\theta^*_{2i})}$ such that $(\theta^*_{1i}, \theta^*_{2i}) \in \overline{\Omega}_1(\epsilon) \times \overline{\Omega}_2(\epsilon)$ and $(\theta^*_{1i}, \theta^*_{2i})$ are the closest points to $(\theta_{1i}, \theta_{2i})$ in this set for $1 \leq i \leq k'$. Additionally, we denote $G^* = \sum_{i=1}^{k'} \pi^*_i \delta_{(\theta^*_{1i},\theta^*_{2i})}$ where $\pi^*_i \in \Delta(\epsilon)$ and π^* are the closest points to π_i in this set for $1 \leq i \leq k'$. From the formulation of G^*, it is clear that $p_{G^*} \in \mathcal{S}$. Invoking triangle inequality with sup-norm, the following inequality holds:

$$\|p_G(X,Y) - p_{G_*}(X,Y)\|_\infty \leq \|p_G(X,Y) - p_{\overline{G}}(X,Y)\|_\infty + \|p_{\overline{G}}(X,Y) - p_{G_*}(X,Y)\|_\infty.$$

According to the definition of $\overline{G}$ and G^*, direct computation leads to

$$\|p_{\overline{G}}(X,Y) - p_{G_*}(X,Y)\|_\infty \leq \sum_{i=1}^{k'} |\pi^*_i - \pi_i| \, \|f\left(Y|h_1(X,\theta^*_{1i}), h_2(X,\theta^*_{2i})\right) \overline{f}(X)\|_\infty \precsim \epsilon. \quad (55)$$

Furthermore, given the formulation of $\overline{G}$, we obtain that

$$\begin{aligned}
\|p_G(X,Y) - p_{\overline{G}}(X,Y)\|_\infty &\leq \sum_{i=1}^{k'} \pi_i \|\overline{f}(X)\big[f\left(Y|h_1(X,\theta^*_{1i}), h_2(X,\theta^*_{2i})\right) \\
&\qquad - f\left(Y|h_1(X,\theta_{1i}), h_2(X,\theta_{2i})\right)\big]\|_\infty \\
&\precsim \sum_{i=1}^{k'} \pi_i \left(\|\theta^*_{1i} - \theta_{1i}\|_2 + \|\theta^*_{2i} - \theta_{2i}\|_2\right) \precsim \epsilon,
\end{aligned}$$

where the second inequality is due to the fact that the expert functions h_1 and h_2 are twice differentiable with respect to their parameters θ_1 and θ_2 and the space $\mathcal{X}$ is a bounded set. This inequality implies that the covering number for metric space $(\mathcal{P}_k(\Omega), \|.\|_\infty)$ will be upper bounded by the cardinality of $\mathcal{S}$. More precisely, we obtain the following bound

$$N\left(\epsilon, \mathcal{P}_k(\Omega), \|.\|_\infty\right) \leq c_1(q_1, k, \Omega_1) c_2(q_2, k, \Omega_2) \left(\frac{5}{\epsilon}\right)^k \left(\frac{1}{\epsilon}\right)^{(q_1+q_2)k}.$$

Putting the above results together, we reach to the conclusion of the bound with covering number (52).

Proof for bracketing entropy control (53) Recall that, from the assumption with expert functions h_1 and h_2, we have $h_1(X,\theta_1) \in [-a, a]$ and $h_2(X,\theta_2) \in [\underline{\gamma}, \overline{\gamma}]$ for all $X \in \mathcal{X}$, $\theta_1 \in \Omega_1$, and $\theta_2 \in \Omega_2$ where a is some positive constant depending only on $\mathcal{X}$ and Ω_1.

Now, let $\eta \leq \epsilon$ to be some positive number that we will chose later. From the formulation of univariate location-scale Gaussian distribution, we can check that

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

for any $|Y| \geq 2a$ and $X \in \mathcal{X}$. Therefore, if we define

$$H(X,Y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\underline{\gamma}} \exp\left(-Y^2/(8\overline{\gamma}^2)\right) \overline{f}(X), & \text{for } |Y| \geq 2a \\ \dfrac{1}{\sqrt{2\pi}\underline{\gamma}} \overline{f}(X), & \text{for } |Y| < 2a, \end{cases} \quad (56)$$

then we can verify that that $H(X,Y)$ is an envelope of $\mathcal{P}_k(\Omega)$. We denote $g_1,\ldots,g_N$ an η-net over $\mathcal{P}_k(\Omega)$. Then, we construct the brackets $[p_i^L(X,Y), p_i^U(X,Y)]$ as follows:

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

for $1 \le i \le N$. We can verify that $\mathcal{P}_k(\Omega) \subset \cup_{i=1}^N [p_i^L(X,Y), p_i^U(X,Y)]$ and $p_i^U(X,Y) - p_i^L(X,Y) \le \min\{2\eta, H(X,Y)\}$. Direct computations lead to

$$\begin{aligned}\int \left(p_i^U(X,Y) - p_i^L(X,Y)\right) d(X,Y) \le & \int\limits_{|Y|<2a} \left(p_i^U(X,Y) - p_i^L(X,Y)\right) d(X,Y) \\ & + \int\limits_{|Y|\ge 2a} \left(p_i^U(X,Y) - p_i^L(X,Y)\right) d(X,Y) \\ & \le \overline{C}\eta + \exp\left(-\overline{C}^2/(2\overline{\gamma}^2)\right) \le c\eta,\end{aligned}$$

where $\overline{C} = \max\{2a, \sqrt{8\overline{\gamma}}\}\log(1/\eta)$ and c is some positive universal constant. The above bound leads to

$$H_B(c\eta, \mathcal{P}_k(\Omega), \|.\|_1) \le N \precsim \log(1/\eta).$$

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

$$H_B(\epsilon, \mathcal{P}_k(\Omega), \|.\|_1) \precsim \log(1/\epsilon).$$

Due to the inequality $h^2 \le \|.\|_1$ between Hellinger distance and total variational distance, we reach the conclusion of bracketing entropy bound (53).