Is Temperature Sample Efficient for Softmax Gaussian Mixture of Experts?

Huy Nguyen [1] Pedram Akbarian [2] Nhat Ho [1]

Abstract

Dense-to-sparse gating mixture of experts (MoE) has recently become an effective alternative to a well-known sparse MoE. Rather than fixing the number of activated experts as in the latter model, which could limit the investigation of potential experts, the former model utilizes the temperature to control the softmax weight distribution and the sparsity of the MoE during training in order to stabilize the expert specialization. Nevertheless, while there are previous attempts to theoretically comprehend the sparse MoE, a comprehensive analysis of the dense-to-sparse gating MoE has remained elusive. Therefore, we aim to explore the impacts of the dense-to-sparse gate on the maximum likelihood estimation under the Gaussian MoE in this paper. We demonstrate that due to interactions between the temperature and other model parameters via some partial differential equations, the convergence rates of parameter estimations are slower than any polynomial rates, and could be as slow as $\mathcal{O}(1/\log(n))$, where n denotes the sample size. To address this issue, we propose using a novel activation dense-to-sparse gate, which routes the output of a linear layer to an activation function before delivering them to the softmax function. By imposing linearly independence conditions on the activation function and its derivatives, we show that the parameter estimation rates are significantly improved to polynomial rates. Finally, we conduct a simulation study to empirically validate our theoretical results.

1. Introduction

Mixture of experts (MoE) (Jacobs et al., 1991; Jordan & Jacobs, 1994) is a statistical machine learning framework that aggregates the power of multiple expert networks using softmax as a gating function (weight function) to create a more sophisticated model than a single network. To scale up the model capacity (the number of model parameters) given a fixed computational cost, (Shazeer et al., 2017) have introduced a sparse variant of the MoE model, which turns on only one or a few experts for each input. Thanks to its scalability, sparse MoE models have been widely used in several applications, namely large language models (Lepikhin et al., 2021; Du et al., 2022; Fedus et al., 2022; Zhou et al., 2023; Jiang et al., 2024), computer vision (Dosovitskiy et al., 2021; Liang et al., 2022; Riquelme et al., 2021), multi-task learning (Hazimeh et al., 2021) and speech recognition (Gulati et al., 2020; You et al., 2021).

In sparse MoE models, gating functions are concurrently trained with expert networks to route inputs effectively. However, early in training, gating networks may exhibit instability in expert selection due to their inexperience. Additionally, pre-determining the number of activated experts per input can limit exploration of potential experts. To address this, (Nie et al., 2022) introduced a dense-to-sparse gate, initially routing inputs to all experts and gradually becoming sparser. This approach strategically controls the temperature of a softmax-based gating function to adjust weight distribution among experts and regulate sparsity in MoE models, promoting stability in expert specialization.

From a theoretical perspective, there have been attempts to comprehend the properties of MoE models. Firstly, (Chen et al., 2022) studied how sparse MoE layers enhanced the efficacy of neural network learning and explained why they would not collapse into a single model. Another line of work tried to understand the effects of gating functions on the convergence rates of maximum likelihood estimation under Gaussian MoE models. In particular, when the gating function was independent of input, (Ho et al., 2022) showed an interaction among expert parameters, which made those rates inversely proportional to the number of over-specified experts. Next, (Nguyen et al., 2023) considered a dense softmax gating function, and demonstrated that parameter estimation rates were determined by the solvability of an intricate system of polynomial equations due to another interaction between gating and expert parameters. Subsequently, (Nguyen et al., 2024b) explored a general Top-K sparse softmax gating function. They revealed that acti-

[1] Department of Statistics and Data Sciences [2] Department of Electrical and Computer Engineering, The University of Texas at Austin. Correspondence to: Huy Nguyen <huynm@utexas.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

vating only one expert, i.e., $K = 1$, makes the previous interaction between expert and gating parameters disappear, and thus, improved the parameter estimation rates significantly. However, a comprehensive theoretical analysis of the dense-to-sparse gate has remained missing in the literature.

In this work, we focus on investigating whether the temperature in the dense-to-sparse gate is sample efficient or not under the parameter estimation problem of the Gaussian MoE model. For that purpose, we now present the formulation of that model formally.

Problem setup. Suppose that the data $\{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ are i.i.d sampled from the dense-to-sparse gating Gaussian mixture of experts, which is associated with the conditional density function $g_{G_*}(Y|X)$ defined as:

$$\sum_{i=1}^{k_*} \mathrm{Softmax}\Big(\frac{(\beta_{1i}^*)^\top X + \beta_{0i}^*}{\tau^*}\Big) \cdot f(Y|(a_i^*)^\top X + b_i^*, \nu_i^*), \tag{1}$$

where $G_* := \sum_{i=1}^{k_*} \exp(\beta_{0i}^*/\tau^*)\delta_{(\beta_{1i}^*,\tau^*,a_i^*,b_i^*,\nu_i^*)}$ is a true but unknown *mixing measure* (i.e., a weighted sum of Dirac measures δ) associated with true parameters $(\beta_{0i}^*, \beta_{1i}^*, \tau^*, a_i^*, b_i^*, \nu_i^*)$ for $i \in \{1, 2, \dots, k_*\}$. Here, τ^* is the softmax temperature which adjusts the sparsity of the MoE models. When τ^* increases, the weight distribution becomes more uniform. On the other hand, when τ^* approaches zero, that distribution turns into one-hot. Meanwhile, $f(\cdot|\mu, \nu)$ stands for a univariate Gaussian density function with mean μ and variance ν. For ease of presentation, we consider k_* linear experts of the form $a^\top X + b$ as the results for general expert settings, including deep neural network, can be achieved in a similar fashion. Additionally, we define for any vector $v = (v_1, \dots, v_{k_*}) \in \mathbb{R}^{k_*}$ that $\mathrm{Softmax}(v_i) := \exp(v_i)/\sum_{j=1}^{k_*}\exp(v_j)$.

Maximum likelihood estimation. To estimate the parameters of model (1), we propose using the maximum likelihood method as follows:

$$\widehat{G}_n := \arg\max_{G} \frac{1}{n}\sum_{i=1}^{n} \log(g_G(Y_i|X_i)). \tag{2}$$

When the true number of expert k_* is known (*exact-specified settings*), the above maximum is taken over the set of all mixing measures of order k_* denoted by $\mathcal{E}_{k_*}(\Theta) := \{G = \sum_{i=1}^{k_*}\exp(\beta_{0i}/\tau)\delta_{(\beta_{1i},\tau,a_i,b_i,\nu_i)} : (\beta_{0i}, \beta_{1i}, \tau, a_i, b_i, \nu_i) \in \Theta\}$. Conversely, when k_* is unknown and the true model (1) is over-specified by a Gaussian mixture of k experts where $k > k_*$ (*over-specified settings*), the maximum is subject to the set of all mixing measures of order at most k, i.e., $\mathcal{G}_k(\Theta) := \{G = \sum_{i=1}^{k'}\exp(\beta_{0i}/\tau)\delta_{(\beta_{1i},\tau,a_i,b_i,\nu_i)} : 1 \leq k' \leq k,\ (\beta_{0i}, \beta_{1i}, \tau, a_i, b_i, \nu_i) \in \Theta\}$.

Assumptions. In our analysis, we have four main assumptions on the parameters:

(A.1) The parameter space Θ is a compact subset of $\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}_+ \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+$, and the input space $\mathcal{X}$ is bounded;

(A.2) $\beta_{1k_*}^* = \mathbf{0}_d$ and $\beta_{0k_*}^* = 0$;

(A.3) $(a_1^*, b_1^*, \nu_1^*), \dots, (a_{k_*}^*, b_{k_*}^*, \nu_{k_*}^*)$ are pairwise distinct;

(A.4) At least one among $\beta_{11}^*, \dots, \beta_{1k_*}^*$ is non-zero.

Above, the first assumption helps ensure the convergence of parameter estimation, while the second guarantees that the dense-to-sparse gating Gaussian MoE model is identifiable (see Appendix C). Next, the third one is to make experts in model (1) pairwise distinct. Finally, the last assumption makes sure that the gating function hinges on the input X.

Technical challenges. The softmax temperature leads to two fundamental challenges in theory:

(C.1) Temperature's interaction with other parameters. To establish a parameter estimation rate given a density estimation rate, we need to decompose the density discrepancy $g_{\widehat{G}_n}(Y|X) - g_{G_*}(Y|X)$ into a combination of linearly independent terms. This can be done by applying Taylor expansions to the softmax's numerator $F(Y|X, \omega) := \exp(\frac{\beta_1^\top X}{\tau})f(Y|a^\top X + b, \nu)$, where $\omega := (\beta_1, \tau, a, b, \nu)$. However, we realize that the temperature interacts with both gating and expert parameters via two following partial differential equations (PDEs), which induce a number of linearly dependent terms:

$$\frac{\partial F}{\partial \tau} = \frac{1}{\tau} \cdot \beta_1^\top \frac{\partial F}{\partial \beta_1}, \tag{3}$$

$$\frac{\partial^2 F}{\partial \tau\, \partial b} = \frac{1}{\tau^2} \cdot \beta_1^\top \frac{\partial F}{\partial a}. \tag{4}$$

Intuitively, the first PDE reveals that there is an intrinsic interaction between the temperature τ and the gating parameter β_1. Meanwhile, the second PDE indicates that the temperature also interacts with the expert parameters a and b. Although parameter interactions expressed in the language of PDEs have been observed in (Nguyen et al., 2023), the structures of the above interactions are Furthermore, these interactions are substantially more serious than those in (Ho et al., 2022; Nguyen et al., 2023; 2024b). More specifically, we will show in Section 2 that due to the above PDEs, the parameter estimation rates are slower than any polynomial rates, and thus, could be as slow as $1/\log(n)$, where n denotes the sample size. Such phenomenon has never been observed in previous work.

(C.2) Rate improvement. From the previous observation, it is essential to propose a method to accelerate the parameter estimation rates. To enhance slow rates caused by the interaction between gating and expert parameters, (Nguyen et al., 2024a) suggested transforming the inputs using a 'modified' function M, e.g. $\log(|\cdot|)$, $\cos(\cdot)$, prior to delivering them to the gating network, i.e. $\mathrm{Softmax}\Big(\frac{(\beta_{1i}^*)^\top M(X) + \beta_{0i}^*}{\tau^*}\Big)$.

However, it can be verified that the PDEs (3) and (4) still holds true with the corresponding function $F(Y|X,\omega) := \exp(\frac{\beta_1^\top M(X)}{\tau})f(Y|a^\top X + b,\nu)$. Therefore, we have to come up with a novel solution in this work.

Main contributions. In this paper, we conduct a convergence analysis of density and parameter estimations under the dense-to-sparse gating Gaussian MoE. Our contributions are two-fold and can be summarized as follows:

1. Dense-to-sparse gating function: Equipped with this gating function, we first establish the convergence rate of density estimation under the Total Variation distance $\mathbb{E}_X[V(g_{\widehat{G}_n}(\cdot|X), g_{G_*}(\cdot|X))] = \widetilde{\mathcal{O}}(n^{-1/2})$, which is parametric on the sample size n. Given this result, we then demonstrate that under the exact-specified settings, the estimation rates for β^*_{1i},τ^* are slower than any polynomial rates owing to the PDE (3), and therefore, could be $1/\log(n)$. Meanwhile, those for a^*_i, b^*_i, ν^*_i are significantly faster, standing at $\widetilde{\mathcal{O}}(n^{-1/2})$. Under the over-specified settings, we show that the rates for estimating β^*_{1i},τ^* remain unchanged, whereas that for a^*_i becomes slower than any polynomial rates due to the PDE (4). Additionally, the estimation rates for b^*_i,ν^*_i depend on the solvability of a system of polynomial equations.

2. Activation dense-to-sparse gating function. To enhance the previous slow rates, we propose a novel class of gating functions called *activation dense-to-sparse* given by $\mathrm{Softmax}\Big(\frac{\sigma((\beta^*_{1i})^\top X)+\beta^*_{0i}}{\tau^*}\Big)$. Here, $\sigma(\cdot)$ is an activation function satisfying conditions in Definition 3.2 (resp. Definition 3.4), which make the interactions of temperature with other parameters in equations (3) and (4) vanish under the exact-specified (resp. over-specified) settings. As a consequence, we rigorously prove that β^*_{1i}, τ^* and a^*_i share the same considerably improved estimation rates of orders $\widetilde{\mathcal{O}}(n^{-1/2})$ and $\widetilde{\mathcal{O}}(n^{-1/4})$ under those settings, respectively.

Outline. The paper proceeds as follows. In Section 2, we derive the convergence rates of density estimation and parameter estimation under the dense-to-sparse gating Gaussian MoE. Subsequently, we carry out the previous analysis for the Gaussian MoE with the novel activation dense-to-sparse gate in Section 3. Then, we run some numerical experiments in Section 4 to empirically verify our theoretical results before concluding the paper in Section 6. Finally, rigorous proofs and further details of experiments are provided in the supplementary material.

Notations. We denote $[n] := \{1,2,\ldots,n\}$ for any positive integer n. Additionally, the notation $|S|$ indicates the cardinality of any set S. For any vectors $v := (v_1,v_2,\ldots,v_d) \in \mathbb{R}^d$ and $\alpha := (\alpha_1,\alpha_2,\ldots,\alpha_d) \in \mathbb{N}^d$, we let $v^\alpha = v_1^{\alpha_1}v_2^{\alpha_2}\ldots v_d^{\alpha_d}$, $|v| := v_1 + v_2 + \ldots + v_d$ and $\alpha! := \alpha_1!\alpha_2!\ldots\alpha_d!$, while $\|v\|$ stands for its 2-norm value. Given any two positive sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, we write $a_n = \mathcal{O}(b_n)$ or $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for all $n \in \mathbb{N}$, where $C > 0$ is some universal constant. Furthermore, we write $a_n = \widetilde{\mathcal{O}}(b_n)$ to indicate $a_n \widetilde{\lesssim} b_n$ up to some logarithmic factors. Finally, for any two probability density functions p,q dominated by the Lebesgue measure μ, we denote $h(p,q) = \left(\frac{1}{2}\int(\sqrt{p}-\sqrt{q})^2 d\mu\right)^{1/2}$ as their Hellinger distance and $V(p,q) = \frac{1}{2}\int|p-q|d\mu$ as their Total Variation distance.

2. Dense-to-sparse Gating Function

In this section, we characterize the density and parameter estimation rates for the dense-to-sparse gating Gaussian MoE under both the exact-specified and over-specified settings.

We start with providing the convergence rate of the density estimation $g_{\widehat{G}_n}$ to the true density g_{G_*} under the Total Variation distance in the following theorem:

Theorem 2.1. *Under the Total Variation distance, the density estimation $g_{\widehat{G}_n}(Y|X)$ converges to the true density $g_{G_*}(Y|X)$ at the following rate:*

$$\mathbb{E}_X[V(g_{\widehat{G}_n}(\cdot|X), g_{G_*}(\cdot|X))] = \widetilde{\mathcal{O}}(n^{-1/2}).$$

We leverage fundamental results on density estimation for M-estimator in (van de Geer, 2000) to prove Theorem 2.1 in Appendix A.1. It follows from the above bound that the density estimation rate is parametric on the sample size n. This results also indicates that if the Total Variation lower bound $\mathbb{E}_X[V(g_{\widehat{G}_n}(\cdot|X), g_{G_*}(\cdot|X))] \gtrsim \mathcal{D}(\widehat{G}_n, G_*)$, where $\mathcal{D}$ is some loss function among parameters, then we obtain the parameter estimation rate $\mathcal{D}(\widehat{G}_n, G_*) = \widetilde{\mathcal{O}}(n^{-1/2})$. Now, we are ready to precisely capture those rates under the exact-specified and over-specified settings in Section 2.1 and Section 2.2, respectively.

2.1. Exact-specified Settings

Before diving deeper into the parameter estimation problem under the exact-specified settings, let us introduce a notion of Voronoi cells (Manole & Ho, 2022), which are then used to construct our loss functions.

Voronoi cells. Assume that a mixing measure G has k' components $\omega_i := (\beta_{1i},\tau,a_i,b_i,\nu_i)$. Then, we distribute these components to the Voronoi cells $\mathcal{A}_j \equiv \mathcal{A}_j(G)$ generated by the components $\omega^*_j := (\beta^*_{1j},\tau^*,a^*_j,b^*_j,\nu^*_j)$ of G_*, which are defined as

$$\mathcal{A}_j := \{i \in [k'] : \|\omega_i - \omega^*_j\| \leq \|\omega_i - \omega^*_\ell\|, \forall \ell \neq j\}. \quad (5)$$

For instance, since the MLE $\widehat{G}_n$ has k_* components under this setting, each Voronoi cell $\mathcal{A}_j(\widehat{G}_n)$ has exactly one element when the sample size n is sufficiently large.

Voronoi loss. Let us define $K_{ij}(\kappa_1,\kappa_2,\kappa_3,\kappa_4,\kappa_5) := \|\Delta\beta_{1ij}\|^{\kappa_1} + |\Delta\tau|^{\kappa_2} + \|\Delta a_{ij}\|^{\kappa_3} + |\Delta b_{ij}|^{\kappa_4} + |\Delta\nu_{ij}|^{\kappa_5}$,

Table 1. Summary of density estimation rates and parameter estimation rates under the (activation) dense-to-sparse gating Gaussian MoE. In this table, the function $\bar{r}(\cdot)$ represents for the solvability of the system of polynomial equations (11) with $\bar{r}(2) = 4$ and $\bar{r}(3) = 6$. Additionally, $\mathcal{A}_j$ denotes a Voronoi cell given in equation (5).

Dense-to-sparse gating Gaussian MoE					
Setting	$g_{G_*}(Y\|X)$	β^*_{1j}, τ^*	a^*_j	b^*_j	ν^*_j
Exact-specified	$\widetilde{\mathcal{O}}(n^{-1/2})$	Slower than $\widetilde{\mathcal{O}}(n^{-1/2r}), \forall r \geq 1$	$\widetilde{\mathcal{O}}(n^{-1/2})$	$\widetilde{\mathcal{O}}(n^{-1/2})$	$\widetilde{\mathcal{O}}(n^{-1/2})$
Over-specified	$\widetilde{\mathcal{O}}(n^{-1/2})$	Slower than $\widetilde{\mathcal{O}}(n^{-1/2r}), \forall r \geq 1$		$\widetilde{\mathcal{O}}(n^{-1/2\bar{r}(\|\mathcal{A}_j\|)})$	$\widetilde{\mathcal{O}}(n^{-1/\bar{r}(\|\mathcal{A}_j\|)})$
Activation Dense-to-sparse gating Gaussian MoE					
Setting	$p_{G_*}(Y\|X)$	β^*_{1j}, τ^*	a^*_j	b^*_j	ν^*_j
Exact-specified	$\widetilde{\mathcal{O}}(n^{-1/2})$	$\widetilde{\mathcal{O}}(n^{-1/2})$			
Over-specified	$\widetilde{\mathcal{O}}(n^{-1/2})$	$\widetilde{\mathcal{O}}(n^{-1/4})$		$\widetilde{\mathcal{O}}(n^{-1/2\bar{r}(\|\mathcal{A}_j\|)})$	$\widetilde{\mathcal{O}}(n^{-1/\bar{r}(\|\mathcal{A}_j\|)})$

where $\Delta\beta_{1ij} := \beta_{1i} - \beta_{1j}$, $\Delta\tau := \tau - \tau^*$, $\Delta a_{ij} := a_i - a^*_j$, $\Delta b_{ij} := b_i - b^*_j$ and $\Delta\nu_{ij} := \nu_i - \nu^*_j$. Then, the Voronoi loss of interest is given by

$$\mathcal{D}_{1,r}(G, G_*) := \sum_{j=1}^{k_*} \Big| \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_{0i}}{\tau}\Big) - \exp\Big(\frac{\beta^*_{0j}}{\tau^*}\Big) \Big| + \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_{0i}}{\tau}\Big) K_{ij}(r, r, r, r, r) \quad (6)$$

Next, let us recall that when using the dense-to-sparse gate, there are two interactions of the softmax temperature τ with gating parameter β_1 and expert parameters a, b:

$$\frac{\partial F}{\partial \tau} = \frac{1}{\tau} \cdot \beta_1^\top \frac{\partial F}{\partial \beta_1}; \quad \frac{\partial^2 F}{\partial \tau\, \partial b} = \frac{1}{\tau^2} \cdot \beta_1^\top \frac{\partial F}{\partial a}, \quad (7)$$

where $F(Y|X, \omega) := \exp(\frac{\beta_1^\top X}{\tau}) f(Y|a^\top X + b, \nu)$. Unfortunately, such interactions are so serious that the Total Variation lower bound $\mathbb{E}_X[V(g_{\widehat{G}_n}(\cdot|X), g_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_{1,r}(\widehat{G}_n, G_*)$ does not hold true, and thus, we cannot achieve the bound $\mathcal{D}_{1,r}(\widehat{G}_n, G_*) = \widetilde{\mathcal{O}}(n^{-1/2})$ as discussed below Theorem 2.1. Instead, we show in Appendix B.1 that

$$\inf_{G \in \mathcal{E}_{k_*}(\Theta): \mathcal{D}_{1,r}(G, G_*) \leq \varepsilon} \frac{\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_{1,r}(G, G_*)} \to 0,$$

as $\varepsilon \to 0$, for any $r \geq 1$. This result leads to the following minimax lower bound of parameter estimation:

Theorem 2.2. *Under the exact-specified settings, the following minimax lower bound of estimating G_* holds true for any $r \geq 1$:*

$$\inf_{\overline{G}_n \in \mathcal{E}_{k_*}(\Theta)} \sup_{G \in \mathcal{E}_{k_*}(\Theta)} \mathbb{E}_{g_G}[\mathcal{D}_{1,r}(\overline{G}_n, G)] \gtrsim n^{-1/2}.$$

Here, the notation $\mathbb{E}_{g_G}$ indicates the expectation taken w.r.t the product measure with mixture density g^n_G.

Proof of Theorem 2.2 is in Appendix B.1. The above minimax lower bound suggests that the rates for estimating parameters $\beta^*_{1j}, \tau^*, a^*_j, b^*_j, \nu^*_j$ are slower than any polynomial rates $\widetilde{\mathcal{O}}(n^{-1/2r})$, and therefore, could be as slow as $1/\log(n)$. This convergence behavior has never been captured in previous work on Gaussian MoE models, including (Ho et al., 2022; Nguyen et al., 2023; 2024c;b). Nevertheless, in our arguments, since the true number of experts k_* is known, it is sufficient to apply the first-order Taylor expansion to the gating numerator F. Therefore, the second PDE in equation (7) should not affect the parameter estimation rates under this setting. In other words, parameters a^*_i, b^*_i, ν^*_i should enjoy faster estimation rates than their counterparts β^*_{1i}, τ^*. To illustrate this point, let us take into account another Voronoi loss function.

Voronoi loss. To capture the rates for estimating a^*_j, b^*_j and ν^*_j more accurately, it is essential to consider the projections of previous mixing measures onto the space of those parameters $\Psi := \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+$. In particular, for each $G = \sum_{i=1}^{k} \exp(\beta_{0i}/\tau)\delta_{(\beta_{1i}, \tau, a, b, \nu)}$, we define $G^{|\Psi} := \sum_{i=1}^{k} \exp(\beta_{0i}/\tau)\delta_{(a_i, b_i, \nu_i)}$. Then, the loss function between these projected mixing measures is given by:

$$\mathcal{D}_2(G^{|\Psi}, G^{|\Psi}_*) := \sum_{j=1}^{k_*} \Big| \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_{0i}}{\tau}\Big) - \exp\Big(\frac{\beta^*_{0j}}{\tau^*}\Big) \Big| + \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_{0i}}{\tau}\Big) \Big[\|\Delta a_{ij}\| + |\Delta b_{ij}| + |\Delta \nu_{ij}| \Big]. \quad (8)$$

Theorem 2.3. *Under the exact-specified settings, the following Total Variation lower bound holds true for any $G \in \mathcal{E}_{k_*}(\Theta)$:*

$$\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_2(G^{|\Psi}, G^{|\Psi}_*).$$

This bound together with Theorem 2.1 leads to the parametric convergence rate of MLE: $\mathcal{D}_2(\widehat{G}_n^{|\Psi}, G_*^{|\Psi}) = \widetilde{\mathcal{O}}(n^{-1/2})$.

Proof of Theorem 2.3 is in Appendix B.2. It follows from the above result that a_j^*, b_j^*, ν_j^* share the same estimation rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$, which are significantly faster than those resulting from Theorem 2.2.

2.2. Over-specified Settings

Analogous to the previous section, we first need to design a Voronoi loss function used for the over-specified settings.

Voronoi loss. Let us define for each $r \geq 1$ that

$$\mathcal{D}_{3,r}(G, G_*) := \sum_{j=1}^{k_*} \Big| \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_{0i}}{\tau}\Big) - \exp\Big(\frac{\beta_{0j}^*}{\tau^*}\Big) \Big| + \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_{0i}}{\tau}\Big) K_{ij}(r, r, r, r, r). \quad (9)$$

Recall that under this setting, the true number of experts k_* is unknown, and we assume that the MLE $\widehat{G}_n$ belongs to the set of mixing measures with at most $k > k_*$ components $\mathcal{G}_k(\Theta)$. Thus, from the definition of Voronoi cells, there could be some cells $\mathcal{A}_j(\widehat{G}_n)$ having more than one element. On the other hand, each Voronoi cell $\mathcal{A}_j(\widehat{G}_n)$ under the exact-specified settings has exactly one element. This is the main difference between the Voronoi losses $\mathcal{D}_{3,r}$ and $\mathcal{D}_{1,r}$.

Theorem 2.4. *Under the over-specified settings, the following minimax lower bound of estimating G_* holds true for any $r \geq 1$:*

$$\inf_{\overline{G}_n \in \mathcal{G}_k(\Theta)} \sup_{G \in \mathcal{G}_k(\Theta) \setminus \mathcal{O}_{k_*-1}(\Theta)} \mathbb{E}_{g_G}[\mathcal{D}_{3,r}(\overline{G}_n, G)] \gtrsim n^{-1/2}.$$

Here, the notation $\mathbb{E}_{g_G}$ indicates the expectation taken w.r.t the product measure with mixture density g_G^n.

Proof of Theorem 2.4 is in Appendix B.3. The above minimax lower bound indicates that the estimation rates for parameters $\beta_{1j}^*, \tau^*, a_j^*, b_j^*, \nu_j^*$ are all slower than $\widetilde{\mathcal{O}}(n^{-1/2r})$ for any $r \geq 1$. This means that those rates cannot be faster than polynomial rates and could be as slow as $1/\log(n)$. Such slow rates are caused by the interaction between the softmax temperature and other parameters via the PDEs in equation (7). Despite the issue, not all parameters are negatively impacted. Specifically, our constructed loss function, detailed in Theorem 2.5, demonstrates polynomial estimation rates for b_j^* and ν_j^*.

Voronoi loss. Similar to Section 2.1, for each mixing measure $G = \sum_{i=1}^{k} \exp(\beta_{0i}/\tau)\delta_{(\beta_{1i}, \tau, a, b, \nu)}$, we consider its projection on the space $\Upsilon := \mathbb{R} \times \mathbb{R}_+$ of parameters b_j^*, ν_j^*, that is, $G^{|\Upsilon} := \sum_{i=1}^{k} \exp(\beta_{0i}/\tau)\delta_{(b_i, \nu_i)}$. Then, the loss function of interest is defined as

$$\mathcal{D}_4(G^{|\Upsilon}, G_*^{|\Upsilon}) := \sum_{j=1}^{k_*} \Big| \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_{0i}}{\tau}\Big) - \exp\Big(\frac{\beta_{0j}^*}{\tau^*}\Big) \Big| + \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_{0i}}{\tau}\Big) \Big[|\Delta b_{ij}|^{\bar{r}(|\mathcal{A}_j|)} + |\Delta \nu_{ij}|^{\frac{\bar{r}(|\mathcal{A}_j|)}{2}} \Big] + \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_{0i}}{\tau}\Big) \Big[|\Delta b_{ij}| + |\Delta \nu_{ij}| \Big]. \quad (10)$$

Here, $\bar{r}(|\mathcal{A}_j|)$ stands for the smallest positive integer r such that the following system does not have any non-trivial solutions for the unknown variables $\{p_l, q_{1l}, q_{2l}\}_{l=1}^{m}.$:

$$\sum_{l=1}^{|\mathcal{A}_j|} \sum_{\substack{n_1, n_2 \in \mathbb{N}: \\ n_1 + 2n_2 = s}} \frac{p_l^2 \, q_{1l}^{n_1} \, q_{2l}^{n_2}}{n_1! \; n_2!} = 0, \quad s = 1, 2, \ldots, r, \quad (11)$$

A solution is called non-trivial if all the values of p_l are different from zero, whereas at least one among q_{1l} is nonzero. (Ho & Nguyen, 2016) demonstrate that $\bar{r}(2) = 4$, $\bar{r}(3) = 6$ and $\bar{r}(m) \geq 7$ when $m \geq 4$.

Theorem 2.5. *Under the over-specified settings, the following Total Variation lower bound holds true for any $G \in \mathcal{G}_k(\Theta)$:*

$$\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_4(G^{|\Upsilon}, G_*^{|\Upsilon}).$$

This bound together with Theorem 2.1 leads to the parametric convergence rate of MLE: $\mathcal{D}_4(\widehat{G}_n^{|\Upsilon}, G_*^{|\Upsilon}) = \widetilde{\mathcal{O}}(n^{-1/2})$.

Proof of Theorem 2.5 is in Appendix B.4. The above result reveals that the MLE $\widehat{G}_n$ converges to the true mixing measure G_* under the loss $\mathcal{D}_4$ at the parametric rate $\widetilde{\mathcal{O}}(n^{-1/2})$, which implies the followings:

(i) The rates for estimating parameters β_{1j}^*, ν_j^* which are fitted by one component, i.e. $|\mathcal{A}_j(\widehat{G}_n)| = 1$, are of the same order $\widetilde{\mathcal{O}}(n^{-1/2})$. Compared to the rates resulted from Theorem 2.3 under the exact-specified settings, those rates remain unchanged under the over-specified settings.

(ii) For parameters β_{1j}^*, ν_j^* which are approximated by more than one component, i.e. $|\mathcal{A}_j(\widehat{G}_n)| > 1$, their estimation rates are of orders $\widetilde{\mathcal{O}}(n^{-1/2\bar{r}(|\mathcal{A}_j(\widehat{G}_n)|)})$ and $\widetilde{\mathcal{O}}(n^{-1/\bar{r}(|\mathcal{A}_j(\widehat{G}_n)|)})$, respectively. For instance, if those parameters are fitted by two components, that is, $|\mathcal{A}_j(\widehat{G}_n)| = 2$, then the previous rates become $\widetilde{\mathcal{O}}(n^{-1/8})$ and $\widetilde{\mathcal{O}}(n^{-1/4})$. On the other hand, if $|\mathcal{A}_j(\widehat{G}_n)| = 3$, then the rates for estimating β_{1j}^*, ν_j^* are of orders $\widetilde{\mathcal{O}}(n^{-1/12})$ and $\widetilde{\mathcal{O}}(n^{-1/6})$.

3. Activation Dense-to-sparse Gating Function

In this section, we propose a novel class of gating functions named activation dense-to-sparse in order to improve the slow parameter estimation rates when using the dense-to-sparse gate in Section 2.

To begin with, let us present the formulation of a Gaussian MoE with the activation dense-to-sparse gating function.

Problem setup. Suppose that the data $\{(X_i, Y_i)\}_{i=1}^{n} \subset \mathbb{R}^d \times \mathbb{R}$ are i.i.d sampled from the activation dense-to-sparse Gaussian MoE, whose conditional density function $p_{G_*}(Y|X)$ is defined as:

$$\sum_{i=1}^{k_*} \mathrm{Softmax}\Big(\frac{\sigma((\beta_{1i}^*)^\top X) + \beta_{0i}^*}{\tau^*}\Big) \times f(Y|(a_i^*)^\top X + b_i^*, \nu_i^*). \quad (12)$$

In the model's gating network, the output of a linear layer undergoes an activation function σ before entering the softmax function. The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ must satisfy the conditions in Definition 3.2 and Definition 3.4 for exact-specified and over-specified settings, respectively. These conditions mitigate the interaction of the softmax temperature with other parameters in equation (7), addressing the slow rates discussed in Section 2. We maintain the assumptions on the parameters outlined in Section 1, unless explicitly stated otherwise.

Maximum likelihood estimation. According to the change of gating function, let us re-define the MLE corresponding to the model (12) as follows:

$$\widetilde{G}_n := \arg\max_{G} \frac{1}{n}\sum_{i=1}^{n} \log(p_G(Y_i|X_i)). \quad (13)$$

Subsequently, we provide in the following theorem a convergence rate of density estimation under the Gaussian MoE model with the activation dense-to-sparse gate.

Theorem 3.1. *Under the Total Variation distance, the density estimation $p_{\widehat{G}_n}(Y|X)$ converges to the true density $p_{G_*}(Y|X)$ at the following rate:*

$$\mathbb{E}_X[V(p_{\widetilde{G}_n}(\cdot|X), p_{G_*}(\cdot|X))] = \widetilde{\mathcal{O}}(n^{-1/2}).$$

Proof of Theorem 3.1 is in Appendix A.2. The above bound confirms that the density estimation rate under the Gaussian MoE with the activation dense-to-sparse gate is of the same order as that with the standard dense-to-sparse gate in Theorem 2.1, which is $\widetilde{\mathcal{O}}(n^{-1/2})$. Next, we utilize this result to derive the parameter estimation rates for the model (12) under the exact-specified and over-specified settings.

3.1. Exact-specified Setting

First of all, we introduce the conditions on the activation function σ in the model (12) under this setting.

Definition 3.2 (First-order Independence)**.** Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a differentiable function, and $\bar{\sigma}(X, w) := \sigma(w^\top X)$. We say that σ is first-order independent if the set

$$\Big\{1,\ \bar{\sigma}(X, w),\ \frac{\partial \bar{\sigma}}{\partial w^{(u)}}(X, w) : 1 \leq u \leq d\Big\} \quad (14)$$

is linearly independent w.r.t X for any $w \in \mathbb{R}^d$.

Example. It can be verified that the functions $\mathrm{sigmoid}(\cdot)$ and Gaussian error linear units $\mathrm{GELU}(\cdot)$ (Hendrycks & Gimpel, 2023) are first-order independent. On the other hand, the function $z \mapsto z^p$ for $p \geq 1$ does not satisfy the first-order independence condition in Definition 3.2.

Denote $\widetilde{F}(Y|X, \omega) := \exp(\frac{\sigma(\beta_1^\top X)}{\tau}) f(Y|a^\top X + b, \nu)$. Then, the first-order independence condition on the activation function σ guarantees that the interaction between τ and β_1 in equation (7) no longer holds true, that is,

$$\frac{\partial \widetilde{F}}{\partial \tau} \neq \frac{1}{\tau} \cdot \beta_1^\top \frac{\partial \widetilde{F}}{\partial \beta_1}.$$

As a result, the estimation rates for parameters β_{1j}^* and τ^* should be improved in comparison with those in Section 2. To certify this point, we design the following Voronoi loss function, and then provide in Theorem 3.3 the convergence rate of the MLE under the exact-specified settings.

Voronoi loss. The Voronoi loss of interest is given by

$$\mathcal{D}_5(G, G_*) := \sum_{j=1}^{k_*} \Big| \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_{0i}}{\tau}\Big) - \exp\Big(\frac{\beta_{0j}^*}{\tau^*}\Big)\Big| + \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_{0i}}{\tau}\Big) K_{ij}(1,1,1,1,1). \quad (15)$$

Theorem 3.3. *Under the exact-specified settings, the following Total Variation lower bound holds true for any $G \in \mathcal{E}_{k_*}(\Theta)$:*

$$\mathbb{E}_X[V(p_G(\cdot|X), p_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_5(G, G_*).$$

This bound together with Theorem 3.1 leads to the parametric convergence rate of MLE: $\mathcal{D}_5(\widetilde{G}_n, G_*) = \widetilde{\mathcal{O}}(n^{-1/2})$.

Proof of Theorem 3.3 is in Appendix B.5. Theorem 3.3 implies that all the rates for estimating parameters $\beta_{1j}^*, \tau^*, a_j^*, b_j^*, \nu_j^*$ are parametric on the sample size, standing at order $\widetilde{\mathcal{O}}(n^{-1/2})$. It can be seen that the estimation rates for β_{1j}^* and τ^* when using the activation dense-to-sparse gate become substantially faster than their counterparts when using the standard dense-to-sparse gate, which are slower than $\widetilde{\mathcal{O}}(n^{-1/2r})$ for any $r \geq 1$. This highlights the benefits of our proposed activation dense-to-sparse gate.

3.2. Over-specified Setting

In this section, we continue to characterize conditions for the activation function σ under the over-specified settings for the sake of enhancing the parameter estimation rates.

Definition 3.4 (Second-order Independence)**.** Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a twice differentiable function and $\bar{\sigma}(X, w) := \sigma(\omega^\top X)$. We say that σ is second-order independent if the set

$$\Big\{1,\ \bar{\sigma}(X,w),\ \bar{\sigma}^2(X,w),\ \frac{\partial \bar{\sigma}}{\partial w^{(u)}}(X,w),\ \Big(\bar{\sigma} \cdot \frac{\partial \bar{\sigma}}{\partial w^{(u)}}\Big)(X,w),\ \Big(\frac{\partial \bar{\sigma}}{\partial w^{(u)}} \cdot \frac{\partial \sigma}{\partial w^{(v)}}\Big)(X,w),\ \frac{\partial^2 \bar{\sigma}}{\partial w^{(u)} \partial w^{(v)}}(X,w) : 1 \leq u, v \leq d\Big\}$$

is linearly independent for almost surely X for any $w \in \mathbb{R}^d$.

Example. We can validate that the function $\mathrm{sigmoid}(\cdot)$ and Gaussian error linear units $\mathrm{GELU}(\cdot)$ (Hendrycks & Gimpel, 2023) also meet the second-order independence condition. Additionally, since the second-order independence condition implies the first-order one, the function $z \mapsto z^p$, for $p \geq 1$, is still not second-order independent.

The second-order independence condition on the activation function σ ensures that there are no interactions of the softmax temperature with other parameters as in equation (7), i.e.

$$\frac{\partial \widetilde{F}}{\partial \tau} \neq \frac{1}{\tau} \cdot \beta_1^\top \frac{\partial \widetilde{F}}{\partial \beta_1}; \quad \frac{\partial^2 \widetilde{F}}{\partial \tau\, \partial b} \neq \frac{1}{\tau^2} \cdot \beta_1^\top \frac{\partial \widetilde{F}}{\partial a}.$$

Consequently, not only the rates for estimating β_{1j}^* and τ^* should be improved under the over-specified settings as in Section 3.1 but also those for parameters a_j^*. Now, it is necessary to build a Voronoi loss function to give a theoretical guarantee for that claim in Theorem 3.5.

Voronoi loss. Then, the Voronoi loss of interest is given by

$$\begin{aligned}\mathcal{D}_6(G, G_*) &:= \sum_{j=1}^{k_*} \Big| \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_{0i}}{\tau}\Big) - \exp\Big(\frac{\beta_{0j}^*}{\tau^*}\Big) \Big| \\ &+ \sum_{j: |\mathcal{A}_j| > 1} \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_{0i}}{\tau}\Big) K_{ij}\Big(2, 2, 2, \bar{r}(|\mathcal{A}_j|), \frac{\bar{r}(|\mathcal{A}_j|)}{2}\Big) \\ &+ \sum_{j: |\mathcal{A}_j| = 1} \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_{0i}}{\tau}\Big) K_{ij}(1, 1, 1, 1, 1). \qquad (16)\end{aligned}$$

Theorem 3.5. *Under the over-specified settings, the following Total Variation lower bound holds true for any $G \in \mathcal{G}_k(\Theta)$:*

$$\mathbb{E}_X[V(p_G(\cdot|X), p_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_6(G, G_*)$$

This bound together with Theorem 3.1 leads to the parametric convergence rate of MLE: $\mathcal{D}_6(\widetilde{G}_n, G_*) = \widetilde{\mathcal{O}}(n^{-1/2})$.

Proof of Theorem 3.5 is in Appendix B.6. The above parametric convergence rate $\widetilde{\mathcal{O}}(n^{-1/2})$ of the MLE $\widetilde{G}$ to G_* under the loss function $\mathcal{D}_6$ gives us the followings:

(i) Under the over-specified settings, parameters $\beta_{1j}^*, \tau^*, a_j^*, b_j^*, \nu_j^*$ which are fitted by one component, i.e. $|\mathcal{A}_j(\widetilde{G}_n)| = 1$, enjoy the same estimation rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$. This result aligns with that under the exact-specified settings in Theorem 3.3.

(ii) On the other hand, those for parameters approximated by more than one component, i.e. $|\mathcal{A}_j(\widetilde{G}_n)| > 1$, are no longer homogeneous. In particular, the rates for estimating $\beta_{1j}^*, \tau^*, a_j^*$ are of order $\widetilde{\mathcal{O}}(n^{-1/4})$. At the same time, the estimation rates for b_j^* and ν_j^* become $\widetilde{\mathcal{O}}(n^{-1/2\bar{r}(|\mathcal{A}_j(\widetilde{G}_n)|)})$ and $\widetilde{\mathcal{O}}(n^{-1/\bar{r}(|\mathcal{A}_j(\widetilde{G}_n)|)})$, respectively.

4. Numerical Experiments

In this section, we perform numerical experiments to empirically confirm the theoretical convergence rates of maximum likelihood estimation (MLE) in both standard and activation dense-to-sparse gating MoE models. We employ an Expectation-Maximization (EM) algorithm (Dempster et al., 1977) for parameter estimation, utilizing synthetic datasets generated from the true models in equation (1) and equation (12). Further details on the experimental setups are deferred to Appendix D due to the space limit.

Throughout the experiments under the over-specified setting, we set true number of experts $k_* = 2$ and the estimated number of experts $k = k_* + 1 = 3$. We perform 40 experiments for each sample size n in different setups, encompassing a range of 20 distinct sample sizes whose values vary from $n = 10^4$ to $n = 10^5$. Moreover, we use the sigmoid function as an activation for the activation dense-to-sparse gate. The graphs in Figure 1 illustrate that the empirical convergence rates of the MLE $\widehat{G}_n$ to the true mixing measure G_* under different Voronoi metrics.

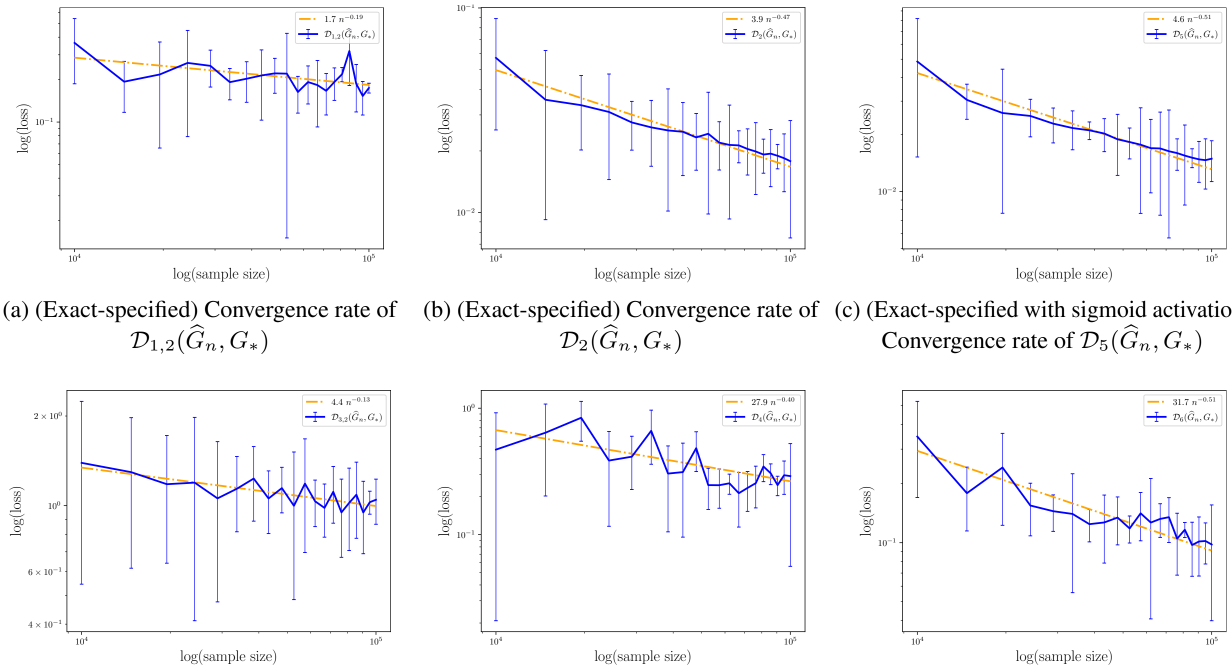
4.1. Standard Dense-to-sparse Gating Function

Exact-specified setting. As illustrated in Figures 1a and 1b, the convergence rate of $\mathcal{D}_{1,2}(\widehat{G}_n, G_*)$ is significantly slow as indicated by Theorem 2.2. Additionally, that of $\mathcal{D}_2(\widehat{G}_n, G_*)$ admits a much faster rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$, which aligns with the result in Theorem 2.3.

Over-specified setting. Similar to the exact-specified setting, as depicted in Figures 1d and 1e, the convergence rate for $\mathcal{D}_{3,2}(\widehat{G}_n, G_*)$ is notably slower than that of $\mathcal{D}_4(\widehat{G}_n, G_*)$, where $\mathcal{D}_4(\widehat{G}_n, G_*)$ exhibits a parametric rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$ per Theorem 2.4 and Theorem 2.5.

4.2. Activation Dense-to-sparse Gating Function

Illustrated in Figures 1c and 1f, the use of sigmoid activation in the softmax gate leads to an enhanced rate of $\widetilde{\mathcal{O}}(n^{-1/2})$ for both exact-specified and over-specified settings. These results totally match those in Theorem 3.3 and Theorem 3.5.



(a) (Exact-specified) Convergence rate of $\mathcal{D}_{1,2}(\widehat{G}_n, G_*)$

(b) (Exact-specified) Convergence rate of $\mathcal{D}_2(\widehat{G}_n, G_*)$

(c) (Exact-specified with sigmoid activation) Convergence rate of $\mathcal{D}_5(\widehat{G}_n, G_*)$

(d) (Over-specified) Convergence rate of $\mathcal{D}_{3,2}(\widehat{G}_n, G_*)$

(e) (Over-specified) Convergence rate of $\mathcal{D}_4(\widehat{G}_n, G_*)$

(f) (Over-specified with sigmoid activation) Convergence rate of $\mathcal{D}_6(\widehat{G}_n, G_*)$

Figure 1. Log-log scaled plots for the empirical convergence rates of the MLE $\widehat{G}_n$ for exact and over-specified settings. In these figures, the corresponding empirical discrepancies are illustrated by the blue curves, while the orange dash-dotted lines represent for the least-squares fitted linear regression lines. The error bars represent two times the empirical standard deviation under the exact-specified setting.

5. Practical Implications

In this section, we provide four main practical implications from our theoretical results:

1. Expert Selection: An important application of the dense-to-sparse MoE is to select the important experts (similar to the popular top-K sparse MoE for large language models in the literature). Our theory for parameter estimation have direct indications for the sample efficiency of choosing the important experts. In particular, it suggests that we need an exponential number of data (roughly $\exp(1/\epsilon^{\eta})$ for some $\eta > 0$ where ϵ is the desired approximation error) to estimate the softmax gating function with temperature, which directly implies that we need an exponential number of data to select the important experts. This is undesirable in practice and would lead to potential wrong choice of important experts when we only have a finite number of data. On the other hand, for the proposed activation dense-to-sparse gate, we can reduce that exponential number of data to only a polynomial number of data (roughly $\epsilon^{-\bar{\eta}}$ for some $\bar{\eta} > 0$) to select the important experts, which is a considerable improvement in practice.

2. Expert Estimation: Apart from the benefit of sample complexity from dense-to-sparse to activation dense-to-sparse models, the parameter estimation rates also directly imply the convergence behavior of estimating expert networks, which is of practical interest. In particular, we consider linear experts of the form $a^{\top}x + b$ in our work, and show that the estimation rate of the expert $a^{\top}x + b$ is the slowest between the estimation rates of its parameters a and b. For instance, it can be seen from Table 1 that under the over-specified setting of the dense-to-sparse gate MoE model, the rate for estimating b^*_j is faster than that for a^*_j. Consequently, the expert estimation rate is equal to the rate for estimating a^*_j, which is slower than any polynomial rates. By contrast, under the over-specified setting of the activation dense-to-sparse gate MoE model, since the estimation rate of a^*_j is improved to be faster than that of b^*_j, the expert admits the same estimation rate as b^*_j, which is of order $\widetilde{\mathcal{O}}(n^{-1/2\bar{r}})$, faster the previous one. From this observation, we see that the parameter estimation problem also provides useful insights in designing the gating mechanism of MoE models, which helps enhance the performance of expert networks significantly.

3. Misspecified settings: In this paper, we consider a well-specified setting where the data are assumed to be sampled from the (activation) dense-to-sparse gating Gaussian MoE in order to lay the foundation for a more realistic yet challenging misspecified setting where the data are not necessarily generated from these models. Under that misspecified setting, we assume that the data are generated from the true but unknown conditional distribution $Q(Y|X)$, which is not either dense-to-sparse or activation dense-to-sparse mixture of experts. Then, we can demonstrate that the MLE $\widehat{G}_n$ converges to a mixing measure $\overline{G} \in \arg\min_{G \in \mathcal{G}_k(\Theta)} \mathrm{KL}(Q(Y|X)||g_G(Y|X))$ where KL stands for the Kullback-Leibler divergence, and $g_G(Y|X)$ is the conditional density of the (activation) dense-to-sparse gate Gaussian MoE. As $Q(Y|X)$ does not belong to the (activation) dense-to-sparse MoEs, the optimal mixing measure will be in the boundary of the parameter space, namely, the number of experts in $\overline{G}$ is k. Therefore, as n is sufficiently large, $\widehat{G}_n$ also has k experts.

The insights from our theories in the well-specified setting indicate that the Voronoi losses can be used to obtain the estimation rates of individual parameters of the MLE $\widehat{G}_n$ to those of $\overline{G}$. From Table 1 in the manuscript, we can see that

(3.1) Dense-to-sparse MoE: the worst parameter estimation rate of $\widehat{G}_n$ to $\overline{G}$ could be as slow as $1/\log^{\eta}(n)$ for some $\eta > 0$. It indicates that we still need exponential number of data (roughly $\exp(1/\epsilon^{\eta})$ where ϵ is the desired approximation error) to estimate $\overline{G}$.

(3.2) Activation dense-to-sparse MoE: Under the misspecified setting, the parameter estimation rate of $\widehat{G}_n$ to $\overline{G}$ is $n^{-1/2}$. It indicates that we only need roughly ϵ^{-2} where ϵ is the desired approximation error to estimate $\overline{G}$.

As a consequence, under the misspecified settings, the parameter estimation rates achieved when we use the activation dense-to-sparse gate MoE in the above KL divergence should be faster than those obtained when we use the dense-to-sparse gate MoE. This explains why the activation dense-to-sparse gate is a solution to the parameter estimation problem, or more generally, the expert estimation and selection problem of the MoE models.

4. Model design: From the benefits of the activation dense-to-sparse gate for the expert estimation of MoE models when using the temperature to smooth the expert selection process, we deduce that it would be better to use a gating network with sufficiently sophisticated activation functions (e.g. $\mathrm{sigmoid}$, GeLU, etc) rather than a simple linear gating network.

6. Concluding Remarks

In this paper, we investigate the effects of the dense-to-sparse gate on the convergence rates of maximum likelihood estimation under the Gaussian mixture of experts. We discover that the density estimation rate is parametric on the sample size. On the other hand, due to the interactions of the temperature with both gating and expert parameters via two partial differential equations, the rates for estimating them are slower than any polynomial rates, and therefore, could be as slow as $\mathcal{O}(1/\log(n))$. To enhance the sample efficiency of the temperature for the Gaussian mixture of experts, we design a novel gating function called activation dense-to-sparse, which routes the outputs of a linear layer to an activation function before sending them to the softmax function. We demonstrate that if the activation function meets the first-order (second-order) independence condition, then the aforementioned interactions disappear, and the parameter estimation rates become polynomial under the exact-specified (over-specified) settings.

Following from the results in this work, there are a few promising directions that we leave for future development:

Firstly, the convergence analysis of maximum likelihood estimation in this paper is carried out under the assumption that the data are sampled from the (activation) dense-to-sparse gating Gaussian mixture of experts. However, when the data are not necessarily generated from that model, such theoretical analysis has remained missing in the literature. More concretely, under that setting, the MLE converges to a mixing measure $\overline{G} \in \arg\min_{G \in \mathcal{G}_k(\Theta)} \mathrm{KL}(Q(Y|X)||g_G(Y|X))$ where $Q(Y|X)$ stands for the true conditional distribution of Y given X and KL denotes the Kullback-Leibler divergence. It is worth noting that current techniques for analyzing the convergence of the MLE apply only for the setting when the space of mixing measures is convex. Since the space $\mathcal{G}_k(\Theta)$ is non-convex, it is essential to develop further technical tools to establish the convergence rate of the MLE to the set of $\overline{G}$.

Secondly, the current theories are for probabilistic settings of dense-to-sparse gating mixture of experts, namely, when the expert functions are means of the Gaussian distribution. In practical applications, we usually consider the deterministic settings of dense-to-sparse gating mixture of experts of the form: $\sum_{i=1}^{k^*} \mathrm{Softmax}\left(\frac{(\beta_{1i}^*)^\top X + \beta_{0i}^*}{\tau}\right) h(X, \eta_i^*)$ where $h(., \eta_i^*)$ are general expert functions for all $i \in [k]$ and utilize the least square loss function to estimate the true parameters. Therefore, extending the current theories under the probabilistic settings to those under the deterministic settings is also practically important.

Acknowledgements

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

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. Given the theoretical nature of the paper, we believe that there are no potential societal consequences of our work.

References

Chen, Z., Deng, Y., Wu, Y., Gu, Q., and Li, Y. Towards understanding the mixture-of-experts layer in deep learning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23049–23062. Curran Associates, Inc., 2022.

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

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

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

Gulati, A., Qin, J., Chiu, C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pp. 5036–5040, 2020. doi: 10.21437/Interspeech.2020-3015.

Hazimeh, H., Zhao, Z., Chowdhery, A., Sathiamoorthy, M., Chen, Y., Mazumder, R., Hong, L., and Chi, E. DSelect-k: Differentiable Selection in the Mixture of Experts with Applications to Multi-Task Learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29335–29347. Curran Associates, Inc., 2021.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arxiv 1606.08415*, 2023.

Ho, N. and Nguyen, X. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44:2726–2755, 2016.

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

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural Computation*, 3, 1991.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mixtral of experts. *arxiv preprint arxiv 2401.04088*, 2024.

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

Jordan, M. I. and Xu, L. Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks*, 8(9):1409–1431, 1995. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(95)00014-3.

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., and Chen, Z. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In *International Conference on Learning Representations*, 2021.

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

Manole, T. and Ho, N. Refined convergence rates for maximum likelihood estimation under finite mixture models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14979–15006. PMLR, 17–23 Jul 2022.

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

Nguyen, H., Akbarian, P., Nguyen, T., and Ho, N. A general theory for softmax gating multinomial logistic mixture of experts. In *Proceedings of the ICML*, 2024a.

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

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

Nie, X., Miao, X., Cao, S., Ma, L., Liu, Q., Xue, J., Miao, Y., Liu, Y., Yang, Z., and Cui, B. Evomoe: An evolutional mixture-of-experts training framework via dense-to-sparse gate. *arXiv preprint arXiv:2112.14397*, 2022.

Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Pint, A. S., Keysers, D., and Houlsby, N. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8583–8595. Curran Associates, Inc., 2021.

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

Teicher, H. On the mixture of distributions. *Annals of Statistics*, 31:55–73, 1960.

Teicher, H. Identifiability of mixtures. *Annals of Statistics*, 32:244–248, 1961.

Teicher, H. Identifiability of finite mixtures. *Ann. Math. Statist.*, 32:1265–1269, 1963.

van de Geer, S. *Empirical processes in M-estimation*. Cambridge University Press, 2000.

You, Z., Feng, S., Su, D., and Yu, D. Speechmoe: Scaling to large acoustic models with dynamic routing mixture of experts. In *Interspeech*, 2021.

Zhou, Y., Du, N., Huang, Y., Peng, D., Lan, C., Huang, D., Shakeri, S., So, D., Dai, A., Lu, Y., Chen, Z., Le, Q., Cui, C., Laudon, J., and Dean, J. Brainformers: Trading simplicity for efficiency. In *International Conference on Machine Learning*, pp. 42531–42542. PMLR, 2023.

Supplementary Material for "Is Temperature Sample Efficient for Softmax Gaussian Mixture of Experts?"

In this supplementary material, we provide rigorous theoretical guarantee for the convergence rate of density estimation in Appendix A, while we leave that for parameter estimation in Appendix B. We study the identifiability of the (activation) dense-to-sparse gating Gaussian mixture of experts (MoE) in Appendix C. Finally, additional experiment setups are presented in Appendix D.

A. Proofs for Density Estimation Rates

A.1. Proof of Theorem 2.1

In this proof, we will leverage results regarding the convergence rates of density estimation from MLE in [Theorem 7.4, (van de Geer, 2000)]. Prior to presenting those result here, it is necessary to introduce some notations. Firstly, we denote by $\mathcal{G}_k(\Theta)$ the set of conditional densities of all mixing measures in $\mathcal{O}_k(\Theta)$, that is, $\mathcal{G}_k(\Theta) := \{g_G(Y|X) : G \in \mathcal{G}_k(\Theta)\}$. Next, we define

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

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

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

In order to measure the size of the above set, Geer et. al. (van de Geer, 2000) suggest using the following quantity:

$$\mathcal{J}_B(\delta, \widetilde{\mathcal{G}}_k^{1/2}(\Theta,\delta)) := \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(t, \widetilde{\mathcal{G}}_k^{1/2}(\Theta,t), \|\cdot\|)\,\mathrm{d}t \vee \delta, \tag{17}$$

where $H_B(t, \widetilde{\mathcal{G}}_k^{1/2}(\Theta,t), \|\cdot\|)$ stands for the bracketing entropy (van de Geer, 2000) of $\widetilde{\mathcal{G}}_k^{1/2}(\Theta,u)$ under the ℓ_2-norm, and $t \vee \delta := \max\{t,\delta\}$. Now, let us recall the statement of Theorem 7.4 in (van de Geer, 2000) with notations being adapted to this work.

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

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

for all $\delta \geq \delta_n$, where $h(g_1,g_2) := \left(\frac{1}{2}\int(\sqrt{g_1}-\sqrt{g_2})^2\mathrm{d}\mu\right)^{1/2}$ is the Hellinger distance w.r.t Lebesgue measure μ.

Proof of Lemma A.1 can be found in (van de Geer, 2000). Subsequently, we provide below a result on the bound for the bracketing entropy, which is essential for the proof of Theorem 2.1.

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

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

Proof of Lemma A.2 is deferred to Appendix A.1.2. Equipped with the results in Lemma A.1 and Lemma A.2, we present the proof of Theorem 2.1 in Appendix A.1.1.

A.1.1. Main Proof

It is worth noting that

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

for any $t > 0$. Then, we deduce from equation (17) that

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

where the second inequality occurs due to the upper bound of a bracketing entropy in Lemma A.2.

Denote $\Psi(\delta) = \delta \cdot [\log(1/\delta)]^{1/2}$, it is clear that $\Psi(\delta)/\delta^2$ is a non-increasing function of θ. Furthermore, it follows the above inequality that $\Psi(\delta) \geq \mathcal{J}_B(\delta, \widetilde{\mathcal{G}}_k^{1/2}(\Theta, \delta))$. Additionally, let $\delta_n = \sqrt{\log(n)/n}$, we get that $\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n)$ for some universal constant c. Now, by applying Lemma A.1, we obtain that

$$\mathbb{P}\Big(\mathbb{E}_X[h(g_{\widehat{G}_n}(\cdot|X), g_{G_*}(\cdot|X))] > C(\log(n)/n)^{1/2}\Big) \lesssim \exp(-c\log(n)),$$

for some universal constant C that depends only on Θ. Since the Hellinger distance is lower bounded by the Total Variation distance, i.e. $h \geq V$, we also achieve that

$$\mathbb{P}\Big(\mathbb{E}_X[V(g_{\widehat{G}_n}(\cdot|X), g_{G_*}(\cdot|X))] > C(\log(n)/n)^{1/2}\Big) \lesssim \exp(-c\log(n)).$$

Hence, we reach the conclusion that $\mathbb{E}_X[V(g_{\widehat{G}_n}(\cdot|X), g_{G_*}(\cdot|X))] = \widetilde{\mathcal{O}}(n^{-1/2})$.

A.1.2. Proof of Lemma A.2

First of all, we aim to derive an upper bound for the Gaussian density $f(Y|a^\top X + b, \nu)$. As both $\mathcal{X}$ and Θ are bounded sets, we can find positive constants κ, u, ℓ that satisfy $-\kappa \leq a^\top X + b \leq \kappa$ and $\ell \leq \nu \leq u$. Therefore, we have that

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

For any $|Y| \geq 2\kappa$, we get that $\frac{(Y-a^\top X-b)^2}{2\nu} \geq \frac{Y^2}{8u}$, implying that

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

Putting the above results together, it follows that $f(Y|a^\top X + b, \nu) \leq B(Y|X)$, where we define

$$B(Y|X) = \begin{cases} \frac{1}{\sqrt{2\pi\ell}} \exp\Big(-\frac{Y^2}{8u}\Big), & |Y| \geq 2\kappa; \\ \\ \frac{1}{\sqrt{2\pi\ell}}, & \text{otherwise.} \end{cases}$$

Let $\eta \leq \varepsilon$, we assume that the set $\mathcal{G}_k(\Theta)$ has an η-cover (under ℓ_1-norm) denoted by $\{\pi_1, \dots, \pi_N\}$, where $N := N(\eta, \mathcal{G}_k(\Theta), \|\cdot\|_1)$ is the η-covering number of the metric space $(\mathcal{G}_k(\Theta), \|\cdot\|_1)$. Then, we construct the brackets of the form $[L_i(Y|X), U_i(Y|X)]$ for all $i \in [N]$ as follows:

$$\begin{aligned} L_i(Y|X) &:= \max\{\pi_i(Y|X) - \eta, 0\}, \\ U_i(Y|X) &:= \min\{\pi_i(Y|X) + \eta, B(Y|X)\}. \end{aligned}$$

We can verify that $\mathcal{G}_k(\Theta) \subset \bigcup_{i=1}^N [L_i(Y|X), U_i(Y|X)]$ with a note that $0 \leq U_i(Y|X) - L_i(Y|X) \leq \min\{2\eta, B(Y|X)\}$. Next, for each $i \in [N]$, the term $\|U_i - L_i\|_1$ is upper bounded as follows:

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

in which $R := \max\{2\kappa, \sqrt{8u}\}\log(1/\eta)$ and $R' > 0$ is a universal constant. From the definition of bracketing entropy, $H_B(R'\eta, \mathcal{G}_k(\Theta), \|\cdot\|_1)$ is the logarithm of the smallest number of brackets of size $R'\eta$ necessary to cover $\mathcal{G}_k(\Theta)$, which leads to

$$H_B(R'\eta, \mathcal{G}_k(\Theta), \|\cdot\|_1) \leq \log N = \log N(\eta, \mathcal{G}_k(\Theta), \|\cdot\|_1). \tag{18}$$

As we demonstrate at the end of this proof, the covering number is bounded as $\log N(\eta, \mathcal{G}_k(\Theta), \|\cdot\|_1) \lesssim \log(1/\eta)$. This bound together with the result in equation (18) implies that

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

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

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

Upper bound of the covering number. For completion, we establish the following upper bound for the covering number:

$$\log N(\eta, \mathcal{G}_k(\Theta), \|\cdot\|_1) \lesssim \log(1/\eta).$$

Since Θ is a compact set, it follows that $\Delta := \{(\beta_0, \beta_1, \tau) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}_+ : (\beta_0, \beta_1, \tau, a, b, \nu) \in \Theta\}$ and $\Omega := \{(a, b, \nu) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+ : (\beta_0, \beta_1, \tau, a, b, \nu) \in \Theta\}$ are also compact. Therefore, there exist η-covers Δ_η and Ω_η for Δ and Ω, respectively. Additionally, we can validate that $|\Delta_\eta| \leq \mathcal{O}(\eta^{-(d+2)k})$ and $|\Omega_\eta| \leq \mathcal{O}(\eta^{-(d+2)k})$.

Subsequently, for each mixing measure $G = \sum_{i=1}^{k} \exp\left(\frac{\beta_{0i}}{\tau}\right)\delta_{(\beta_{1i},\tau,a_i,b_i,\nu_i)} \in \mathcal{O}_k(\Theta)$, we consider another one denoted by $\widetilde{G} := \sum_{i=1}^{k} \exp\left(\frac{\beta_{0i}}{\tau}\right)\delta_{(\beta_{1i},\tau,\overline{a}_i,\overline{b}_i,\overline{\nu}_i)}$, where $(\overline{a}_i, \overline{b}_i, \overline{\nu}_i) \in \Omega_\eta$ such that $(\overline{a}_i, \overline{b}_i, \overline{\nu}_i)$ are the closest to (a_i, b_i, ν_i) in that set for all $i \in [k]$. Besides, we also take into account the mixing measure $\overline{G} := \sum_{i=1}^{k} \exp\left(\frac{\overline{\beta}_{0i}}{\overline{\tau}}\right)\delta_{(\overline{\beta}_{1i},\overline{\tau},\overline{a}_i,\overline{b}_i,\overline{\nu}_i)}$, where $(\overline{\beta}_{0i}, \overline{\beta}_{1i}, \overline{\tau}) \in \Delta_\eta$ are the closest to $(\beta_{0i}, \beta_{1i}, \tau)$ in that set. It can be verified that the conditional density $g_{\overline{G}}$ belongs to the following set:

$$\mathcal{R} := \left\{g_G \in \mathcal{G}_k(\Theta) : (\beta_{0i}, \beta_{1i}, \tau) \in \Delta_\eta,\ (a_i, b_i, \nu_i) \in \Omega_\eta,\ \forall i \in [k]\right\}.$$

It follows from the formulation of $\widetilde{G}$ that

$$\begin{aligned}
\|g_G - g_{\widetilde{G}}\|_1 &\leq \sum_{i=1}^{k} \int \mathrm{Softmax}\Big(\frac{(\beta_{1i})^\top X + \beta_{0i}}{\tau}\Big) \cdot \Big|f(Y|(a_i)^\top X + b_i, \nu_i) - f(Y|(\overline{a}_i)^\top X + \overline{b}_i, \overline{\nu}_i)\Big| \mathrm{d}(X, Y) \\
&\leq \sum_{i=1}^{k} \int \Big|f(Y|(a_i)^\top X + b_i, \nu_i) - f(Y|(\overline{a}_i)^\top X + \overline{b}_i, \overline{\nu}_i)\Big| \mathrm{d}(X, Y) \\
&\lesssim \sum_{i=1}^{k} (\|a_i - \overline{a}_i\| + |b_i - \overline{b}_i| + |\nu_i - \overline{\nu}_i|) \\
&\lesssim \eta,
\end{aligned} \tag{19}$$

Since $\mathrm{Softmax}$ is a Lipschitz function with Lipschitz constant $L \geq 0$, we get

$$\begin{aligned}
\|g_{\widetilde{G}} - g_{\overline{G}}\|_1 &\leq \sum_{i=1}^{k} \int \Big|\mathrm{Softmax}\Big(\frac{(\beta_{1i})^\top X + \beta_{0i}}{\tau}\Big) - \mathrm{Softmax}\Big(\frac{(\overline{\beta}_{1i})^\top X + \overline{\beta}_{0i}}{\overline{\tau}}\Big)\Big| \cdot f(Y|(\overline{a}_i)^\top X + \overline{b}_i, \overline{\nu}_i) \mathrm{d}(X, Y) \\
&\lesssim L \cdot \sum_{i=1}^{k} \int \Big(\Big\|\frac{\beta_{1i}}{\tau} - \frac{\overline{\beta}_{1i}}{\overline{\tau}}\Big\| \cdot \|X\| + \Big|\frac{\beta_{0i}}{\tau} - \frac{\overline{\beta}_{0i}}{\overline{\tau}}\Big|\Big) \mathrm{d}(X, Y)
\end{aligned}$$

where the second inequality follows from the fact that the Gaussian density $f(Y|(\overline{a}_i)^\top X + \overline{b}_i, \overline{\nu}_i)$ is bounded. Note that

$$\Big\|\frac{\beta_{1i}}{\tau} - \frac{\overline{\beta}_{1i}}{\overline{\tau}}\Big\| = \Big\|\beta_{1i}\Big(\frac{1}{\tau} - \frac{1}{\overline{\tau}}\Big) + \frac{\beta_{1i} - \overline{\beta}_{1i}}{\overline{\tau}}\Big\| \leq \frac{\|\beta_{1i}\|}{\tau\overline{\tau}} \cdot |\tau - \overline{\tau}| + \frac{\|\beta_{1i} - \overline{\beta}_{1i}\|}{\overline{\tau}} \lesssim \eta.$$

Similarly, we also get that $\left|\frac{\beta_{0i}}{\tau} - \frac{\overline{\beta}_{0i}}{\overline{\tau}}\right| \lesssim \eta$. Moreover, since $\mathcal{X}$ is a bounded set, there exists a constant $B > 0$ such that $\|X\| \leq B$ for any $X \in \mathcal{X}$. As a result,

$$\|g_{\widetilde{G}} - g_{\overline{G}}\|_1 \lesssim L \cdot \sum_{i=1}^{k} \int (\eta \cdot B + \eta) \mathrm{d}(X,Y) \leq Lk\eta(B+1). \tag{20}$$

Putting the bounds in equations (19) and (20) together with the triangle inequality, we receive that

$$\|g_G - g_{\overline{G}}\|_1 \leq \|g_G - g_{\widetilde{G}}\|_1 + \|g_{\widetilde{G}} - g_{\overline{G}}\|_1 \lesssim \eta,$$

which means that $\mathcal{R}$ is an η-cover (not necessarily smallest) of the metric space $(\mathcal{G}_k(\Theta), \|\cdot\|_1)$. By definition of the covering number, we know that

$$N(\eta, \mathcal{G}_k(\Theta), \|\cdot\|_1) \leq |\mathcal{R}| = |\Delta_\eta| \times |\Omega_\eta| \leq \mathcal{O}(\eta^{-(d+2)k}) \cdot \mathcal{O}(\eta^{(-d+2)k}) \leq \mathcal{O}(\eta^{-(2d+4)k}),$$

which implies that

$$\log N(\eta, \mathcal{G}_k(\Theta), \|\cdot\|_1) \lesssim \log(1/\eta).$$

Hence, the proof is completed.

A.2. Proof of Theorem 3.1

Based on the proof of Theorem 2.1 in Appendix A.1, it suffices to establish the following upper bound for the covering number of the metric space $(\mathcal{P}_k(\Theta), \|\cdot\|_1)$, where $\mathcal{P}_k(\Theta) := \{p_G(Y|X) : G \in \mathcal{O}_k(\Theta)\}$, while other results can be demonstrated in a similar fashion:

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

Recall that Θ is a compact set, it follows that $\Delta := \{(\beta_0, \beta_1, \tau) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}_+ : (\beta_0, \beta_1, \tau, a, b, \nu) \in \Theta\}$ and $\Omega := \{(a, b, \nu) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+ : (\beta_0, \beta_1, \tau, a, b, \nu) \in \Theta\}$ are also compact. Therefore, there exist η-covers Δ_η and Ω_η for Δ and Ω, respectively, with a note that $|\Delta_\eta| \leq \mathcal{O}(\eta^{-(d+2)k})$ and $|\Omega_\eta| \leq \mathcal{O}(\eta^{-(d+2)k})$.

Next, for each mixing measure $G = \sum_{i=1}^{k} \exp\left(\frac{\beta_{0i}}{\tau}\right) \delta_{(\beta_{1i}, \tau, a_i, b_i, \nu_i)} \in \mathcal{O}_k(\Theta)$, we consider another one denoted by $\widetilde{G} := \sum_{i=1}^{k} \exp\left(\frac{\beta_{0i}}{\tau}\right) \delta_{(\beta_{1i}, \tau, \overline{a}_i, \overline{b}_i, \overline{\nu}_i)}$, where $(\overline{a}_i, \overline{b}_i, \overline{\nu}_i) \in \Omega_\eta$ such that $(\overline{a}_i, \overline{b}_i, \overline{\nu}_i)$ are the closest to (a_i, b_i, ν_i) in that set for all $i \in [k]$. Additionally, we also take into account the mixing measure $\overline{G} := \sum_{i=1}^{k} \exp\left(\frac{\overline{\beta}_{0i}}{\overline{\tau}}\right) \delta_{(\overline{\beta}_{1i}, \overline{\tau}, \overline{a}_i, \overline{b}_i, \overline{\nu}_i)}$, where $(\overline{\beta}_{0i}, \overline{\beta}_{1i}, \overline{\tau}) \in \Delta_\eta$ are the closest to $(\beta_{0i}, \beta_{1i}, \tau)$ in that set. It can be verified that the conditional density $p_{\overline{G}}$ belongs to the following set:

$$\mathcal{R} := \left\{p_G \in \mathcal{G}_k(\Theta) : (\beta_{0i}, \beta_{1i}, \tau) \in \Delta_\eta,\ (a_i, b_i, \nu_i) \in \Omega_\eta,\ \forall i \in [k]\right\}.$$

From the formulation of $\widetilde{G}$, we have that

$$\begin{aligned}
\|p_G - p_{\widetilde{G}}\|_1 &\leq \sum_{i=1}^{k} \int \mathrm{Softmax}\Big(\frac{\sigma((\beta_{1i})^\top X) + \beta_{0i}}{\tau}\Big) \cdot \Big|f(Y|(a_i)^\top X + b_i, \nu_i) - f(Y|(\overline{a}_i)^\top X + \overline{b}_i, \overline{\nu}_i)\Big| \mathrm{d}(X,Y) \\
&\leq \sum_{i=1}^{k} \int \Big|f(Y|(a_i)^\top X + b_i, \nu_i) - f(Y|(\overline{a}_i)^\top X + \overline{b}_i, \overline{\nu}_i)\Big| \mathrm{d}(X,Y) \\
&\lesssim \sum_{i=1}^{k} (\|a_i - \overline{a}_i\| + |b_i - \overline{b}_i| + |\nu_i - \overline{\nu}_i|) \\
&\lesssim \eta,
\end{aligned} \tag{21}$$

Since $\mathrm{Softmax}$ is a Lipschitz function with Lipschitz constant $L_1 \geq 0$, we get

$$\begin{aligned}\|p_{\widetilde{G}} - p_{\overline{G}}\|_1 &\leq \sum_{i=1}^{k} \int \Big|\mathrm{Softmax}\Big(\frac{\sigma((\beta_{1i})^\top X) + \beta_{0i}}{\tau}\Big) - \mathrm{Softmax}\Big(\frac{\sigma((\overline{\beta}_{1i})^\top X) + \overline{\beta}_{0i}}{\overline{\tau}}\Big)\Big| \cdot f(Y|(\overline{a}_i)^\top X + \overline{b}_i, \overline{\nu}_i)\mathrm{d}(X,Y)\\ &\lesssim L_1 \cdot \sum_{i=1}^{k} \int \Big(\Big|\frac{\sigma((\beta_{1i})^\top X)}{\tau} - \frac{\sigma((\overline{\beta}_{1i})^\top X)}{\overline{\tau}}\Big| + \Big|\frac{\beta_{0i}}{\tau} - \frac{\overline{\beta}_{0i}}{\overline{\tau}}\Big|\Big)\mathrm{d}(X,Y)\end{aligned}$$

where the second inequality follows from the fact that the Gaussian density $f(Y|(\overline{a}_i)^\top X + \overline{b}_i, \overline{\nu}_i)$ is bounded. Note that

$$\begin{aligned}&\Big|\frac{\sigma((\beta_{1i})^\top X)}{\tau} - \frac{\sigma((\overline{\beta}_{1i})^\top X)}{\overline{\tau}}\Big|\\ &= \Big|\sigma((\beta_{1i})^\top X)\Big(\frac{1}{\tau} - \frac{1}{\overline{\tau}}\Big) + \frac{\sigma((\beta_{1i})^\top X) - \sigma((\overline{\beta}_{1i})^\top X)}{\overline{\tau}}\Big|\\ &\leq \frac{|\sigma((\beta_{1i})^\top X)|}{\tau\overline{\tau}} \cdot |\tau - \overline{\tau}| + \frac{|\sigma((\beta_{1i})^\top X) - \sigma((\overline{\beta}_{1i})^\top X)|}{\overline{\tau}}.\end{aligned}$$

Since the function σ is differentiable, it is also Lipschitz with some Lipschitz constant $L_2 > 0$ and $|\sigma((\beta_{1i})^\top X)|$ is bounded. Furthermore, as $\mathcal{X}$ is a bounded set, it follows that $\|X\|$ is also bounded. Thus, we get

$$\begin{aligned}&\Big|\frac{\sigma((\beta_{1i})^\top X)}{\tau} - \frac{\sigma((\overline{\beta}_{1i})^\top X)}{\overline{\tau}}\Big|\\ &\leq \frac{|\sigma((\beta_{1i})^\top X)|}{\tau\overline{\tau}} \cdot |\tau - \overline{\tau}| + L_2 \cdot \frac{\|\beta_{1i} - \overline{\beta}_{1i}\| \cdot \|X\|}{\overline{\tau}}\\ &\leq \frac{|\sigma((\beta_{1i})^\top X)|}{\tau\overline{\tau}} \cdot \eta + L_2 \cdot \frac{\|X\|}{\overline{\tau}} \cdot \eta\\ &\lesssim \eta.\end{aligned}$$

Analogously, we also have that $\Big|\frac{\beta_{0i}}{\tau} - \frac{\overline{\beta}_{0i}}{\overline{\tau}}\Big| \lesssim \eta$. Consequently,

$$\|p_{\widetilde{G}} - p_{\overline{G}}\|_1 \lesssim L_1 \cdot \sum_{i=1}^{k} \int (\eta + \eta)\mathrm{d}(X,Y) \leq 2L_1 k\eta. \tag{22}$$

Putting the bounds in equations (21) and (22) together with the triangle inequality, we receive that

$$\|p_G - p_{\overline{G}}\|_1 \leq \|p_G - p_{\widetilde{G}}\|_1 + \|p_{\widetilde{G}} - p_{\overline{G}}\|_1 \lesssim \eta,$$

which means that $\mathcal{R}$ is an η-cover (not necessarily smallest) of the metric space $(\mathcal{P}_k(\Theta), \|\cdot\|_1)$. By definition of the covering number, we know that

$$N(\eta, \mathcal{P}_k(\Theta), \|\cdot\|_1) \leq |\mathcal{R}| = |\Delta_\eta| \times |\Omega_\eta| \leq \mathcal{O}(\eta^{-(d+2)k}) \cdot \mathcal{O}(\eta^{(-d+2)k}) \leq \mathcal{O}(\eta^{-(2d+4)k}),$$

which implies that

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

Hence, the proof is completed.

B. Proofs for Parameter Estimation Rates

B.1. Proof of Theorem 2.2

Before going to the main proof of Theorem 2.2 in Appendix B.1.1, let us introduce a key lemma for that proof as follows:

Lemma B.1. *For any $r \geq 1$, if the following holds :*

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{E}_{k_*}(\Theta): \mathcal{D}_{1,r}(G,G_*) \leq \varepsilon} \frac{\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_{1,r}(G,G_*)} = 0,$$

then we achieve that

$$\inf_{\overline{G}_n \in \mathcal{E}_{k_*}(\Theta)} \sup_{G \in \mathcal{E}_{k_*}(\Theta)} \mathbb{E}_{g_G}[\mathcal{D}_{1,r}(\overline{G}_n, G)] \gtrsim n^{-1/2}.$$

Proof of Lemma B.1 is deferred to Appendix B.1.2. Now, we are ready to present the main proof of Theorem 2.2.

B.1.1. Main Proof

Based on the result of Lemma B.1, it is sufficient to construct a sequence of mixing measures G_n such that $\mathcal{D}_{1,r}(G_n, G_*) \to 0$ and

$$\frac{\mathbb{E}_X[V(g_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_{1,r}(G_n, G_*)} \to 0, \tag{23}$$

as $n \to \infty$. For that purpose, we choose the following sequence: $G_n = \sum_{i=1}^{k_*} \exp\left(\frac{\beta_{0i}^n}{\tau^n}\right) \delta_{(\beta_{1i}^n, \tau^n, a_i^n, b_i^n, \nu_i^n)}$, where

- $a_i^n = a_i^*$, $b_i^n = b_i^*$, $\nu_i^n = \nu_i^*$ for any $i \in [k_*]$;
- $\beta_{1i}^n = \beta_{1i}^* + s_{n,i}$, for any $i \in [k_*]$;
- $\tau^n = \tau^* + t_n$;
- $\beta_{0i}^n = \Big(1 + \dfrac{t_n}{\tau^*}\Big)\beta_{0i}^*$, which implies that $\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right) = \exp\left(\frac{\beta_{0i}^*}{\tau^*}\right)$, for any $i \in [k_*]$

where $s_{n,i} := (s_{n,i}^{(1)}, \ldots, s_{n,i}^{(d)}) \in \mathbb{R}^d$ and $t_n \in \mathbb{R}$ will be chosen later such that $s_{n,i}^{(u)} \to 0$ and $t_n \to 0$ as $n \to \infty$ for any $u \in [d]$ and $i \in [k_*]$. Then, the loss function $\mathcal{D}_{1,r}$ is reduced to

$$\mathcal{D}_{1,r}(G_n, G_*) = \sum_{i=1}^{k_*} \exp\left(\frac{\beta_{0i}^*}{\tau^*}\right)(\|s_{n,i}\|^r + t_n^r). \tag{24}$$

It is clear that $\mathcal{D}_{1,r}(G_n, G_*) \to 0$ as $n \to \infty$. Now, we will show that $\mathbb{E}_X[V(g_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_{1,r}(G_n, G_*) \to 0$ as $n \to \infty$. Let us consider the quantity $Q_n := \Big[\sum_{i=1}^{k_*} \exp\Big(\dfrac{(\beta_{1i}^*)^\top X + \beta_{0i}^*}{\tau^*}\Big)\Big] \cdot \Big[g_{G_n}(Y|X) - g_{G_*}(Y|X)\Big]$, which can be decomposed as follows:

$$\begin{aligned} Q_n &= \sum_{i=1}^{k_*} \exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\Big[\exp\Big(\frac{(\beta_{1i}^n)^\top X}{\tau^n}\Big) f(Y|(a_i^n)^\top X + b_i^n, \nu_i^n) - \exp\Big(\frac{(\beta_{1i}^*)^\top X}{\tau^*}\Big) f(Y|(a_i^*)^\top X + b_i^*, \nu_i^*)\Big] \\ &- \sum_{i=1}^{k_*} \exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\Big[\exp\Big(\frac{(\beta_{1i}^n)^\top X}{\tau^n}\Big) g_{G_n}(Y|X) - \exp\Big(\frac{(\beta_{1i}^*)^\top X}{\tau^*}\Big) g_{G_n}(Y|X)\Big] \\ &+ \sum_{i=1}^{k_*} \Big[\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right) - \exp\left(\frac{\beta_{0i}^*}{\tau^*}\right)\Big]\Big[\exp\Big(\frac{(\beta_{1i}^*)^\top X}{\tau^*}\Big) f(Y|(a_i^*)^\top X + b_i^*, \nu_i^*) - \exp\Big(\frac{(\beta_{1i}^*)^\top X}{\tau^*}\Big) g_{G_n}(Y|X)\Big] \\ &:= A_n - B_n + E_n. \end{aligned}$$

Given the formulation of G_n, the term A_n can be simplified as

$$A_n = \sum_{i=1}^{k_*} \exp\left(\frac{\beta_{0i}^*}{\tau^*}\right)\Big[\exp\Big(\frac{(\beta_{1i}^n)^\top X}{\tau^n}\Big) - \exp\Big(\frac{(\beta_{1i}^*)^\top X}{\tau^*}\Big)\Big] f(Y|(a_i^*)^\top X + b_i^*, \nu_i^*)$$

By means of first-order Taylor expansions, we can rewrite A_n as

$$A_n = \sum_{i=1}^{k_*}\sum_{u=1}^{d}\exp\Big(\frac{\beta_{0i}^*}{\tau^*}\Big)\Big[\frac{s_{n,i}^{(u)}}{\tau^*}-\frac{t_n(\beta_{1i}^*)^{(u)}}{(\tau^*)^2}\Big]\cdot X^{(u)}\exp\Big(\frac{(\beta_{1i}^*)^\top X}{\tau^*}\Big)f(Y|(a_i^*)^\top X+b_i^*,\nu_i^*)+R_1(X,Y),$$

where $R_1(X,Y)$ is a Taylor remainder such that $R_1(X,Y)/\mathcal{D}_{1,r}(G_n,G_*)\to 0$ as $n\to\infty$. Then, by choosing

$$t_n=\frac{1}{n};\qquad s_{n,i}^{(u)}=\frac{t_n(\beta_{1i}^*)^{(u)}}{\tau^*}=\frac{(\beta_{1i}^*)^{(u)}}{n\tau^*},$$

we obtain that $A_n/\mathcal{D}_{1,r}(G_n,G_*)\to 0$ as $n\to\infty$.

Next, we consider the term B_n:

$$B_n=\sum_{i=1}^{k_*}\exp\Big(\frac{\beta_{0i}^*}{\tau^*}\Big)\Big[\exp\Big(\frac{(\beta_{1i}^n)^\top X}{\tau^n}\Big)-\exp\Big(\frac{(\beta_{1i}^*)^\top X}{\tau^*}\Big)\Big]g_{G_n}(Y|X).$$

By arguing similarly, we also get that $B_n/\mathcal{D}_{1,r}(G_n,G_*)\to 0$ as $n\to\infty$. Since we have $E_n=0$, it follows that $Q_n/\mathcal{D}_{1,r}(G_n,G_*)\to 0$ as $n\to\infty$. Moreover, as the term $\Big[\sum_{i=1}^{k_*}\exp\Big(\frac{(\beta_{1i}^*)^\top X+\beta_{0i}^*}{\tau^*}\Big)\Big]$ is bounded, we can deduce that $|g_{G_n}(\cdot|X)-g_{G_*}(\cdot|X)|/\mathcal{D}_{1,r}(G_n,G_*)\to 0$ as $n\to\infty$ for almost surely X. As a consequence, we satisfy the condition in equation (23). Hence, the proof is completed.

B.1.2. Proof of Lemma B.1

For a sufficiently small $\varepsilon>0$ and a fixed constant $C_1>0$ that we will choose later, it follows from the assumption that we can find a mixing measure $G'_*\in\mathcal{E}_{k_*}(\Theta)$ that satisfies $\mathcal{D}_{1,r}(G'_*,G_*)=2\varepsilon$ and $\mathbb{E}_X[V(g_{G'_*}(\cdot|X),g_{G_*}(\cdot|X))\leq C_1\varepsilon$. Additionally, for any sequence $\overline{G}_n\in\mathcal{E}_{k_*}(\Theta)$, we have

$$2\max_{G\in\{G'_*,G_*\}}\mathbb{E}_{g_G}[\mathcal{D}_{1,r}(\overline{G}_n,G)]\geq\mathbb{E}_{g_{G_*}}[\mathcal{D}_{1,r}(\overline{G}_n,G_*)]+\mathbb{E}_{g_{G'_*}}[\mathcal{D}_{1,r}(\overline{G}_n,G'_*)],$$

where $\mathbb{E}_{g_G}$ stands for the expectation taken w.r.t the product measure with density g_G^n. Furthermore, since the loss $\mathcal{D}_{1,r}$ satisfies the weak triangle inequality, we can find a constant $C_2>0$ such that

$$\mathcal{D}_{1,r}(\overline{G}_n,G_*)+\mathcal{D}_{1,r}(\overline{G}_n,G'_*)\geq C_2\mathcal{D}_{1,r}(G_*,G'_*)=2C_2\varepsilon.$$

Consequently, it follows that

$$\begin{aligned}\max_{G\in\{G_*,G'_*\}}\mathbb{E}_{g_G}[\mathcal{D}_{1,r}(\overline{G}_n,G)]&\geq\frac{1}{2}\Big(\mathbb{E}_{g_{G_*}}[\mathcal{D}_{1,r}(\overline{G}_n,G_*)]+\mathbb{E}_{g_{G'_*}}[\mathcal{D}_{1,r}(\overline{G}_n,G'_*)]\Big)\\&\geq C_2\varepsilon\cdot\inf_{f_1,f_2}\Big(\mathbb{E}_{g_{G_*}}[f_1]+\mathbb{E}_{g_{G'_*}}[f_2]\Big).\end{aligned}$$

Here, f_1 and f_2 in the above infimum are measurable functions in terms of $X_1,X_2,\ldots,X_n$ that satisfy $f_1+f_2=1$. By the definition of Total Variation distance, the above infimum value is equal to $1-\mathbb{E}_X[V(g_{G_*}^n(\cdot|X),g_{G'_*}^n(\cdot|X))]$. Therefore, we obtain that

$$\begin{aligned}\max_{G\in\{G_*,G'_*\}}\mathbb{E}_{g_G}[\mathcal{D}_{1,r}(\overline{G}_n,G)]&\geq C_2\varepsilon\Big(1-\mathbb{E}_X[V(g_{G_*}^n(\cdot|X),g_{G'_*}^n(\cdot|X))]\Big)\\&\geq C_2\varepsilon\Big[1-\sqrt{1-(1-C_1^2\varepsilon^2)^n}\Big].\end{aligned}$$

By choosing $\varepsilon=n^{-1/2}/C_1$, we have $C_1^2\varepsilon^2=\frac{1}{n}$, which implies that

$$\sup_{G\in\mathcal{E}_{k_*}(\Theta)\setminus\mathcal{O}_{k_*-1}(\Theta)}\mathbb{E}_{g_G}[\mathcal{D}_{1,r}(\overline{G}_n,G)]\geq\max_{G\in\{G_*,G'_*\}}\mathbb{E}_{g_G}[\mathcal{D}_{1,r}(\overline{G}_n,G)]\gtrsim n^{-1/2},$$

for any mixing measure $\overline{G}_n\in\mathcal{E}_{k_*}(\Theta)$. Hence, we reach the conclusion of Lemma B.1, that is,

$$\inf_{\overline{G}_n\in\mathcal{E}_{k_*}(\Theta)}\sup_{G\in\mathcal{E}_{k_*}(\Theta)\setminus\mathcal{O}_{k_*-1}(\Theta)}\mathbb{E}_{g_G}[\mathcal{D}_{1,r}(\overline{G}_n,G)]\gtrsim n^{-1/2},$$

for any $r\geq 1$.

B.2. Proof of Theorem 2.3

In this proof, our main goal is to prove the following inequality:

$$\inf_{G\in\mathcal{E}_{k_*}(\Theta)} \mathbb{E}_X[V(g_G(\cdot|X),g_{G_*}(\cdot|X))]/\mathcal{D}_2(G^{|\Psi},G_*^{|\Psi})>0. \tag{25}$$

For that purpose, we divide the above inequality into local and global parts as below.

Local part: In this part, we aim to establish the following inequality:

$$\lim_{\varepsilon\to 0}\inf_{G\in\mathcal{E}_{k_*}(\Theta):\mathcal{D}_2(G^{|\Psi},G_*^{|\Psi})\leq\varepsilon} \mathbb{E}_X[V(g_G(\cdot|X),g_{G_*}(\cdot|X))]/\mathcal{D}_2(G^{|\Psi},G_*^{|\Psi})>0. \tag{26}$$

Assume by contrary that the above inequality does not hold true, then there exists a sequence of mixing measures $G_n\in\mathcal{E}_{k_*}(\Theta)$ such that $G_n^{|\Psi}=\sum_{i=1}^{k_*}\exp(\beta_{0i}^n/\tau^n)\delta_{(a_i^n,b_i^n,\nu_i^n)}$ which satisfies $\mathcal{D}_{2n}:=\mathcal{D}_2(G_n^{|\Psi},G_*^{|\Psi})\to 0$ and

$$\mathbb{E}_X[V(g_{G_n}(\cdot|X),g_{G_*}(\cdot|X))]/\mathcal{D}_{2n}\to 0, \tag{27}$$

as $n\to\infty$. Recall that under the exact-specified settings, each Voronoi cell $\mathcal{A}_i^n=\mathcal{A}_i(G_n)$ has only one element. Therefore, we may assume without loss of generality (WLOG) that $\mathcal{A}_i^n=\{i\}$ for any $i\in[k_*]$. Thus, the loss function $\mathcal{D}_{2n}$ is reduced to

$$\mathcal{D}_{2n}:=\sum_{i=1}^{k_*}\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\Big[\|\Delta a_i^n\|+|\Delta b_i^n|+|\Delta\nu_i^n|\Big]+\sum_{i=1}^{k_*}\Big|\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)-\exp\Big(\frac{\beta_{0j}^*}{\tau^*}\Big)\Big| \tag{28}$$

Since $\mathcal{D}_{2n}\to 0$, we get that $(a_i^n,b_i^n,\nu_i^n)\to(a_i^*,b_i^*,\nu_i^*)$ and $\exp(\beta_{0i}^n/\tau^n)\to\exp(\beta_{0i}^*/\tau^*)$ as $n\to\infty$. Now, we separate the proof of local part into three steps as follows:

Step 1. In this step, we decompose the quantity $Q_n:=\Big[\sum_{i=1}^{k_*}\exp\Big(\dfrac{(\beta_{1i}^*)^\top X+\beta_{0i}^*}{\tau^*}\Big)\Big]\cdot[g_{G_n}(Y|X)-g_{G_*}(Y|X)]$ into a linear combination of linearly independent terms. For the ease of presentation, let us denote $F(Y;X,\beta_1,\tau,a,b,\nu):=\exp\Big(\frac{\beta_1^\top X}{\tau}\Big)f(Y|a^\top X+b,\nu)$ and $H(Y;X,\beta_1,\tau):=\exp\Big(\frac{\sigma(\beta_1^\top X)}{\tau}\Big)g_{G_n}(Y|X)$. Then, it can be checked that

$$\begin{aligned}
Q_n&=\sum_{i=1}^{k_*}\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\Big[F(Y;X,\beta_{1i}^n,\tau^n,a_i^n,b_i^n,\nu_i^n)-F(Y;X,\beta_{1i}^*,\tau^*,a_i^*,b_i^*,\nu_i^*)\Big]\\
&-\sum_{i=1}^{k_*}\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\Big[H(Y;X,\beta_{1i}^n,\tau^n)-H(Y;X,\beta_{1i}^*,\tau^*)\Big]\\
&+\sum_{i=1}^{k_*}\Big[\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)-\exp\Big(\frac{\beta_{0i}^*}{\tau^*}\Big)\Big]\cdot\exp\Big(\frac{(\beta_{1i}^*)^\top X}{\tau^*}\Big)f(Y|(a_i^*)^\top X+b_i^*,\nu_i^*)\\
&-\sum_{i=1}^{k_*}\Big[\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)-\exp\Big(\frac{\beta_{0i}^*}{\tau^*}\Big)\Big]\cdot\exp\Big(\frac{(\beta_{1i}^*)^\top X}{\tau^*}\Big)g_{G_n}(Y|X)\\
&:=A_n-B_n+E_{n,1}-E_{n,2}.
\end{aligned}$$

Next, by means of the first-order Taylor expansion, we get that

$$\begin{aligned}
A_n=\sum_{i=1}^{k_*}\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)&\sum_{|\alpha|=1}(\Delta\beta_{1i}^n)^{\alpha_1}(\Delta\tau^n)^{\alpha_2}(\Delta a_i^n)^{\alpha_3}(\Delta b_i^n)^{\alpha_4}(\Delta\nu_i^n)^{\alpha_5}\\
&\times\frac{\partial F}{\partial\beta_1^{\alpha_1}\,\partial\tau^{\alpha_2}\,\partial a^{\alpha_3}\,\partial b^{\alpha_4}\,\partial\nu^{\alpha_5}}(Y;X,\beta_{1i}^*,\tau^*,a_i^*,b_i^*,\nu_i^*)+R_1(X,Y),
\end{aligned}$$

where $R_1(X,Y)$ is a Taylor remainder such that $R_1(X,Y)/\mathcal{D}_{2n}\to 0$ as $n\to\infty$. Let us denote

$$F^{(\eta)}(Y;X,\omega_i^*):=\exp\Big(\frac{(\beta_{1i}^*)^\top X}{\tau^*}\Big)\cdot\frac{\partial^\eta f}{\partial h_1^\eta}(Y|(a_i^*)^\top X+b_i^*,\nu_i^*),$$

for any $\eta \in \mathbb{N}$, where $\omega_i^* := (\beta_{1i}^*, \tau^*, a_i^*, b_i^*, \nu_i^*)$. Then, the first derivatives of function F w.r.t its parameters are given by

$$\frac{\partial F}{\partial \beta_1^{(u)}}(Y;X,\omega_i^*) = \frac{X^{(u)}}{\tau^*}F(Y;X,\omega_i^*), \qquad \frac{\partial F}{\partial \tau}(Y;X,\omega_i^*) = -\frac{(\beta_{1i}^*)^\top X}{(\tau^*)^2}F(Y;X,\omega_i^*),$$
$$\frac{\partial F}{\partial a^{(u)}}(Y;X,\omega_i^*) = X^{(u)}F^{(1)}(Y;X,\omega_i^*), \quad \frac{\partial F}{\partial b}(Y;X,\omega_i^*) = F^{(1)}(Y;X,\omega_i^*), \quad \frac{\partial F}{\partial \nu}(Y;X,\omega_i^*) = \frac{1}{2}F^{(2)}(Y;X,\omega_i^*). \tag{29}$$

From this result, we can rewrite A_n as

$$A_n = \sum_{i=1}^{k_*} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\Big[\sum_{u=1}^{d}\Big(\frac{(\Delta\beta_{1i}^n)^{(u)}}{\tau^*} - \frac{(\Delta\tau^n)(\beta_{1i}^*)^{(u)}}{(\tau^*)^2}\Big)X^{(u)}F(Y;X,\omega_i^*) + \sum_{u=1}^{d}(\Delta a_i^n)^{(u)}X^{(u)}F^{(1)}(Y;X,\omega_i^*)$$
$$+ (\Delta b_i^n)F^{(1)}(Y;X,\omega_i^*) + \frac{1}{2}(\Delta\nu_i^n)F^{(2)}(Y;X,\omega_i^*)\Big] + R_1(X,Y),$$

Analogously, we also apply the first-order Taylor expansion to the term B_n and get that

$$\begin{aligned}B_n &= \sum_{i=1}^{k_*}\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\sum_{|\gamma|=1}(\Delta\beta_{1i}^n)^{\gamma_1}(\Delta\tau^n)^{\gamma_2}\cdot\frac{\partial H}{\partial\beta_1^{\gamma_1}\,\partial\tau^{\gamma_2}}(Y;X,\beta_{1i}^*,\tau^*) + R_2(X,Y),\\ &= \sum_{i=1}^{k_*}\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\sum_{u=1}^{d}\Big(\frac{\Delta\beta_{1i}^n}{\tau^*} - \frac{(\beta_{1i}^*)^{(u)}}{(\tau^*)^2}\Big)X^{(u)}H(Y;X,\beta_{1i}^*,\tau^*) + R_2(X,Y)\end{aligned}$$

where $R_2(X,Y)$ is a Taylor remainder such that $R_2(X,Y)/\mathcal{D}_{2n} \to 0$ as $n \to \infty$.

As a result, we can represent Q_n as

$$Q_n = \sum_{i=1}^{k_*}\sum_{\eta=0}^{2} C_{n,\eta,i}(X)F^{(\eta)}(Y;X,\omega_i^*) - \sum_{i=1}^{k_*}C_{n,0,i}(X)H(Y;X,\beta_{1i}^*,\tau^*) + R_1(X,Y) - R_2(X,Y), \tag{30}$$

where we define

$$\begin{aligned}C_{n,0,i}(X) &:= \Big[\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) - \exp\Big(\frac{\beta_{0i}^*}{\tau^*}\Big)\Big] + \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\sum_{u=1}^{d}\Big[\frac{(\Delta\beta_{1i}^n)^{(u)}}{\tau^*} - \frac{(\Delta\tau^n)(\beta_{1i}^*)^{(u)}}{(\tau^*)^2}\Big]\cdot X^{(u)},\\ C_{n,1,i}(X) &:= \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\Big[\sum_{u=1}^{d}(\Delta a_i^n)^{(u)}X^{(u)} + (\Delta b_i^n)\Big],\\ C_{n,2,i}(X) &:= \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\cdot\frac{(\Delta\nu_i^n)}{2}.\end{aligned}$$

From the above results, we can treat $[Q_n - R_1(X,Y) + R_2(X,Y)]/\mathcal{D}_{2n}$ as a combination of elements from the following set:

$$\Big\{F(Y;X,\omega_i^*),\ X^{(u)}F(Y;X,\omega_i^*),\ F^{(1)}(Y;X,\omega_i^*),\ X^{(u)}F^{(1)}(Y;X,\omega_i^*),\ F^{(2)}(Y;X,\omega_i^*) : u \in [d], i \in [k_*]\Big\}$$
$$\cup\Big\{H(Y;X,\beta_{1i}^*,\tau^*),\ X^{(u)}H(Y;X,\beta_{1i}^*,\tau^*) : u \in [d], i \in [k_*]\Big\}.$$

Step 2. In this step, we demonstrate that at least one among the coefficients in the representation of $[Q_n - R_1(X,Y) + R_2(X,Y)]/\mathcal{D}_{2n}$ does not converge to zero when $n \to \infty$. Assume by contrary that all of them go to 0 as $n \to \infty$. By taking the summation of the absolute values of the coefficients of $F(Y;X,\omega_i^*)$, we get that

$$\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{i=1}^{k_*}\Big|\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) - \exp\Big(\frac{\beta_{0i}^*}{\tau^*}\Big)\Big| \to 0. \tag{31}$$

Next, by taking the summation of the absolute values of the coefficients associated with

- $X^{(u)}F^{(1)}(Y;X,\omega_i^*)$: we have that $\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{i=1}^{k_*}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\|\Delta a_i^n\|_1\to 0$;
- $F^{(1)}(Y;X,\omega_i^*)$: we have that $\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{i=1}^{k_*}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)|\Delta b_i^n|\to 0$;
- $F^{(2)}(Y;X,\omega_i^*)$: we have that $\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{i=1}^{k_*}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)|\Delta \nu_i^n|\to 0$;

Due to the topological equivalence between ℓ_1-norm and ℓ_2-norm, it follows that

$$1=\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{i=1}^{k_*}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\Big(\|\Delta a_i^n\|+|\Delta b_i^n|+|\Delta\nu_i^n|\Big)\to 0, \tag{32}$$

which is a contradiction. Consequently, not all the coefficients in the representation of $[Q_n-R_1(X,Y)+R_2(X,Y)]/\mathcal{D}_{2n}$ converge to zero when $n\to\infty$.

Step 3. In this step, we leverage the Fatou's lemma to show a result contradicting to that in Step 2. In particular, by the Fatou's lemma, we have

$$\lim_{n\to\infty}\frac{\mathbb{E}_X[V(g_{G_n}(\cdot|X),g_{G_*}(\cdot|X))]}{\mathcal{D}_{2n}}\geq\int\liminf_{n\to\infty}\frac{|g_{G_n}(Y|X)-g_{G_*}(Y|X)|}{2\mathcal{D}_{2n}}\mathrm{d}(X,Y).$$

Moreover, recall from the hypothesis in equation (27) that $\mathbb{E}_X[V(g_{G_n}(\cdot|X),g_{G_*}(\cdot|X))]/\mathcal{D}_{2n}\to 0$ as $n\to\infty$. Therefore, we deduce that

$$\frac{|g_{G_n}(Y|X)-g_{G_*}(Y|X)|}{\mathcal{D}_{2n}}\to 0,$$

for almost surely (X,Y). Since the term $\left[\sum_{i=1}^{k_*}\exp\left(\frac{(\beta_{1i}^*)^\top X+\beta_{0i}^*}{\tau^*}\right)\right]$ is bounded, we also have that $\frac{Q_n}{\mathcal{D}_{2n}}\to 0$ as $n\to\infty$. Following from the results in equation (30), Q_n can be represented as

$$Q_n=\sum_{i=1}^{k_*}\sum_{\eta=0}^{2}C_{n,\eta,i}(X)F^{(\eta)}(Y;X,\omega_i^*)-\sum_{i=1}^{k_*}C_{n,0,i}(X)H(Y;X,\beta_{1i}^*,\tau^*)+R_1(X,Y)-R_2(X,Y).$$

Since $R_1(X,Y)/\mathcal{D}_{2n}\to 0$ and $R_2(X,Y)/\mathcal{D}_{2n}\to 0$ as $n\to\infty$, we can deduce that $C_{n,\eta,i}(X)/\mathcal{D}_{2n}$ must be bounded for any $\eta\in\{0,1,2\}$. Indeed, if at least one among them is not bounded, then that ratio will go to infinity, implying that $Q_n/\mathcal{D}_{2n}\not\to 0$, which is a contradiction. Thus, for each $\eta\in\{0,1,2\}$, we can replace $C_{n,\eta,i}(X)$ by one of its subsequences such that the ratio $C_{n,\eta,i}(X)/\mathcal{D}_{2n}$ has a finite limit as $n\to\infty$. Let us denote,

$$\begin{aligned}\frac{1}{\mathcal{D}_{2n}}\cdot\left[\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)-\exp\left(\frac{\beta_{0i}^*}{\tau^*}\right)\right]\to\phi_{0,i},&\qquad\frac{1}{\mathcal{D}_{2n}}\cdot\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)(\Delta\beta_{1i}^n)^{(u)}\to\phi_{1,i}^{(u)},\\ \frac{1}{\mathcal{D}_{2n}}\cdot\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)(\Delta\tau^n)\to\phi_{2,i},&\qquad\frac{1}{\mathcal{D}_{2n}}\cdot\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)(\Delta a_i^n)^{(u)}\to\phi_{3,i}^{(u)},\\ \frac{1}{\mathcal{D}_{2n}}\cdot\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)(\Delta b_i^n)\to\phi_{4,i},&\qquad\frac{1}{\mathcal{D}_{2n}}\cdot\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)(\Delta\nu_i^n)\to\phi_{5,i}.\end{aligned}$$

Then, we have

$$\frac{Q_n}{\mathcal{D}_{2n}}\to\sum_{i=1}^{k_*}\sum_{\eta=0}^{2}C_{\eta,i}^*(X)\cdot F^{(\eta)}(Y;X,\omega_i^*)-\sum_{i=1}^{k_*}C_{0,i}^*(X)\cdot H(Y;X,\beta_{1i}^*,\tau^*),$$

as $n\to\infty$, for almost surely (X,Y), where we define

$$C_{0,i}^*(X):=\phi_{0,i}+\sum_{u=1}^{d}\left[\frac{\phi_{1,i}^{(u)}}{\tau^*}-\frac{\phi_{2,i}}{(\tau^*)^2}\cdot(\beta_{1i}^*)^{(u)}\right]\cdot X^{(u)}, \tag{33}$$

$$C_{1,i}^*(X):=\sum_{u=1}^{d}\phi_{3,i}^{(u)}\cdot X^{(u)}+\phi_{4,i}, \tag{34}$$

$$C_{2,i}^*(X):=\frac{1}{2}\phi_{5,i}, \tag{35}$$

for any $i \in [k_*]$. In other words, we have

$$\sum_{i=1}^{k_*}\sum_{\eta=0}^{2} C^*_{\eta,i}(X) \cdot F^{(\eta)}(Y;X,\omega^*_i) - \sum_{i=1}^{k_*} C^*_{0,i}(X) \cdot H(Y;X,\beta^*_{1i},\tau^*) = 0,$$

for almost surely (X,Y). Since the set $\{F^{(\eta)}(Y;X,\omega^*_i),\ H(Y;X,\beta^*_{1i},\tau^*) : \eta \in \{0,1,2\},\ i \in [k_*]\}$ is linearly independent, we achieve that $C^*_{\eta,i}(X) = 0$ for almost surely X for any $\eta \in \{0,1,2\}$. As $C^*_{1,i}(X) = 0$, we deduce that $\phi^{(u)}_{3,i} = \phi_{3,i} = 0$ for any $u \in [d]$ and $i \in [k_*]$. Next, since $C^*_{2,i}(X) = 0$, we have that $\phi_{5,i} = 0$ for any $i \in [k_*]$. However, it follows from the results in Step 2 that at least one among $\phi^{(u)}_{3,i}$, $\phi_{4,i}$ and $\phi_{5,i}$ must be different from zero, which is a contradiction. Hence, we achieve the local inequality in equation (26). Therefore, we can find a constant $\varepsilon' > 0$ such that

$$\inf_{G \in \mathcal{E}_{k_*}(\Theta): \mathcal{D}_2(G^{|\Psi}, G_*^{|\Psi}) \leq \varepsilon'} \mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_2(G^{|\Psi}, G_*^{|\Psi}) > 0.$$

Global part. As a consequence, it suffices to demonstrate the following inequality:

$$\inf_{G \in \mathcal{E}_{k_*}(\Theta): \mathcal{D}_2(G^{|\Psi}, G_*^{|\Psi}) > \varepsilon'} \mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_2(G^{|\Psi}, G_*^{|\Psi}) > 0. \tag{36}$$

Assume by contrary that the above claim does not hold true, then we can seek a sequence of mixing measures $G'_n \in \mathcal{E}_{k_*}(\Omega)$ such that $\mathcal{D}_2((G'_n)^{|\Psi}, G_*^{|\Psi}) > \varepsilon'$ and

$$\lim_{n\to\infty} \frac{\mathbb{E}_X[V(g_{G'_n}(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_2((G'_n)^{|\Psi}, G_*^{|\Psi})} = 0,$$

which directly implies that $\mathbb{E}_X[V(g_{G'_n}(\cdot|X), g_{G_*}(\cdot|X))] \to 0$ as $n \to \infty$. Recall that Θ is a compact set, therefore, we can replace the sequence G'_n by one of its subsequences that converges to a mixing measure $G' \in \mathcal{E}_{k_*}(\Omega)$. Since $\mathcal{D}_2((G'_n)^{|\Psi}, G_*^{|\Psi}) > \varepsilon'$, this result induces that $\mathcal{D}_2((G')^{|\Psi}, G_*^{|\Psi}) > \varepsilon'$.

Next, by invoking the Fatou's lemma, it follows that

$$0 = \lim_{n\to\infty} \mathbb{E}_X[2V(g_{G'_n}(\cdot|X), g_{G_*}(\cdot|X))] \geq \int \liminf_{n\to\infty} \Big| g_{G'_n}(Y|X) - g_{G_*}(Y|X)\Big| \, \mathrm{d}(X,Y).$$

Thus, we get that $p_{G'}(Y|X) = g_{G_*}(Y|X)$ for almost surely (X,Y). From Proposition C.1, we know that the model (1) is identifiable, which indicates that $G' \equiv G_*$. As a consequence, we have that $\mathcal{D}_2((G')^{|\Psi}, G_*^{|\Psi}) = 0$, contradicting the fact that $\mathcal{D}_2((G')^{|\Psi}, G_*^{|\Psi}) > \varepsilon' > 0$.

Hence, the proof is completed.

B.3. Proof of Theorem 2.4

First of all, we provide a useful lemma that will be utilized for this proof as follows:

Lemma B.2. *For any $r \geq 1$, if the following holds :*

$$\lim_{\varepsilon\to 0} \inf_{G \in \mathcal{G}_k(\Theta): \mathcal{D}_{3,r}(G,G_*) \leq \varepsilon} \frac{\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_{3,r}(G,G_*)} = 0,$$

then we achieve that

$$\inf_{\overline{G}_n \in \mathcal{G}_k(\Theta)} \sup_{G \in \mathcal{G}_k(\Theta) \setminus \mathcal{O}_{k_*-1}(\Theta)} \mathbb{E}_{g_G}[\mathcal{D}_{3,r}(\overline{G}_n, G)] \gtrsim n^{-1/2}.$$

The proof of Lemma B.2 can be done similarly as in Appendix B.1.2. Following from this lemma, it suffices to build a sequence of mixing measures G_n that satisfies $\mathcal{D}_{3,r}(G_n, G_*) \to 0$ and

$$\frac{\mathbb{E}_X[V(g_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_{3,r}(G_n, G_*)} \to 0, \tag{37}$$

as $n \to \infty$. To this end, we take into account the mixing measure sequence $G_n = \sum_{i=1}^{k_*} \exp\left(\frac{\beta^n_{0i}}{\tau^n}\right)\delta_{(\beta^n_{1i},\tau^n,a^n_i,b^n_i,\nu^n_i)}$, where we define for any $j \in [k_*]$ that

- $a_i^n = a_j^*$, $b_i^n = b_j^*$, $\nu_i^n = \nu_j^*$ for any $i \in \mathcal{A}_j$;
- $\beta_{1i}^n = \beta_{1j}^* + s_{n,j}$, for any $i \in \mathcal{A}_j$;
- $\tau^n = \tau^* + t_n$;
- $\beta_{0i}^n = \tau^n \cdot \Big[\frac{\beta_{0j}^*}{\tau^*} - \log(|\mathcal{A}_j|)\Big]$, which implies that $\sum_{i\in\mathcal{A}_j} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) = \exp\Big(\frac{\beta_{0j}^*}{\tau^*}\Big)$, for any $i \in \mathcal{A}_j$,

where $s_{n,j} := (s_{n,j}^{(1)}, \ldots, s_{n,j}^{(d)}) \in \mathbb{R}^d$ and $t_n \in \mathbb{R}$ will be chosen later such that $s_{n,j}^{(u)} \to 0$ and $t_n \to 0$ as $n \to \infty$ for any $u \in [d]$ and $j \in [k_*]$. Then, the loss function $\mathcal{D}_{3,r}$ is reduced to

$$\mathcal{D}_{3,r}(G_n, G_*) = \sum_{j=1}^{k_*} |\mathcal{A}_j| \cdot \exp\Big(\frac{\beta_{0j}^*}{\tau^*}\Big)(\|s_{n,j}\|^r + t_n^r).$$

Obviously, we have that $\mathcal{D}_{3,r}(G_n, G_*) \to 0$ as $n \to \infty$.

Now, we will show that $\mathbb{E}_X[V(g_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_{3,r}(G_n, G_*) \to 0$ as $n \to \infty$. Let us consider the quantity $Q_n := \Big[\sum_{j=1}^{k_*} \exp\Big(\dfrac{(\beta_{1i}^*)^\top X + \beta_{0i}^*}{\tau^*}\Big)\Big] \cdot \Big[g_{G_n}(Y|X) - g_{G_*}(Y|X)\Big]$, which can be represented as as follows:

$$\begin{aligned}
Q_n &= \sum_{j=1}^{k_*} \sum_{i\in\mathcal{A}_j} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\Big[\exp\Big(\frac{(\beta_{1i}^n)^\top X}{\tau^n}\Big) f(Y|(a_i^n)^\top X + b_i^n, \nu_i^n) - \exp\Big(\frac{(\beta_{1j}^*)^\top X}{\tau^*}\Big) f(Y|(a_j^*)^\top X + b_j^*, \nu_j^*)\Big] \\
&\quad - \sum_{j=1}^{k_*} \sum_{i\in\mathcal{A}_j} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\Big[\exp\Big(\frac{(\beta_{1i}^n)^\top X}{\tau^n}\Big) g_{G_n}(Y|X) - \exp\Big(\frac{(\beta_{1j}^*)^\top X}{\tau^*}\Big) g_{G_n}(Y|X)\Big] \\
&\quad + \sum_{j=1}^{k_*} \Big[\sum_{i\in\mathcal{A}_j} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) - \exp\Big(\frac{\beta_{0j}^*}{\tau^*}\Big)\Big]\Big[\exp\Big(\frac{(\beta_{1j}^*)^\top X}{\tau^*}\Big) f(Y|(a_j^*)^\top X + b_j^*, \nu_j^*) - \exp\Big(\frac{(\beta_{1j}^*)^\top X}{\tau^*}\Big) g_{G_n}(Y|X)\Big] \\
&:= A_n - B_n + E_n.
\end{aligned}$$

Following from the formulation of G_n, we can rewrite the term A_n as

$$A_n = \sum_{i=1}^{k_*} \sum_{i\in\mathcal{A}_j} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\Big[\exp\Big(\frac{(\beta_{1i}^n)^\top X}{\tau^n}\Big) - \exp\Big(\frac{(\beta_{1j}^*)^\top X}{\tau^*}\Big)\Big] f(Y|(a_j^*)^\top X + b_j^*, \nu_j^*).$$

By means of first-order Taylor expansions, we can rewrite A_n as

$$A_n = \sum_{j=1}^{k_*} \sum_{i\in\mathcal{A}_j} \sum_{u=1}^{d} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\Big[\frac{s_{n,j}^{(u)}}{\tau^*} - \frac{t_n(\beta_{1j}^*)^{(u)}}{(\tau^*)^2}\Big] \cdot X^{(u)} \exp\Big(\frac{(\beta_{1j}^*)^\top X}{\tau^*}\Big) f(Y|(a_j^*)^\top X + b_j^*, \nu_j^*) + R_1(X, Y),$$

where $R_1(X, Y)$ is a Taylor remainder such that $R_1(X, Y)/\mathcal{D}_{1,r}(G_n, G_*) \to 0$ as $n \to \infty$. Then, by choosing

$$t_n = \frac{1}{n}; \qquad s_{n,j}^{(u)} = \frac{t_n(\beta_{1j}^*)^{(u)}}{\tau^*} = \frac{(\beta_{1j}^*)^{(u)}}{n\tau^*},$$

we obtain that $A_n/\mathcal{D}_{3,r}(G_n, G_*) \to 0$ as $n \to \infty$.

By arguing in the same fashion, we also get that $B_n/\mathcal{D}_{3,r}(G_n, G_*) \to 0$ as $n \to \infty$. As we have $E_n = 0$, it follows that $Q_n/\mathcal{D}_{3,r}(G_n, G_*) \to 0$ as $n \to \infty$. Moreover, since the term $\Big[\sum_{j=1}^{k_*} \exp\Big(\dfrac{(\beta_{1j}^*)^\top X + \beta_{0j}^*}{\tau^*}\Big)\Big]$ is bounded, we can deduce that $|g_{G_n}(Y|X) - g_{G_*}(Y|X)|/\mathcal{D}_{3,r}(G_n, G_*) \to 0$ as $n \to \infty$ for almost surely (X, Y). As a consequence, we satisfy the condition in equation (37). Hence, the proof is completed.

B.4. Proof of Theorem 2.5

Analogous to the proof of Theorem 2.3 in Appendix B.2, we aim to prove the following inequality:

$$\inf_{G\in\mathcal{G}_k(\Theta)} \mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_4(G^{|\Upsilon}, G_*^{|\Upsilon}) > 0. \tag{38}$$

Moreover, we also divide the above inequality into local and global parts. Since the global part can be argued in the same fashion as in Appendix B.2, we will demonstrate only the local part, that is

$$\lim_{\varepsilon\to 0} \inf_{G\in\mathcal{G}_k(\Theta):\mathcal{D}_4(G^{|\Upsilon},G_*^{|\Upsilon})\leq\varepsilon} \mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_4(G^{|\Upsilon}, G_*^{|\Upsilon}) > 0. \tag{39}$$

Assume by contrary that the above claim does not hold true, then we can find a sequence of mixing measures $G_n = \sum_{i=1}^{k_*} \exp(\beta_{0i}^n/\tau^n)\delta_{(\beta_{1i}^n,\tau^n,a_i^n,b_i^n,\nu_i^n)}$ in $\mathcal{G}_k(\Theta)$ that satisfies $\mathcal{D}_{4n} := \mathcal{D}_4(G_n^{|\Upsilon}, G_*^{|\Upsilon}) \to 0$ and

$$\mathbb{E}_X[V(g_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_{4n} \to 0, \tag{40}$$

as $n\to\infty$. Let us denote $\mathcal{A}_j^n = \mathcal{A}_j(G_n)$, then the loss function $\mathcal{D}_{6n}$ is reduced to

$$\begin{aligned}\mathcal{D}_{4n} := &\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j} \exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\Big[|\Delta b_{ij}^n|^{\bar{r}_j} + |\Delta\nu_{ij}^n|^{\bar{r}_j/2}\Big] \\ &+ \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j} \exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\Big[|\Delta b_{ij}^n| + |\Delta\nu_{ij}^n|\Big] + \sum_{j=1}^{k_*}\Big|\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right) - \exp\left(\frac{\beta_{0j}^*}{\tau^*}\right)\Big|\end{aligned} \tag{41}$$

As $\mathcal{D}_{4n}\to 0$, we get that $(b_i^n,\nu_i^n)\to(b_j^*,\nu_j^*)$ and $\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n/\tau^n)\to\exp(\beta_{0j}^*/\tau^*)$ as $n\to\infty$ for any $i\in\mathcal{A}_j$ and $j\in[k_*]$. Now, we divide the proof of local part into three steps as follows:

Step 1. In this step, we decompose the quantity $Q_n := \Big[\sum_{j=1}^{k_*}\exp\Big(\dfrac{\beta_{1j}^*X+\beta_{0j}^*}{\tau^*}\Big)\Big]\cdot[g_{G_n}(Y|X) - g_{G_*}(Y|X)]$ into a linear combination of linearly independent terms. Firstly, let $F(Y;X,\beta_1,\tau,a,b,\nu) := \exp\Big(\frac{\beta_1 X}{\tau}\Big)f(Y|aX+b,\nu)$ and $H(Y;X,\beta_1,\tau) := \exp\Big(\frac{\beta_1 X}{\tau}\Big)g_{G_n}(Y|X)$. Then, it can be verified that

$$\begin{aligned}Q_n = &\sum_{j=1}^{k_*}\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\Big[F(Y;X,\beta_{1i}^n,\tau^n,a_i^n,b_i^n,\nu_i^n) - F(Y;X,\beta_{1j}^*,\tau^*,a_j^*,b_j^*,\nu_j^*)\Big] \\ &-\sum_{j=1}^{k_*}\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\Big[H(Y;X,\beta_{1i}^n,\tau^n) - H(Y;X,\beta_{1j}^*,\tau^*)\Big] \\ &+\sum_{j=1}^{k_*}\Big[\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right) - \exp\left(\frac{\beta_{0i}^*}{\tau^*}\right)\Big]\cdot\exp(\beta_{1j}^*X)f(Y|a_j^*X+b_j^*,\nu_j^*) \\ &-\sum_{j=1}^{k_*}\Big[\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right) - \exp\left(\frac{\beta_{0i}^*}{\tau^*}\right)\Big]\cdot\exp(\beta_{1j}^*X)g_{G_n}(Y|X) \\ :&= A_n - B_n + E_{n,1} - E_{n,2}.\end{aligned} \tag{42}$$

Next, we continue to separate A_n into two terms as follows:

$$\begin{aligned}A_n : = &\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\Big[F(Y;X,\beta_{1i}^n,\tau^n,a_i^n,b_i^n,\nu_i^n) - F(Y;X,\beta_{1j}^*,\tau^*,a_j^*,b_j^*,\nu_j^*)\Big] \\ &+\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\Big[F(Y;X,\beta_{1i}^n,\tau^n,a_i^n,b_i^n,\nu_i^n) - F(Y;X,\beta_{1j}^*,\tau^*,a_j^*,b_j^*,\nu_j^*)\Big] \\ :&= A_{n,1} + A_{n,2}.\end{aligned}$$

Let us denote

$$F^{(\eta)}(Y;X,\omega_j^*) := \exp\left(\frac{\beta_{1j}^* X}{\tau^*}\right)\frac{\partial^\eta f}{\partial h_1^\eta}(Y|a_j^*X+b_j^*,\nu_j^*),$$

for any $\eta \in \mathbb{N}$, where $\omega_j^* := (\beta_{1j}^*, \tau^*, a_j^*, b_j^*, \nu_j^*)$. Then, by applying the first-order Taylor expansion as in equation (57), the term $A_{n,1}$ can be decomposed as

$$\begin{aligned}
A_{n,1} &= \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\sum_{|\alpha|=1}\frac{1}{\alpha!}(\Delta\beta_{1ij}^n)^{\alpha_1}(\Delta\tau^n)^{\alpha_2}(\Delta a_{ij}^n)^{\alpha_3}(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5}\\
&\qquad\qquad\qquad\times\frac{\partial F}{\partial\beta_1^{\alpha_1}\,\partial\tau^{\alpha_2}\,\partial a^{\alpha_3}\,\partial b^{\alpha_4}\,\partial\nu^{\alpha_5}}(Y;X,\omega_j^*)+R_1(X,Y)\\
&= \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\sum_{|\alpha|=1}\frac{1}{\alpha!}(\Delta\beta_{1ij}^n)^{\alpha_1}(\Delta\tau^n)^{\alpha_2}(\Delta a_{ij}^n)^{\alpha_3}(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5}\\
&\qquad\times\frac{X^{\alpha_1}}{(\tau^*)^{\alpha_1}}\cdot\Big(\sum_{w=1}^{\alpha_2}\frac{c_{w,\beta_{1j}^*,\tau^*}}{(\tau^*)^{\alpha_2}}X^w\Big)\cdot X^{\alpha_3}F^{(\alpha_3+\alpha_4+2\alpha_5)}(Y;X,\omega_j^*)+R_1(X,Y),
\end{aligned}$$

where $R_1(X,Y)$ is a Taylor remainder such that $R_1(X,Y)/\mathcal{D}_{4n} \to 0$ as $n \to \infty$. By letting $\ell_1 = \alpha_1 + w + \alpha_3$ and $\ell_2 = \alpha_3 + \alpha_4 + 2\alpha_5$, we obtain that

$$\begin{aligned}
A_{n,1} = \sum_{j:|\mathcal{A}_j|=1}\sum_{\ell_1+\ell_2=1}^{2}\sum_{i\in\mathcal{A}_j}\sum_{\alpha\in\mathcal{I}_{\ell_1,\ell_2}}&\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\frac{c_{\ell_1-\alpha_1-\alpha_3,\beta_{1j}^*,\tau^*}}{\alpha!2^{\alpha_5}(\tau^*)^{\alpha_1+\alpha_2}}(\Delta\beta_{1ij}^n)^{\alpha_1}(\Delta\tau^n)^{\alpha_2}(\Delta a_{ij}^n)^{\alpha_3}(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5}\\
&\times X^{\ell_1}F^{(\ell_2)}(Y;X,\omega_j^*)+R_1(X,Y),
\end{aligned}\tag{43}$$

where

$$\mathcal{I}_{\ell_1,\ell_2} := \{(\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5)\in\mathbb{N}^5 : \alpha_1+\alpha_2+\alpha_3\geq\ell_1,\ \alpha_3+\alpha_4+\alpha_5=\ell_2\}.$$

Regarding $A_{n,2}$, for each $j : |\mathcal{A}_j| > 1$, by means of Taylor expansion of order $\bar{r}_j$, we have

$$\begin{aligned}
A_{n,2} = \sum_{j:|\mathcal{A}_j|>1}\sum_{\ell_1+\ell_2=1}^{2\bar{r}_j}\sum_{i\in\mathcal{A}_j}\sum_{\alpha\in\mathcal{I}_{\ell_1,\ell_2}}&\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\frac{c_{\ell_1-\alpha_1-\alpha_3,\beta_{1j}^*,\tau^*}}{\alpha!2^{\alpha_5}(\tau^*)^{\alpha_1+\alpha_2}}(\Delta\beta_{1ij}^n)^{\alpha_1}(\Delta\tau^n)^{\alpha_2}(\Delta a_{ij}^n)^{\alpha_3}(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5}\\
&\times X^{\ell_1}F^{(\ell_2)}(Y;X,\omega_j^*)+R_2(X,Y),
\end{aligned}\tag{44}$$

where $R_2(X,Y)$ is a Taylor remainder such that $R_2(X,Y)/\mathcal{D}_{4n} \to 0$ as $n \to \infty$. From the results in equations (43), (44) and the definition of $E_{n,1}$, we get

$$A_n + E_{n,1} = \sum_{j=1}^{k_*}\sum_{\ell_1+\ell_2=0}^{2\bar{r}_j} Z_{\ell_1,\ell_2,j}^n \cdot X^{\ell_1}F^{(\ell_2)}(Y;X,\omega_j^*)+R_1(X,Y)+R_2(X,Y),\tag{45}$$

where

$$Z_{\ell_1,\ell_2,j}^n := \begin{cases}\sum_{i\in\mathcal{A}_j}\sum_{\alpha\in\mathcal{I}_{\ell_1,\ell_2}}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\frac{c_{\ell_1-\alpha_1-\alpha_3,\beta_{1j}^*,\tau^*}}{\alpha!2^{\alpha_5}(\tau^*)^{\alpha_1+\alpha_2}}(\Delta\beta_{1ij}^n)^{\alpha_1}(\Delta\tau^n)^{\alpha_2}(\Delta a_{ij}^n)^{\alpha_3}(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5}, \\ \hfill (\ell_1,\ell_2)\neq(0,0); \\ \\ \sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)-\exp\left(\frac{\beta_{0i}^*}{\tau^*}\right), \hfill (\ell_1,\ell_2)=(0,0).\end{cases}$$

Subsequently, we also separate B_n into two terms:

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

By applying the first-order Taylor expansion to $B_{n,1}$, we have

$$\begin{aligned}B_{n,1} &= \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)\sum_{|\gamma|=1}\frac{1}{\gamma!}(\Delta\beta^n_{1ij})^{\gamma_1}(\Delta\tau^n)^{\gamma_2}\cdot\frac{\partial^{\gamma_1+\gamma_2}H}{\partial\beta_1^{\gamma_1}\partial\tau^{\gamma_2}}(Y;X,\beta^*_{1j},\tau^*)+R_3(X,Y)\\ &= \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)\sum_{|\gamma|=1}\frac{1}{\gamma!}(\Delta\beta^n_{1ij})^{\gamma_1}(\Delta\tau^n)^{\gamma_2}\\ &\qquad\times\frac{X^{\gamma_1}}{(\tau^*)^{\gamma_1}}\cdot\Big(\sum_{w=1}^{\gamma_2}\frac{c_{w,\beta^*_{1j},\tau^*}}{(\tau^*)^{\gamma_2}}X^w\Big)H(Y;X,\beta^*_{1j},\tau^*)+R_3(X,Y),\end{aligned}$$

where $R_3(X,Y)$ is a Taylor remainder such that $R_3(X,Y)/\mathcal{D}_{4n}\to 0$ as $n\to\infty$. By letting $\ell=\gamma_1+w$, we rewrite $B_{n,1}$ as

$$B_{n,1}=\sum_{j:|\mathcal{A}_j|=1}\sum_{\ell=1}\sum_{i\in\mathcal{A}_j}\sum_{\gamma\in\mathcal{J}_\ell}\frac{c_{\ell-\gamma_1,\beta^*_{1j},\tau^*}}{\gamma!(\tau^*)^{\gamma_1+\gamma_2}}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)(\Delta\beta^n_{1ij})^{\gamma_1}(\Delta\tau^n)^{\gamma_2}\cdot X^\ell H(Y;X,\beta^*_{1j},\tau^*)+R_3(X,Y). \tag{46}$$

Regarding $B_{n,2}$, by means of the second-order Taylor expansion, we get

$$B_{n,2}=\sum_{j:|\mathcal{A}_j|>1}\sum_{\ell=1}^{2}\sum_{i\in\mathcal{A}_j}\sum_{\gamma\in\mathcal{J}_\ell}\frac{c_{\ell-\gamma_1,\beta^*_{1j},\tau^*}}{\gamma!(\tau^*)^{\gamma_1+\gamma_2}}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)(\Delta\beta^n_{1ij})^{\gamma_1}(\Delta\tau^n)^{\gamma_2}\cdot X^\ell H(Y;X,\beta^*_{1j},\tau^*)+R_4(X,Y), \tag{47}$$

where $R_4(X,Y)$ is a Taylor remainder such that $R_4(X,Y)/\mathcal{D}_{4n}\to 0$ as $n\to\infty$. From the results in equations (46), (47) and the definition of $E_{n,2}$, we obtain that

$$B_n+E_{n,2}=\sum_{j=1}^{k_*}\sum_{\ell=0}^{1+\mathbf{1}_{|\mathcal{A}_j|>1}}Z^n_{\ell,0,j}\cdot X^\ell H(Y;X,\beta^*_{1j},\tau^*)+R_3(X,Y)+R_4(X,Y). \tag{48}$$

Combine equation (45) with equation (48), we have

$$\begin{aligned}Q_n=\sum_{j=1}^{k_*}\sum_{\ell_1+\ell_2=0}^{2\bar{r}_j}Z^n_{\ell_1,\ell_2,j}\cdot X^{\ell_1}F^{(\ell_2)}(Y;X,\omega^*_j)-\sum_{j=1}^{k_*}\sum_{\ell=0}^{1+\mathbf{1}_{|\mathcal{A}_j|>1}}Z^n_{\ell,0,j}\cdot X^\ell H(Y;X,\beta^*_{1j},\tau^*)&\\ +R_1(X,Y)+R_2(X,Y)-R_3(X,Y)-R_4(X,Y).&\end{aligned} \tag{49}$$

As a consequence, we can view $[Q_n-R_1(X,Y)-R_2(X,Y)+R_3(X,Y)+R_4(X,Y)]/\mathcal{D}_{4n}$ as a combination of elements from the following set:

$$\mathcal{S}:=\Big\{X^{\ell_1}F^{(\ell_2)}(Y;X,\omega^*_j),\ X^\ell H(Y;X,\beta^*_{1j},\tau^*):j\in[k_*],\ 0\leq\ell_1+\ell_2\leq 2\bar{r}_j,\ 0\leq\ell\leq 1+\mathbf{1}_{|\mathcal{A}_j|>1}\Big\}.$$

Step 2. In this step, we show that at least one among the ratios $Z^n_{\ell_1,\ell_2,j}/\mathcal{D}_{4n}$ does not converge to zero as $n\to\infty$. Assume by contrary that all of them go to zero. Then, by taking the summation of the absolute values of

- $Z^n_{0,0,j}/\mathcal{D}_{4n}$ for $j\in[k_*]$, we have $\frac{1}{\mathcal{D}_{4n}}\cdot\sum_{j=1}^{k_*}\Big|\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)-\exp\Big(\frac{\beta^*_{0j}}{\tau^*}\Big)\Big|\to 0$;
- $Z^n_{0,1,j}/\mathcal{D}_{4n}$ for $j:|\mathcal{A}_j|=1$, we have $\frac{1}{\mathcal{D}_{4n}}\cdot\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)|\Delta b^n_{ij}|\to 0$;
- $Z^n_{0,2,j}/\mathcal{D}_{4n}$ for $j:|\mathcal{A}_j|=1$, we have $\frac{1}{\mathcal{D}_{4n}}\cdot\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)|\Delta\nu^n_{ij}|\to 0$.

From the above limits and the formulation of $\mathcal{D}_{4n}$ in equation (41), we deduce that

$$\frac{1}{\mathcal{D}_{4n}}\cdot\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)\Big(|\Delta b^n_{ij}|+|\Delta\nu^n_{ij}|\Big)\to 1.$$

This implies that there exists an index $j^*:|\mathcal{A}_j|>1$ (WLOG assume that $j^*=1$) such that

$$\frac{1}{\mathcal{D}_{4n}}\cdot\sum_{i\in\mathcal{A}_1}\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\Big(|\Delta b_{i1}^n|+|\Delta\nu_{i1}^n|\Big)\not\to 0.$$

Moreover, since

$$\frac{Z_{0,\ell_2,1}^n}{\mathcal{D}_{4n}}-\frac{Z_{1,\ell_2,1}^n}{\mathcal{D}_{4n}}=\frac{1}{\mathcal{D}_{4n}}\cdot\sum_{i\in\mathcal{A}_1}\sum_{\substack{\alpha_4+2\alpha_5=\ell_2,\\1\leq\alpha_4+\alpha_5\leq\bar{r}_1}}\frac{1}{\alpha_4!\alpha_5!2^{\alpha_5}}(\Delta b_{i1}^n)^{\alpha_4}(\Delta\nu_{i1}^n)^{\alpha_5}\to 0,$$

for any $1\leq\ell_2\leq\bar{r}_1$, we obtain that

$$\frac{\sum_{i\in\mathcal{A}_1}\sum_{\substack{\alpha_4+2\alpha_5=\ell_2,\\1\leq\alpha_4+\alpha_5\leq\bar{r}_1}}\frac{1}{\alpha_4!\alpha_5!2^{\alpha_5}}(\Delta b_{i1}^n)^{\alpha_4}(\Delta\nu_{i1}^n)^{\alpha_5}}{\sum_{i\in\mathcal{A}_1}\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\Big(|\Delta b_{i1}^n|+|\Delta\nu_{i1}^n|\Big)}\to 0. \tag{50}$$

Let us define $\overline{M}_n:=\max\{|\Delta b_{i1}^n|,|\Delta\nu_{i1}^n|^{1/2}:i\in\mathcal{A}_1\}$ and $\overline{\pi}_n:=\max_{i\in\mathcal{A}_1}\exp(\frac{\beta_{0i}^n}{\tau^n})$. Since the sequence $\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)/\overline{\pi}_n$ is bounded, it is possible to replace it by its subsequence that has a positive limit $q_{3i}^2:=\lim_{n\to\infty}\exp(\frac{\beta_{0i}^n}{\tau^n})/\overline{\pi}_n$. Thus, at least one among q_{3i}^2, for $i\in\mathcal{A}_1$, is equal to one.

In addition, we also define

$$(\Delta b_{i1}^n)/\overline{M}_n\to q_{4i},\quad(\Delta\nu_{i1}^n)/[2\overline{M}_n]\to q_{5i}.$$

It is worth noting that at least one among q_{4i} and q_{5i} for $i\in\mathcal{A}_1$ is equal to either 1 or -1. Subsequently, we divide both the numerator and the denominator of the ratio in equation (50) by $\overline{\pi}_n\overline{M}_n^{\ell_2}$, and then obtain the following system of polynomial equations:

$$\sum_{i\in\mathcal{A}_1}\sum_{\substack{\alpha_4+2\alpha_5=\ell,\\1\leq\alpha_4+\alpha_5\leq\bar{r}_1}}\frac{q_{3i}^2\,q_{4i}^{\alpha_4}\,q_{5i}^{\alpha_5}}{\alpha_4!\,\alpha_5!}=0,$$

for all $1\leq\ell_2\leq\bar{r}_1$. However, from the definition of $\bar{r}(|\mathcal{A}_1|)$, the above system does not have any non-trivial solutions, which contradicts to the fact that at least one among q_{4i} and q_{5i} for $i\in\mathcal{A}_1$ is non-zero. Therefore, not all the ratios $Z_{\ell_1,\ell_2,j}^n/\mathcal{D}_{4n}$ converge to zero as $n\to\infty$.

Step 3. In this step, we leverage the Fatou's lemma to show a result contradicting to that in Step 2. In particular, by the Fatou's lemma, we have

$$\lim_{n\to\infty}\frac{\mathbb{E}_X[V(g_{G_n}(\cdot|X),g_{G_*}(\cdot|X))]}{\mathcal{D}_{4n}}\geq\int\liminf_{n\to\infty}\frac{|g_{G_n}(Y|X)-g_{G_*}(Y|X)|}{2\mathcal{D}_{4n}}\mathrm{d}(X,Y).$$

Moreover, recall from the hypothesis in equation (40) that $\mathbb{E}_X[V(g_{G_n}(\cdot|X),g_{G_*}(\cdot|X))]/\mathcal{D}_{4n}\to 0$ as $n\to\infty$. Therefore, we deduce that

$$\frac{|g_{G_n}(Y|X)-g_{G_*}(Y|X)|}{\mathcal{D}_{4n}}\to 0,$$

for almost surely (X,Y). Since the term $\Big[\sum_{j=1}^{k_*}\exp\Big(\dfrac{\beta_{1j}^*X+\beta_{0j}^*}{\tau^*}\Big)\Big]$ is bounded, we also have that $\dfrac{Q_n}{\mathcal{D}_{4n}}\to 0$ as $n\to\infty$. Following from the results in equation (49), we have

$$\sum_{j=1}^{k_*}\sum_{\ell_1+\ell_2=0}^{2\bar{r}_j}\frac{Z_{\ell_1,\ell_2,j}^n}{\mathcal{D}_{4n}}\cdot X^{\ell_1}F^{(\ell_2)}(Y;X,\omega_j^*)-\sum_{j=1}^{k_*}\sum_{\ell=0}^{1+\mathbf{1}_{|\mathcal{A}_j|>1}}\frac{Z_{\ell,0,j}^n}{\mathcal{D}_{4n}}\cdot X^{\ell}H(Y;X,\beta_{1j}^*,\tau^*)\to 0. \tag{51}$$

Therefore, $Z_{\ell_1,\ell_2,j}^n/\mathcal{D}_{4n}$ must be bounded for any $j\in[k_*]$ and $0\leq\ell_1+\ell_2\leq 2\bar{r}_j$. Indeed, if at least one among them is not bounded, then the ratio $Z_{\ell_1,\ell_2,j}^n/\mathcal{D}_{4n}$ will go to infinity, implying that the left hand side of equation (51) does not go to zero, which is a contradiction. Thus, for each $j \in [k_*]$ and $0 \leq \ell_1 + \ell_2 \leq 2\bar{r}_j$, we can replace $Z^n_{\ell_1,\ell_2,j}$ by one of its subsequences such that the ratio $Z^n_{\ell_1,\ell_2,j}/\mathcal{D}_{4n}$ has a finite limit as $n \to \infty$. Let us denote $Z^n_{\ell_1,\ell_2,j}/\mathcal{D}_{4n} \to Z^*_{\ell_1,\ell_2,j}$, then it follows from the results in Step 2 that at least one among them is non-zero. Additionally, equation (51) indicates that

$$\sum_{j=1}^{k_*}\sum_{\ell_1+\ell_2=0}^{2\bar{r}_j} Z^*_{\ell_1,\ell_2,j} \cdot X^{\ell_1} F^{(\ell_2)}(Y;X,\omega^*_j) - \sum_{j=1}^{k_*}\sum_{\ell=0}^{1+\mathbf{1}_{|\mathcal{A}_j|>1}} Z^*_{\ell,0,j} \cdot X^{\ell} H(Y;X,\beta^*_{1j},\tau^*) = 0,$$

for almost surely (X, Y). Since the set

$$\mathcal{S} = \Big\{X^{\ell_1}F^{(\ell_2)}(Y;X,\omega^*_j),\ X^{\ell}H(Y;X,\beta^*_{1j},\tau^*) : j \in [k_*],\ 0 \leq \ell_1+\ell_2 \leq 2\bar{r}_j,\ 0 \leq \ell \leq 1 + \mathbf{1}_{|\mathcal{A}_j|>1}\Big\}$$

is linearly independent, we deduce that $Z^*_{\ell_1,\ell_2,j} = 0$ for any $j \in [k_*]$ and $0 \leq \ell_1 + \ell_2 \leq 2\bar{r}_j$, which contradicts the fact that at least one among them is different from zero. Hence, the proof is completed.

B.5. Proof of Theorem 3.3

In this proof, our main goal is to demonstrate the following inequality:

$$\inf_{G\in\mathcal{E}_{k_*}(\Theta)} \mathbb{E}_X[V(p_G(\cdot|X), p_{G_*}(\cdot|X))]/\mathcal{D}_5(G, G_*) > 0. \tag{52}$$

For that purpose, we separate the above inequality into local and global parts.

Local part: In this part, we aim to show that

$$\lim_{\varepsilon\to 0}\inf_{G\in\mathcal{E}_{k_*}(\Theta):\mathcal{D}_5(G,G_*)\leq\varepsilon} \mathbb{E}_X[V(p_G(\cdot|X), p_{G_*}(\cdot|X))]/\mathcal{D}_5(G, G_*) > 0. \tag{53}$$

Assume by contrary that the above claim does not hold true, then we can find a sequence of mixing measures $G_n = \sum_{i=1}^{k_*} \exp(\beta^n_{0i}/\tau^n)\delta_{(\beta^n_{1i},\tau^n,a^n_i,b^n_i,\nu^n_i)}$ in $\mathcal{E}_{k_*}(\Theta)$ that satisfies $\mathcal{D}_{5n} := \mathcal{D}_5(G_n, G_*) \to 0$ and

$$\mathbb{E}_X[V(p_{G_n}(\cdot|X), p_{G_*}(\cdot|X))]/\mathcal{D}_{5n} \to 0, \tag{54}$$

as $n \to \infty$. Recall that under the exact-specified settings, each Voronoi cell $\mathcal{A}^n_i = \mathcal{A}_i(G_n)$ has only one element. Therefore, we may assume without loss of generality (WLOG) that $\mathcal{A}^n_i = \{i\}$ for any $i \in [k_*]$. Thus, the loss function $\mathcal{D}_{5n}$ is reduced to

$$\mathcal{D}_{5n} := \sum_{i=1}^{k_*} \exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)\Big[\|\Delta\beta^n_{1i}\| + |\Delta\tau^n| + \|\Delta a^n_i\| + |\Delta b^n_i| + |\Delta\nu^n_i|\Big] + \sum_{i=1}^{k_*}\Big|\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big) - \exp\Big(\frac{\beta^*_{0j}}{\tau^*}\Big)\Big| \tag{55}$$

Since $\mathcal{D}_{5n} \to 0$, we get that $(\beta^n_{1i}, \tau^n, a^n_i, b^n_i, \nu^n_i) \to (\beta^*_{1i}, \tau^*, a^*_i, b^*_i, \nu^*_i)$ and $\exp(\beta^n_{0i}/\tau^n) \to \exp(\beta^*_{0i}/\tau^*)$ as $n \to \infty$. Now, we divide the proof of local part into three steps as follows:

Step 1. In this step, we decompose the quantity $Q_n := \Big[\sum_{i=1}^{k_*} \exp\Big(\frac{\sigma((\beta^*_{1i})^\top X) + \beta^*_{0i}}{\tau^*}\Big)\Big] \cdot [p_{G_n}(Y|X) - p_{G_*}(Y|X)]$ into a linear combination of linearly independent terms. Firstly, let $\bar{\sigma}(X, \beta_1) := \sigma(\beta_1^\top X)$, $F(Y;X,\beta_1,\tau,a,b,\nu) := \exp\Big(\frac{\bar{\sigma}(X,\beta_1)}{\tau}\Big)f(Y|a^\top X + b, \nu)$ and $H(Y;X,\beta_1,\tau) := \exp\Big(\frac{\bar{\sigma}(X,\beta_1)}{\tau}\Big)p_{G_n}(Y|X)$. Then, it can be verified that

$$\begin{aligned}
Q_n &= \sum_{i=1}^{k_*}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)\Big[F(Y;X,\beta^n_{1i},\tau^n,a^n_i,b^n_i,\nu^n_i) - F(Y;X,\beta^*_{1i},\tau^*,a^*_i,b^*_i,\nu^*_i)\Big]\\
&- \sum_{i=1}^{k_*}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)\Big[H(Y;X,\beta^n_{1i},\tau^n) - H(Y;X,\beta^*_{1i},\tau^*)\Big]\\
&+ \sum_{i=1}^{k_*}\Big[\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big) - \exp\Big(\frac{\beta^*_{0i}}{\tau^*}\Big)\Big]\cdot\exp\Big(\frac{\bar{\sigma}(X,\beta^*_{1i})}{\tau^*}\Big)f(Y|(a^*_i)^\top X + b^*_i, \nu^*_i)\\
&- \sum_{i=1}^{k_*}\Big[\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big) - \exp\Big(\frac{\beta^*_{0i}}{\tau^*}\Big)\Big]\cdot\exp\Big(\frac{\bar{\sigma}(X,\beta^*_{1i})}{\tau^*}\Big)p_{G_n}(Y|X)\\
&:= A_n - B_n + E_{n,1} - E_{n,2}.
\end{aligned}$$

Next, by means of the first-order Taylor expansion, we get that

$$A_n = \sum_{i=1}^{k_*} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) \sum_{|\alpha|=1} \frac{1}{\alpha!} (\Delta\beta_{1i}^n)^{\alpha_1} (\Delta\tau^n)^{\alpha_2} (\Delta a_i^n)^{\alpha_3} (\Delta b_i^n)^{\alpha_4} (\Delta\nu_i^n)^{\alpha_5} \\ \times \frac{\partial F}{\partial\beta_1^{\alpha_1}\,\partial\tau^{\alpha_2}\,\partial a^{\alpha_3}\,\partial b^{\alpha_4}\,\partial\nu^{\alpha_5}}(Y; X, \beta_{1i}^*, \tau^*, a_i^*, b_i^*, \nu_i^*) + R_1(X, Y),$$

where $R_1(X,Y)$ is a Taylor remainder such that $R_1(X,Y)/\mathcal{D}_{5n} \to 0$ as $n \to \infty$. Let us denote

$$F^{(\eta)}(Y; X, \omega_i^*) := \exp\Big(\frac{\bar{\sigma}(X, \beta_{1i}^*)}{\tau^*}\Big) \frac{\partial^\eta f}{\partial h_1^\eta}(Y | (a_i^*)^\top X + b_i^*, \nu_i^*),$$

for any $\eta \in \mathbb{N}$, where $\omega_i^* := (\beta_{1i}^*, \tau^*, a_i^*, b_i^*, \nu_i^*)$. Then, the first derivatives of function F w.r.t its parameters are given by

$$\begin{aligned} \frac{\partial F}{\partial\beta_1^{(u)}}(Y; X, \omega_i^*) &= \frac{1}{\tau^*} \cdot \frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X, \beta_{1i}^*) F(Y; X, \omega_i^*), \\ \frac{\partial F}{\partial\tau}(Y; X, \omega_i^*) &= -\frac{\bar{\sigma}(X, \beta_{1i}^*)}{(\tau^*)^2} F(Y; X, \omega_i^*), \\ \frac{\partial F}{\partial a^{(u)}}(Y; X, \omega_i^*) &= X^{(u)} F^{(1)}(Y; X, \omega_i^*), \\ \frac{\partial F}{\partial b}(Y; X, \omega_i^*) &= F^{(1)}(Y; X, \omega_i^*), \\ \frac{\partial F}{\partial\nu}(Y; X, \omega_i^*) &= \frac{1}{2} F^{(2)}(Y; X, \omega_i^*), \end{aligned} \tag{56}$$

for any $u \in [d]$ and $i \in [k_*]$. Then, the terms A_n and $E_{n,1}$ can be represented as

$$A_n = \sum_{i=1}^{k_*} \sum_{\eta=0}^{2} C_{n,\eta,i}(X) F^{(\eta)}(Y; X, \omega_i^*) + R_1(X, Y), \tag{57}$$

where we define for any $i \in [k_*]$ and $X \in \mathcal{X}$ that

$$\begin{aligned} C_{n,0,i}(X) &:= \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) \Big[\sum_{u=1}^{d} \frac{(\Delta\beta_{1i}^n)^{(u)}}{\tau^*} \cdot \frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X, \beta_{1i}^*) - \frac{(\Delta\tau^n)}{(\tau^*)^2} \cdot \bar{\sigma}(X, \beta_{1i}^*)\Big], \\ C_{n,1,i}(X) &:= \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) \Big[\sum_{u=1}^{d} (\Delta a_i^n)^{(u)} X^{(u)} + (\Delta b_i^n)\Big], \\ C_{n,2,i}(X) &:= \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) \cdot \frac{(\Delta\nu_i^n)}{2}. \end{aligned} \tag{58}$$

Thus, we can view the terms $[A_n - R_1(X,Y)]/\mathcal{D}_{5n}$ and $E_{n,1}/\mathcal{D}_{5n}$ as a linear combination of elements from the set $\mathcal{F} := \cup_{i=1}^{k_*} \cup_{\eta=0}^{2} \mathcal{F}_{i,\eta}$ in which

$$\begin{aligned} \mathcal{F}_{i,0} &:= \Big\{\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X, \beta_{1i}^*) F(Y; X, \omega_i^*) : u \in [d]\Big\} \cup \Big\{\bar{\sigma}(X, \beta_{1i}^*) F(Y; X, \omega_i^*)\Big\} \cup \Big\{F(Y; X, \omega_i^*)\Big\}, \\ \mathcal{F}_{i,1} &:= \Big\{X^{(u)} \cdot F^{(1)}(Y; X, \omega_i^*) : u \in [d]\Big\} \cup \Big\{F^{(1)}(Y; X, \omega_i^*)\Big\}, \\ \mathcal{F}_{i,2} &:= \Big\{F^{(2)}(Y; X, \omega_i^*)\Big\}, \end{aligned}$$

where $\omega_i^* := (\beta_{1i}^*, \tau^*, a_i^*, b_i^*, \nu_i^*)$, for any $i \in [k_*]$. Subsequently, we apply the first-order Taylor expansion to B_n as follows:

$$B_n : = \sum_{i=1}^{k_*} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) \sum_{|\gamma|=1} \frac{1}{\gamma!} (\Delta\beta_{1i}^n)^{\gamma_1} (\Delta\tau^n)^{\gamma_2} \cdot \frac{\partial H}{\partial\beta_1^{\gamma_1}\,\partial\tau^{\gamma_2}}(Y; X, \beta_{1i}^*, \tau^*) + R_2(X, Y),$$

where $R_2(X,Y)$ is a Taylor remainder such that $R_2(X,Y)/\mathcal{D}_{5n} \to 0$ as $n \to \infty$. Then, the term B_n can be represented as

$$B_n = \sum_{i=1}^{k_*} C_{n,0,i}(X) H(Y; X, \beta_{1i}^*, \tau^*) + R_2(X,Y), \tag{59}$$

where $C_{n,0,i}(X)$ is defined in equation (58). Therefore, the terms $[B_n - R_2(X,Y)]/\mathcal{D}_{5n}$ and $E_{n,2}/\mathcal{D}_{5n}$ can be treated as a linear combination of elements from the set $\mathcal{H} := \cup_{i=1}^{k_*} \mathcal{H}_i$, where we define for $i \in [k_*]$ that

$$\mathcal{H}_i := \Big\{ \frac{\partial \bar{\sigma}}{\partial \beta_1}(X, \beta_{1i}^*) H(Y; X, \beta_{1i}^*, \tau^*) : u \in [d] \Big\} \cup \Big\{ \bar{\sigma}(X, \beta_{1i}^*) H(Y; X, \beta_{1i}^*, \tau^*) \Big\} \cup \Big\{ H(Y; X, \beta_{1i}^*, \tau^*) \Big\}.$$

Step 2. In this step, we prove by contradiction that at least one among the coefficients in the representations of $[A_n - R_1(X,Y)]/\mathcal{D}_{5n}$, $[B_n - R_2(X,Y)]/\mathcal{D}_{5n}$, $E_{n,1}/\mathcal{D}_{5n}$ and $E_{n,2}/\mathcal{D}_{5n}$ does not converge to zero when $n \to \infty$. Assume by contrary that all of them go to 0 as $n \to \infty$. In the term $E_{n,1}/\mathcal{D}_{5n}$, by taking the summation of the absolute values of the coefficients of $F(Y; X, \omega_i^*)$, we get that

$$\frac{1}{\mathcal{D}_{5n}} \cdot \sum_{i=1}^{k_*} \Big| \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) - \exp\Big(\frac{\beta_{0i}^*}{\tau^*}\Big) \Big| \to 0. \tag{60}$$

Next, by taking the summation of the absolute values of the coefficients associated with

- $\dfrac{\partial \bar{\sigma}}{\partial \beta_1^{(u)}}(X, \beta_{1i}^*) F(Y; X, \omega_i^*)$ in $\mathcal{F}_0$: we have that $\frac{1}{\mathcal{D}_{5n}} \cdot \sum_{i=1}^{k_*} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) \|\Delta \beta_{1i}^n\|_1 \to 0$;
- $\bar{\sigma}(X, \beta_{1i}^*) F(Y; X, \omega_i^*)$ in $\mathcal{F}_0$: we have that $\frac{1}{\mathcal{D}_{5n}} \cdot \sum_{i=1}^{k_*} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) |\Delta \tau^n| \to 0$;
- $X^{(u)} F^{(1)}(Y; X, \omega_i^*)$ in $\mathcal{F}_1$: we have that $\frac{1}{\mathcal{D}_{5n}} \cdot \sum_{i=1}^{k_*} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) \|\Delta a_i^n\|_1 \to 0$;
- $F^{(1)}(Y; X, \omega_i^*)$ in $\mathcal{F}_1$: we have that $\frac{1}{\mathcal{D}_{5n}} \cdot \sum_{i=1}^{k_*} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) |\Delta b_i^n| \to 0$;
- $F^{(2)}(Y; X, \omega_i^*)$ in $\mathcal{F}_2$: we have that $\frac{1}{\mathcal{D}_{5n}} \cdot \sum_{i=1}^{k_*} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) |\Delta \nu_i^n| \to 0$;

Due to the topological equivalence between ℓ_1-norm and ℓ_2-norm, it follows that

$$\frac{1}{\mathcal{D}_{5n}} \cdot \sum_{i=1}^{k_*} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) \Big(\|\Delta \beta_{1i}^n\| + |\Delta \tau^n| + \|\Delta a_i^n\| + |\Delta b_i^n| + |\Delta \nu_i^n| \Big) \to 0. \tag{61}$$

Putting the results in equations (60) and (61) and the formulation of the loss $\mathcal{D}_{5n}$ in equation (55) together, we deduce that $1 = \mathcal{D}_{5n}/\mathcal{D}_{5n} \to 0$ as $n \to \infty$, which is a contradiction. Consequently, not all the coefficients in the representations of $[A_n - R_1(X,Y)]/\mathcal{D}_{5n}$, $[B_n - R_2(X,Y)]/\mathcal{D}_{5n}$, $E_{n,1}/\mathcal{D}_{5n}$ and $E_{n,2}/\mathcal{D}_{5n}$ converge to zero when $n \to \infty$.

Step 3. In this step, we utilize the Fatou's lemma to demonstrate a result contradicting to that in Step 2. In particular, let us denote m_n as the maximum of the absolute values of the coefficients in the representations of $[A_n - R_1(X,Y)]/\mathcal{D}_{5n}$, $[B_n - R_2(X,Y)]/\mathcal{D}_{5n}$, $E_{n,1}/\mathcal{D}_{5n}$ and $E_{n,2}/\mathcal{D}_{5n}$. From the conclusion of Step 2, we know that $1/m_n \not\to \infty$. Next, we denote

$$\begin{aligned}
\frac{1}{m_n \mathcal{D}_{5n}} \cdot \Big[\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) - \exp\Big(\frac{\beta_{0i}^*}{\tau^*}\Big) \Big] \to \phi_{0,i}, &\qquad \frac{1}{m_n \mathcal{D}_{5n}} \cdot \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) (\Delta \beta_{1i}^n)^{(u)} \to \phi_{1,i}^{(u)}, \\
\frac{1}{m_n \mathcal{D}_{5n}} \cdot \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) (\Delta \tau^n) \to \phi_{2,i}, &\qquad \frac{1}{m_n \mathcal{D}_{5n}} \cdot \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) (\Delta a_i^n)^{(u)} \to \phi_{3,i}^{(u)}, \\
\frac{1}{m_n \mathcal{D}_{5n}} \cdot \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) (\Delta b_i^n) \to \phi_{4,i}, &\qquad \frac{1}{m_n \mathcal{D}_{5n}} \cdot \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) (\Delta \nu_i^n) \to \phi_{5,i},
\end{aligned}$$

as $n \to \infty$ for any $u \in [d]$ and $i \in [k_*]$. Note that at least one among the terms $\phi_{0,i}, \phi_{1,i}^{(u)}, \phi_{2,i}, \phi_{3,i}^{(u)}, \phi_{4,i}$ and $\phi_{5,i}$ is different from zero. By means of the Fatou's lemma, we have that

$$\lim_{n\to\infty} \frac{\mathbb{E}_X[V(p_{G_n}(\cdot|X), p_{G_*}(\cdot|X))]}{m_n \mathcal{D}_{5n}} \geq \int \liminf_{n\to\infty} \frac{|p_{G_n}(Y|X) - p_{G_*}(Y|X)|}{2m_n \mathcal{D}_{5n}} \mathrm{d}(X,Y).$$

Recall from equation (54) that the limit the left hand side is equal to zero, which implies that $\frac{|p_{G_n}(Y|X) - p_{G_*}(Y|X)|}{m_n \mathcal{D}_{5n}} \to 0$. as $n \to \infty$ for almost surely (X,Y). Thus, we also have that $\frac{Q_n}{m_n \mathcal{D}_{5n}} \to 0$ as $n \to \infty$. On the other hand, we have

$$\frac{Q_n}{m_n \mathcal{D}_{5n}} \to \sum_{i=1}^{k_*} \sum_{\eta=0}^{2} C^*_{\eta,i}(X) \cdot F^{(\eta)}(Y; X, \omega^*_i) - \sum_{i=1}^{k_*} C^*_{0,i}(X) \cdot H(Y; X, \beta^*_{1i}, \tau^*),$$

for almost surely (X,Y), where we define

$$C^*_{0,i}(X) := \phi_{0,i} + \sum_{u=1}^{d} \frac{\phi_{1,i}^{(u)}}{\tau^*} \cdot \frac{\partial \bar{\sigma}}{\partial \beta_1^{(u)}}(X, \beta^*_{1i}) - \phi_{2,i} \cdot \frac{\bar{\sigma}(X, \beta^*_{1i})}{(\tau^*)^2}, \tag{62}$$

$$C^*_{1,i}(X) := \sum_{u=1}^{d} \phi_{3,i}^{(u)} \cdot X^{(u)} + \phi_{4,i}, \tag{63}$$

$$C^*_{2,i}(X) := \frac{1}{2} \phi_{5,i}, \tag{64}$$

for any $i \in [k_*]$. As a result, we achieve that

$$\sum_{i=1}^{k_*} \sum_{\eta=0}^{2} C^*_{\eta,i}(X) \cdot F^{(\eta)}(Y; X, \omega^*_i) - \sum_{i=1}^{k_*} C^*_{0,i}(X) \cdot H(Y; X, \beta^*_{1i}, \tau^*) = 0,$$

for almost surely (X,Y). Since the following set is linearly independent w.r.t Y:

$$\left\{ F^{(\eta)}(Y; X, \omega^*_i),\ H(Y; X, \beta^*_{1i}, \tau^*) : 0 \leq \eta \leq 2,\ i \in [k_*] \right\},$$

it leads to $C^*_{\eta,i}(X) = 0$ for any $0 \leq \eta \leq 2$ and $i \in [k_*]$ for almost surely X.

- When $C^*_{0,i}(X) = 0$ for almost surely X: as the function σ satisfies the conditions in Definition 3.2, i.e. the set

  $$\left\{ \frac{\partial \bar{\sigma}}{\partial \beta_1^{(u)}}(X, \beta^*_{1i}),\ \bar{\sigma}(X, \beta^*_{1i}),\ 1 : 1 \leq u \leq d \right\}$$

  is linearly independent w.r.t X, it follows from equation (62) that $\phi_{0,i} = \phi_{1,i}^{(u)} = \phi_{2,i} = 0$ for any $u \in [d]$ and $i \in [k_*]$.
- When $C^*_{1,i}(X) = 0$ for almost surely X: since the set $\{X^{(u)}, 1 : u \in [d]\}$ is linearly independent w.r.t X, equation (63) indicates that $\phi_{3,i}^{(u)} = \phi_{4,i} = 0$ for any $u \in [d]$ and $i \in [k_*]$.
- When $C^*_{2,i}(X) = 0$ for almost surely X: it can be seen from equation (64) that $\phi_{5,i} = 0$ for any $i \in [k_*]$.

However, the above results contradict the fact that at least one among the terms $\phi_{0,i}, \phi_{1,i}^{(u)}, \phi_{2,i}, \phi_{3,i}^{(u)}, \phi_{4,i}$ and $\phi_{5,i}$ is non-zero. Hence, we reach the conclusion of the local part in equation (53), which means that there exists a constant $\varepsilon' > 0$ such that

$$\inf_{G \in \mathcal{E}_{k_*}(\Theta): \mathcal{D}_5(G,G_*) \leq \varepsilon'} \mathbb{E}_X[V(p_G(\cdot|X), p_{G_*}(\cdot|X))] / \mathcal{D}_5(G,G_*) > 0.$$

Global part. As a consequence, it suffices to demonstrate the following inequality:

$$\inf_{G \in \mathcal{E}_{k_*}(\Theta): \mathcal{D}_5(G,G_*) > \varepsilon'} \mathbb{E}_X[V(p_G(\cdot|X), p_{G_*}(\cdot|X))] / \mathcal{D}_5(G,G_*) > 0. \tag{65}$$

Assume by contrary that the above claim does not hold true, then we can seek a sequence of mixing measures $G'_n \in \mathcal{E}_{k_*}(\Omega)$ such that $\mathcal{D}_5(G'_n, G_*) > \varepsilon'$ and

$$\lim_{n\to\infty} \frac{\mathbb{E}_X[V(p_{G'_n}(\cdot|X), p_{G_*}(\cdot|X))]}{\mathcal{D}_5(G'_n, G_*)} = 0,$$

which directly implies that $\mathbb{E}_X[V(p_{G'_n}(\cdot|X), p_{G_*}(\cdot|X))] \to 0$ as $n \to \infty$. Recall that Θ is a compact set, therefore, we can replace the sequence G'_n by one of its subsequences that converges to a mixing measure $G' \in \mathcal{E}_{k_*}(\Omega)$. Since $\mathcal{D}_5(G'_n, G_*) > \varepsilon'$, this result induces that $\mathcal{D}_5(G', G_*) > \varepsilon'$.

Next, by invoking the Fatou's lemma, it follows that

$$0 = \lim_{n\to\infty} \mathbb{E}_X[2V(p_{G'_n}(\cdot|X), p_{G_*}(\cdot|X))] \geq \int \liminf_{n\to\infty} \Big| p_{G'_n}(Y|X) - p_{G_*}(Y|X) \Big| \,\mathrm{d}(X, Y).$$

Thus, we get that $p_{G'}(Y|X) = p_{G_*}(Y|X)$ for almost surely (X, Y). From Proposition C.2, we know that the model (12) is identifiable, which indicates that $G' \equiv G_*$. As a consequence, we have that $\mathcal{D}_5(G', G_*) = 0$, contradicting the fact that $\mathcal{D}_5(G', G_*) > \varepsilon' > 0$.

Hence, the proof is completed.

B.6. Proof of Theorem 3.5

Similar to the proof of Theorem 3.5 in Appendix B.5, we aim to prove the following inequality:

$$\inf_{G\in\mathcal{G}_k(\Theta)} \mathbb{E}_X[V(p_G(\cdot|X), p_{G_*}(\cdot|X))]/\mathcal{D}_6(G, G_*) > 0. \tag{66}$$

Moreover, we also divide the above inequality into local and global parts. Since the global part can be argued in the same fashion as in Appendix B.5, we will demonstrate only the local part, that is

$$\lim_{\varepsilon\to 0} \inf_{G\in\mathcal{G}_k(\Theta):\mathcal{D}_6(G,G_*)\leq\varepsilon} \mathbb{E}_X[V(p_G(\cdot|X), p_{G_*}(\cdot|X))]/\mathcal{D}_6(G, G_*) > 0. \tag{67}$$

Assume by contrary that the above claim does not hold true, then we can find a sequence of mixing measures $G_n = \sum_{i=1}^{k_*} \exp(\beta_{0i}^n/\tau^n)\delta_{(\beta_{1i}^n, \tau^n, a_i^n, b_i^n, \nu_i^n)}$ in $\mathcal{G}_k(\Theta)$ that satisfies $\mathcal{D}_{6n} := \mathcal{D}_6(G_n, G_*) \to 0$ and

$$\mathbb{E}_X[V(p_{G_n}(\cdot|X), p_{G_*}(\cdot|X))]/\mathcal{D}_{6n} \to 0, \tag{68}$$

as $n \to \infty$. Let us denote $\mathcal{A}_j^n = \mathcal{A}_j(G_n)$, then the loss function $\mathcal{D}_{6n}$ is reduced to

$$\begin{aligned}\mathcal{D}_{6n} := &\sum_{j:|\mathcal{A}_j|>1} \sum_{i\in\mathcal{A}_j} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\Big[\|\Delta\beta_{1ij}^n\|^2 + |\Delta\tau^n|^2 + \|\Delta a_{ij}^n\|^2 + |\Delta b_{ij}^n|^{\bar{r}_j} + |\Delta\nu_{ij}^n|^{\bar{r}_j/2}\Big] \\ &+ \sum_{j:|\mathcal{A}_j|=1} \sum_{i\in\mathcal{A}_j} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\Big[\|\Delta\beta_{1ij}^n\| + |\Delta\tau^n| + \|\Delta a_{ij}^n\| + |\Delta b_{ij}^n| + |\Delta\nu_{ij}^n|\Big] + \sum_{j=1}^{k_*}\Big|\sum_{i\in\mathcal{A}_j} \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big) - \exp\Big(\frac{\beta_{0j}^*}{\tau^*}\Big)\Big|\end{aligned} \tag{69}$$

As $\mathcal{D}_{6n} \to 0$, we get that $(\beta_{1i}^n, \tau^n, a_i^n, b_i^n, \nu_i^n) \to (\beta_{1j}^*, \tau^*, a_j^*, b_j^*, \nu_j^*)$ and $\sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n/\tau^n) \to \exp(\beta_{0j}^*/\tau^*)$ as $n \to \infty$ for any $i \in \mathcal{A}_j$ and $j \in [k_*]$. Now, we divide the proof of local part into three steps as follows:

Step 1. In this step, we decompose the quantity $Q_n := \Big[\sum_{j=1}^{k_*} \exp\Big(\dfrac{\sigma((\beta_{1j}^*)^\top X) + \beta_{0j}^*}{\tau}\Big)\Big] \cdot [p_{G_n}(Y|X) - p_{G_*}(Y|X)]$ into a linear combination of linearly independent terms. Firstly, let $\bar{\sigma}(X, w) := \sigma(w^\top X)$, $F(Y; X, \beta_1, \tau, a, b, \nu) :=$

$\exp\left(\frac{\bar{\sigma}(X,\beta_1)}{\tau}\right)f(Y|a^{\top}X+b,\nu)$ and $H(Y;X,\beta_1,\tau):=\exp\left(\frac{\bar{\sigma}(X,\beta_1)}{\tau}\right)p_{G_n}(Y|X)$. Then, it can be verified that

$$\begin{aligned}
Q_n &= \sum_{j=1}^{k_*}\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\Big[F(Y;X,\beta_{1i}^n,\tau^n,a_i^n,b_i^n,\nu_i^n)-F(Y;X,\beta_{1j}^*,\tau^*,a_j^*,b_j^*,\nu_j^*)\Big]\\
&\quad-\sum_{j=1}^{k_*}\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\Big[H(Y;X,\beta_{1i}^n,\tau^n)-H(Y;X,\beta_{1j}^*,\tau^*)\Big]\\
&\quad+\sum_{j=1}^{k_*}\Big[\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)-\exp\left(\frac{\beta_{0i}^*}{\tau^*}\right)\Big]\cdot\exp(\bar{\sigma}(X,\beta_{1j}^*))f(Y|(a_j^*)^{\top}X+b_j^*,\nu_j^*)\\
&\quad-\sum_{j=1}^{k_*}\Big[\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)-\exp\left(\frac{\beta_{0i}^*}{\tau^*}\right)\Big]\cdot\exp(\bar{\sigma}(X,\beta_{1j}^*))p_{G_n}(Y|X)\\
&:=A_n-B_n+E_{n,1}-E_{n,2}.
\end{aligned}\tag{70}$$

Next, we continue to separate A_n into two terms as follows:

$$\begin{aligned}
A_n &:= \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\Big[F(Y;X,\beta_{1i}^n,\tau^n,a_i^n,b_i^n,\nu_i^n)-F(Y;X,\beta_{1j}^*,\tau^*,a_j^*,b_j^*,\nu_j^*)\Big]\\
&\quad+\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\Big[F(Y;X,\beta_{1i}^n,\tau^n,a_i^n,b_i^n,\nu_i^n)-F(Y;X,\beta_{1j}^*,\tau^*,a_j^*,b_j^*,\nu_j^*)\Big]\\
&:=A_{n,1}+A_{n,2}.
\end{aligned}$$

Let us denote

$$F^{(\eta)}(Y;X,\beta_{1j}^*,\tau^*,a_j^*,b_j^*,\nu_j^*):=\exp\left(\frac{\bar{\sigma}(X,\beta_{1j}^*)}{\tau^*}\right)\frac{\partial^{\eta}f}{\partial h_1^{\eta}}(Y|(a_j^*)^{\top}X+b_j^*,\nu_j^*),$$

for any $\eta\in\mathbb{N}$. Then, by applying the first-order Taylor expansion as in equation (57), the term $A_{n,1}$ can be decomposed as

$$\begin{align}
A_{n,1} &= \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\sum_{|\alpha|=1}\frac{1}{\alpha!}(\Delta\beta_{1ij}^n)^{\alpha_1}(\Delta\tau^n)^{\alpha_2}(\Delta a_{ij}^n)^{\alpha_3}(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5} \nonumber\\
&\qquad\times\frac{\partial F}{\partial\beta_1^{\alpha_1}\,\partial\tau^{\alpha_2}\,\partial a^{\alpha_3}\,\partial b^{\alpha_4}\,\partial\nu^{\alpha_5}}(Y;X,\beta_{1j}^*,\tau^*,a_j^*,b_j^*,\nu_j^*)+R_1(X,Y) \nonumber\\
&=\sum_{j:|\mathcal{A}_j|=1}\sum_{\eta=0}^{2}C_{n,\eta,j}(X)F^{(\eta)}(Y;X,\beta_{1j}^*,\tau^*,a_j^*,b_j^*,\nu_j^*)+R_1(X,Y), \tag{71}\\
& \tag{72}
\end{align}$$

where $R_1(X,Y)$ is a Taylor remainder such that $R_1(X,Y)/\mathcal{D}_{6n}\to 0$ as $n\to\infty$ and

$$\begin{aligned}
C_{n,0,j}(X)&:=\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\Big[\frac{\Delta\beta_{1ij}^n}{\tau^*}\cdot\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta_{1j}^*)-\frac{\Delta\tau^n}{(\tau^*)^2}\bar{\sigma}(X,\beta_{1j}^*)\Big],\\
C_{n,1,j}(X)&:=\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\Big[\Delta a_{ij}^n\cdot X^{(u)}+\Delta b_{ij}^n\Big],\\
C_{n,2,j}(X)&:=\sum_{i\in\mathcal{A}_j}\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)\cdot\frac{\Delta\nu_{ij}^n}{2},
\end{aligned}\tag{73}$$

for any $j\in[k_*]:|\mathcal{A}_j|=1$. Meanwhile, for each $j:|\mathcal{A}_j|>1$, by means of the Taylor expansion of order $\bar{r}_j$, we can rewrite

$A_{n,2}$ as

$$\begin{aligned}A_{n,2} &= \sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)\sum_{|\alpha|=1}^{\bar{r}_j}\frac{1}{\alpha!}(\Delta\beta^n_{1ij})^{\alpha_1}(\Delta\tau^n)^{\alpha_2}(\Delta a^n_{ij})^{\alpha_3}(\Delta b^n_{ij})^{\alpha_4}(\Delta\nu^n_{ij})^{\alpha_5}\\
&\qquad\times\frac{\partial^{|\alpha_1|+\alpha_2+|\alpha_3|+\alpha_4+\alpha_5}F}{\partial\beta_1^{\alpha_1}\,\partial\tau^{\alpha_2}\,\partial a^{\alpha_3}\,\partial b^{\alpha_4}\,\partial\nu^{\alpha_5}}(Y;X,\beta^*_{1j},\tau^*,a^*_j,b^*_j,\nu^*_j)+R_2(X,Y),\\
&= \sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)\sum_{|\alpha|=1}^{\bar{r}_j}\frac{1}{\alpha!}(\Delta\beta^n_{1ij})^{\alpha_1}(\Delta\tau^n)^{\alpha_2}(\Delta a^n_{ij})^{\alpha_3}(\Delta b^n_{ij})^{\alpha_4}(\Delta\nu^n_{ij})^{\alpha_5}\\
&\qquad\times\frac{X^{\alpha_3}}{2^{\alpha_5}}\cdot\frac{\partial^{|\alpha_1|+\alpha_2}L}{\partial\beta_1^{\alpha_1}\partial\tau^{\alpha_2}}(X,\beta^*_{1j},\tau^*)\cdot\frac{\partial^{|\alpha_3|+\alpha_4+2\alpha_5}f}{\partial h_1^{|\alpha_3|+\alpha_4+2\alpha_5}}(Y;X,\beta^*_{1j},\tau^*,a^*_j,b^*_j,\nu^*_j)+R_2(X,Y).\end{aligned}$$

where $L(X,\beta_1,\tau):=\exp\Big(\frac{\bar{\sigma}(X,\beta_1)}{\tau}\Big)$ and $R_2(X,Y)$ is a Taylor remainder such that $R_2(X,Y)/\mathcal{D}_{6n}\to 0$ as $n\to\infty$. For each $j:|\mathcal{A}_j|>1$, by letting $|\alpha_1|+\alpha_2=s$, where $0\leq s\leq\bar{r}_j$, then we have $1-s\leq|\alpha_3|+\alpha_4+\alpha_5\leq\bar{r}_j-s$. Next, we denote $\alpha_4+2\alpha_5=\ell$, where $0\leq\ell\leq 2(\bar{r}_j-s-|\alpha_3|)$. Then, $A_{n,2}$ can be represented as

$$A_{n,2}=\sum_{j:|\mathcal{A}_j|>1}\sum_{s=0}^{\bar{r}_j}\sum_{|\alpha_3|=0}^{\bar{r}_j-s}\sum_{\ell=0}^{2(\bar{r}_j-s-|\alpha_3|)}T_{n,s,\alpha_3,\ell,j}(X)\cdot X^{\alpha_3}F^{(|\alpha_3|+\ell)}(Y;X,\beta^*_{1j},\tau^*,a^*_j,b^*_j,\nu^*_j)+R_2(X,Y),\tag{74}$$

where

$$\begin{aligned}T_{n,s,\alpha_3,\ell,j}(X) &:= \sum_{\substack{\alpha_4+2\alpha_5=\ell,\\ 1-s\leq\alpha_4+\alpha_5\leq\bar{r}_j-s}}\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)\frac{1}{2^{\alpha_5}\alpha_3!\alpha_4!\alpha_5!}(\Delta a^n_{ij})^{\alpha_3}(\Delta b^n_{ij})^{\alpha_4}(\Delta\nu^n_{ij})^{\alpha_5}\\
&\qquad\times\Big[\sum_{|\alpha_1|+\alpha_2=s}\frac{1}{\alpha_1!\alpha_2!}(\Delta\beta^n_{1ij})^{\alpha_1}(\Delta\tau^n)^{\alpha_2}\cdot\frac{\partial^{|\alpha_1|+\alpha_2}L}{\partial\beta_1^{\alpha_1}\partial\tau^{\alpha_2}}(X,\beta^*_{1j},\tau^*)\cdot\frac{1}{L(X,\beta^*_{1j},\tau^*)}\Big].\end{aligned}$$

Now, we provide the explicit formulations of $T_{n,s,\alpha_3,\ell,j}(X)$ for $s\in\{0,1,2\}$. First, the term $T_{n,0,\alpha_3,\ell,j}(X)$ is given by:

$$T_{n,0,\alpha_3,\ell,j}(X):=\sum_{\substack{\alpha_4+2\alpha_5=\ell,\\ 1\leq\alpha_4+\alpha_5\leq\bar{r}_j}}\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)\frac{1}{2^{\alpha_5}\alpha_3!\alpha_4!\alpha_5!}(\Delta a^n_{ij})^{\alpha_3}(\Delta b^n_{ij})^{\alpha_4}(\Delta\nu^n_{ij})^{\alpha_5}$$

For the term $T_{n,1,\alpha_3,\ell,j}(X)$, let us derive the first derivatives of function L w.r.t its parameters as

$$\begin{aligned}\frac{\partial L}{\partial\beta_1^{(u)}}(X,\beta^*_{1j},\tau^*)&=\frac{1}{\tau^*}\cdot\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta^*_{1j})L(X,\beta^*_{1j},\tau^*),\\
\frac{\partial L}{\partial\tau}(X,\beta^*_{1j},\tau^*)&=-\frac{\bar{\sigma}(X,\beta^*_{1j})}{(\tau^*)^2}L(X,\beta^*_{1j},\tau^*).\end{aligned}$$

Thus, the formulation of $T_{n,1,\alpha_3,\ell,j}(X)$ reads as

$$\begin{aligned}T_{n,1,\alpha_3,\ell,j}(X) &= \sum_{\substack{\alpha_4+2\alpha_5=\ell,\\ 1-1\leq\alpha_4+\alpha_5\leq\bar{r}_j-1}}\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)\frac{1}{2^{\alpha_5}\alpha_3!\alpha_4!\alpha_5!}(\Delta a^n_{ij})^{\alpha_3}(\Delta b^n_{ij})^{\alpha_4}(\Delta\nu^n_{ij})^{\alpha_5}\\
&\qquad\times\Big[\sum_{u=1}^{d}\frac{(\Delta\beta^n_{1ij})^{(u)}}{\tau^*}\cdot\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta^*_{1j})-\frac{(\Delta\tau^n)}{(\tau^*)^2}\cdot\bar{\sigma}(X,\beta^*_{1j})\Big].\end{aligned}$$

Similarly, for the term $T_{n,2,\alpha_3,\ell,j}$, we derive the second derivatives of function L w.r.t its parameters as

$$\begin{aligned}
\frac{\partial^2 L}{\partial\beta_1^{(u)}\partial\beta_1^{(v)}}(X,\beta_{1j}^*,\tau^*) &= \frac{1}{\tau^*}\cdot\frac{\partial^2\sigma}{\partial\beta_1^{(u)}\partial\beta_1^{(v)}}((\beta_{1j}^*)^\top X)L(X,\beta_{1j}^*,\tau^*)\\
&\qquad+\frac{1}{(\tau^*)^2}\cdot\Big[\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta_{1j}^*)\Big]\Big[\frac{\partial\sigma}{\partial\beta_1^{(v)}}((\beta_{1j}^*)^\top X)\Big]L(X,\beta_{1j}^*,\tau^*),\\
\frac{\partial^2 L}{\partial\tau^2}(X,\beta_{1j}^*,\tau^*) &= \frac{2}{(\tau^*)^3}\cdot\bar{\sigma}(X,\beta_{1j}^*)L(X,\beta_{1j}^*,\tau^*)+\frac{1}{(\tau^*)^4}\cdot\bar{\sigma}^2(X,\beta_{1j}^*)L(X,\beta_{1j}^*,\tau^*),\\
\frac{\partial^2 L}{\partial\beta_1^{(u)}\partial\tau}(X,\beta_{1j}^*,\tau^*) &= -\frac{1}{(\tau^*)^2}\cdot\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta_{1j}^*)L(X,\beta_{1j}^*,\tau^*)\\
&\qquad-\frac{1}{(\tau^*)^3}\cdot\bar{\sigma}(X,\beta_{1j}^*)\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta_{1j}^*)L(X,\beta_{1j}^*,\tau^*).
\end{aligned}$$

Then, $T_{n,2,\alpha_3,\ell,j}$ can be written as

$$\begin{aligned}
T_{n,2,\alpha_3,\ell,j}(X) := &\sum_{\substack{\alpha_4+2\alpha_5=\ell,\\ 0\leq\alpha_4+\alpha_5\leq\bar{r}_j-2}}\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\frac{1}{2^{\alpha_5}\alpha_3!\alpha_4!\alpha_5!}(\Delta a_{ij}^n)^{\alpha_3}(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5}\\
&\times\Bigg\{\sum_{1\leq u,v\leq d}\frac{(\Delta\beta_{1ij}^n)^{(u)}(\Delta\beta_{1ij}^n)^{(v)}}{1+\mathbf{1}_{\{u=v\}}}\Big[\frac{1}{\tau^*}\cdot\frac{\partial^2\bar{\sigma}}{\partial\beta_1^{(u)}\partial\beta_1^{(v)}}+\frac{1}{(\tau^*)^2}\cdot X^{(u)}X^{(v)}\Big(\frac{\partial\sigma}{\partial g}((\beta_{1j}^*)^\top X)\Big)^2\Big]\\
&\qquad+\frac{1}{2}(\Delta\tau^n)^2\Big[\frac{2}{(\tau^*)^3}\cdot\bar{\sigma}(X,\beta_{1j}^*)+\frac{1}{(\tau^*)^4}\cdot\bar{\sigma}^2(X,\beta_{1j}^*)\Big]\\
&\qquad-\sum_{u=1}^{d}(\Delta\beta_{1ij}^n)^{(u)}(\Delta\tau^n)\Big[\frac{1}{(\tau^*)^2}\cdot\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}+\frac{1}{(\tau^*)^3}\cdot\bar{\sigma}(X,\beta_{1j}^*)\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta_{1j}^*)\Big]\Bigg\}.
\end{aligned}$$

Thus, the term $[A_n-R_1(X,Y)-R_2(X,Y)]/\mathcal{D}_{6n}$ can be viewed as a linear combination of elements from the union of the following sets:

$$
\begin{aligned}
\mathcal{F}_{1,0} &:= \Big\{\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta_{1j}^*)F(Y;X,\omega_j^*) : u\in[d],\ j:|\mathcal{A}_j|=1\Big\}\cup\Big\{\bar{\sigma}(X,\beta_{1j}^*)F(Y;X,\omega_j^*) : j:|\mathcal{A}_j|=1\Big\},\\
\mathcal{F}_{1,1} &:= \Big\{X^{(u)}F^{(1)}(Y;X,\omega_j^*) : u\in[d],\ j:|\mathcal{A}_j|=1\Big\}\cup\Big\{F^{(1)}(Y;X,\omega_j^*) : j:|\mathcal{A}_j|=1\Big\},\\
\mathcal{F}_{1,2} &:= \Big\{F^{(2)}(Y;X,\omega_j^*) : j:|\mathcal{A}_j|=1\Big\},\\
\mathcal{F}_{2,0} &:= \Big\{X^{\alpha_3}F^{(|\alpha_3|+\ell)}(Y;X,\omega_j^*) : j:|\mathcal{A}_j|>1,\ 0\leq|\alpha_3|\leq\bar{r}_j,\ 0\leq\ell\leq 2(\bar{r}_j-|\alpha_3|)\Big\},\\
\mathcal{F}_{2,1} &:= \Big\{X^{\alpha_3}\cdot\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta_{1j}^*)F^{(|\alpha_3|+\ell)}(Y;X,\omega_j^*),\ X^{\alpha_3}\cdot\bar{\sigma}(X,\beta_{1j}^*)F^{(|\alpha_3|+\ell)}(Y;X,\omega_j^*) :\\
&\qquad\qquad u\in[d],\ j:|\mathcal{A}_j|>1,\ 0\leq|\alpha_3|\leq\bar{r}_j-1,\ 0\leq\ell\leq 2(\bar{r}_j-1-|\alpha_3|)\Big\},\\
\mathcal{F}_{2,2} &:= \Big\{X^{\alpha_3}\cdot\frac{\partial^2\sigma}{\partial\beta_1^{(u)}\partial\beta_1^{(v)}}((\beta_{1j}^*)^\top X)F^{(|\alpha_3|+\ell)}(Y;X,\omega_j^*),\\
&\qquad X^{\alpha_3}\cdot\Big[\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta_{1j}^*)\Big]\Big[\frac{\partial\sigma}{\partial\beta_1^{(v)}}((\beta_{1j}^*)^\top X)\Big]F^{(|\alpha_3|+\ell)}(Y;X,\omega_j^*),\\
&\qquad X^{\alpha_3}\bar{\sigma}(X,\beta_{1j}^*)F^{(|\alpha_3|+\ell)}(Y;X,\omega_j^*),\ X^{\alpha_3}\bar{\sigma}^2(X,\beta_{1j}^*)F^{(|\alpha_3|+\ell)}(Y;X,\omega_j^*),\\
&\qquad X^{\alpha_3}\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta_{1j}^*)F^{(|\alpha_3|+\ell)}(Y;X,\omega_j^*),\ X^{\alpha_3}\bar{\sigma}(X,\beta_{1j}^*)\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta_{1j}^*)F^{(|\alpha_3|+\ell)}(Y;X,\omega_j^*) :\\
&\qquad\qquad u,v\in[d],\ j:|\mathcal{A}_j|>1,\ 0\leq|\alpha_3|\leq\bar{r}_j-2,\ 0\leq\ell\leq 2(\bar{r}_j-2-|\alpha_3|)\Big\},\\
\mathcal{F}_{2,s} &:= \Big\{\sum_{|\alpha_1|+\alpha_2=s}X^{\alpha_3}\cdot\frac{\partial^{|\alpha_1|+\alpha_2}L}{\partial\beta_1^{\alpha_1}\partial\tau^{\alpha_2}}(X,\beta_{1j}^*,\tau^*)\cdot\frac{1}{L(X,\beta_{1j}^*,\tau^*)}F^{(|\alpha_3|+\ell)}(Y;X,\omega_j^*) :\\
&\qquad\qquad j:|\mathcal{A}_j|>1,\ 0\leq|\alpha_3|\leq\bar{r}_j-s,\ 0\leq\ell\leq 2(\bar{r}_j-s-|\alpha_3|)\Big\}.
\end{aligned}
$$

for any $s\geq 3$, where $\omega_j^* := (\beta_{1j}^*,\tau^*,a_j^*,b_j^*,\nu_j^*)$ for any $j\in[k_*]$. Additionally, it is also worth noting that $E_{n,1}/\mathcal{D}_{6n}$ can be seen as a linear combination of elements from the set $\{F(Y;X,\omega_j^*) : j\in[k_*]\}$.

Similarly, we also decompose B_n into two terms as follows:

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

Subsequently, we apply the first-order Taylor expansion to $B_{n,1}$ as in equation (59), and get that

$$
\begin{aligned}
B_{n,1} &= \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\sum_{|\gamma|=1}\frac{1}{\gamma!}(\Delta\beta_{1ij}^n)^{\gamma_1}(\Delta\tau^n)^{\gamma_2}\cdot\frac{\partial H}{\partial\beta_1^{\gamma_1}\,\partial\tau^{\gamma_2}}(Y;X,\beta_{1j}^*,\tau^*)+R_3(X,Y)\\
&= \sum_{j:|\mathcal{A}_j|=1}C_{n,0,j}(X)\cdot H(Y;X,\beta_{1j}^*,\tau^*)+R_3(X,Y),
\end{aligned}
\tag{75}
$$

where $C_{n,0,j}(X)$ is defined in equation (58) and $R_3(X,Y)$ is a Taylor remainder such that $R_3(X,Y)/\mathcal{D}_{6n}\to 0$ as $n\to\infty$.

On the other hand, by means of the second-order Taylor expansion, we rewrite $B_{n,2}$ as

$$\begin{aligned} B_{n,2} &= \sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)\sum_{|\gamma|=1}^{2}\frac{1}{\gamma!}(\Delta\beta^n_{1ij})^{\gamma_1}(\Delta\tau^n)^{\gamma_2}\cdot\frac{\partial^{|\gamma_1|+\gamma_2}H}{\partial\beta_1^{\gamma_1}\,\partial\tau^{\gamma_2}}(Y;X,\beta^*_{1j},\tau^*)+R_4(X,Y)\\ &= \sum_{j:|\mathcal{A}_j|>1}S_{n,j}(X)\cdot H(Y;X,\beta^*_{1j},\tau^*)+R_4(X,Y), \end{aligned} \tag{76}$$

where $R_4(X,Y)$ is a Taylor remainder such that $R_4(X,Y)/\mathcal{D}_{6n}\to 0$ as $n\to\infty$ and

$$\begin{aligned} S_{n,j}(X) := \sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)\bigg\{&\sum_{u=1}^{d}\frac{(\Delta\beta^n_{1ij})^{(u)}}{\tau^*}\cdot\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta^*_{1j})-\frac{\Delta\tau^n}{(\tau^*)^2}\cdot\bar{\sigma}(X,\beta^*_{1j})\\ &-\sum_{u=1}^{d}(\Delta\beta^n_{1ij})^{(u)}(\Delta\tau^n)\Big[\frac{1}{(\tau^*)^2}\cdot\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}+\frac{1}{(\tau^*)^3}\cdot\bar{\sigma}(X,\beta^*_{1j})\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta^*_{1j})\Big]\\ &+\sum_{1\leq u,v\leq d}\frac{(\Delta\beta^n_{1ij})^{(u)}(\Delta\beta^n_{1ij})^{(v)}}{1+\mathbf{1}_{\{u=v\}}}\Big[\frac{1}{\tau^*}\cdot\frac{\partial^2\bar{\sigma}}{\partial\beta_1^{(u)}\partial\beta_1^{(v)}}+\frac{1}{(\tau^*)^2}\cdot\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta^*_{1j})\frac{\partial\bar{\sigma}}{\partial\beta_1^{(v)}}(X,\beta^*_{1j})\Big]\\ &+\frac{1}{2}(\Delta\tau^n)^2\Big[\frac{2}{(\tau^*)^3}\cdot\bar{\sigma}(X,\beta^*_{1j})+\frac{1}{(\tau^*)^4}\cdot\bar{\sigma}^2(X,\beta^*_{1j})\Big]\bigg\}, \end{aligned} \tag{77}$$

for any $j:|\mathcal{A}_j|>1$. Therefore, the term $[B_n-R_3(X,Y)-R_4(X,Y)]/\mathcal{D}_{6n}$ can be treated as a linear combination of elements from the following set:

$$\begin{aligned} \mathcal{H} := &\Big\{\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta^*_{1j})H(Y;X,\beta^*_{1j},\tau^*),\ \bar{\sigma}(X,\beta^*_{1j})H(Y;X,\beta^*_{1j},\tau^*):u\in[d],j\in[k_*]\Big\}\\ &\cup\Big\{\bar{\sigma}(X,\beta^*_{1j})\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta^*_{1j})H(Y;X,\beta^*_{1j},\tau^*),\ \bar{\sigma}^2(X,\beta^*_{1j})H(Y;X,\beta^*_{1j},\tau^*):u\in[d],\ j:|\mathcal{A}_j|>1\Big\}\\ &\cup\Big\{\frac{\partial^2\bar{\sigma}}{\partial\beta_1^{(u)}\partial\beta_1^{(v)}}(X,\beta^*_{1j})H(Y;X,\beta^*_{1j},\tau^*),\ \Big[\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta^*_{1j})\Big]\Big[\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta^*_{1j})\Big]H(Y;X,\beta^*_{1j},\tau^*)\\ &\qquad\qquad:u,v\in[d],\ j:|\mathcal{A}_j|>1\Big\}. \end{aligned}$$

In addition, we can view the term $E_{n,2}/\mathcal{D}_{6n}$ as a linear combination of elements from the set $\{H(Y;X,\beta^*_{1j},\tau^*):j\in[k_*]\}$.

Step 2. In this step, we prove by contradiction that at least one among the coefficients in the representations of $[A_n-R_1(X,Y)-R_2(X,Y)]/\mathcal{D}_{6n}$, $[B_n-R_3(X,Y)-R_4(X,Y)]/\mathcal{D}_{6n}$, $E_{n,1}/\mathcal{D}_{6n}$ and $E_{n,2}/\mathcal{D}_{6n}$ does not converge to zero when $n\to\infty$. Assume that all of them go to 0 as $n\to\infty$. By using the same arguments for showing the results in equations (60) and (61), we get that

$$\frac{1}{\mathcal{D}_{6n}}\cdot\sum_{j=1}^{k_*}\Big|\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)-\exp\Big(\frac{\beta^*_{0j}}{\tau^*}\Big)\Big|\to 0, \tag{78}$$

and

$$\frac{1}{\mathcal{D}_{6n}}\cdot\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)\Big(\|\Delta\beta^n_{1ij}\|+|\Delta\tau^n|+\|\Delta a^n_{ij}\|+|\Delta b^n_{ij}|+|\Delta\nu^n_{ij}|\Big)\to 0. \tag{79}$$

Next, by taking the summation of the absolute values of the coefficients associated with

- $\dfrac{\partial^2\sigma}{\partial[\beta_1^{(u)}]^2}(X,\beta^*_{1j})H(Y;X,\beta^*_{1j},\tau^*)$ in $\mathcal{H}$: we have that $\frac{1}{\mathcal{D}_{6n}}\cdot\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)\|\Delta\beta^n_{1ij}\|^2\to 0$;
- $\bar{\sigma}^2(X,\beta^*_{1j})H(Y;X,\beta^*_{1j},\tau^*)$ in $\mathcal{H}$: we have that $\frac{1}{\mathcal{D}_{6n}}\cdot\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)|\Delta\tau^n|^2\to 0$;

- $[X^{(u)}]^2 F^{(2)}(Y;X,\omega_j^*)$ in $\mathcal{F}_{2,0}$: we have that $\frac{1}{\mathcal{D}_{6n}} \cdot \sum_{j:|\mathcal{A}_j|>1} \sum_{i\in\mathcal{A}_j} \exp\left(\frac{\beta_{0i}^n}{\tau^n}\right) \|\Delta a_{ij}^n\|^2 \to 0$.

As a result, we obtain that

$$\frac{1}{\mathcal{D}_{6n}} \cdot \sum_{j:|\mathcal{A}_j|>1} \sum_{i\in\mathcal{A}_j} \exp\left(\frac{\beta_{0i}^n}{\tau^n}\right) \Big[\|\Delta\beta_{1ij}^n\|^2 + |\Delta\tau^n|^2 + \|\Delta a_{ij}^n\|^2\Big] \to 0 \tag{80}$$

From the results in equations (78), (79) and (80), we deduce that

$$\frac{1}{\mathcal{D}_{6n}} \cdot \sum_{j:|\mathcal{A}_j|>1} \sum_{i\in\mathcal{A}_j} \exp\left(\frac{\beta_{0i}^n}{\tau^n}\right) \Big[|\Delta b_{ij}^n|^{\bar{r}_j} + |\Delta\nu_{ij}^n|^{\bar{r}_j/2}\Big] \to 1,$$

which means that there exists an index $j : |\mathcal{A}_j| > 1$, which can be assumed WLOG to be $j = 1$, such that

$$\frac{1}{\mathcal{D}_{6n}} \cdot \sum_{i\in\mathcal{A}_1} \exp\left(\frac{\beta_{0i}^n}{\tau^n}\right) \Big[|\Delta b_{i1}^n|^{\bar{r}_1} + |\Delta\nu_{i1}^n|^{\bar{r}_1/2}\Big] \not\to 0. \tag{81}$$

Moreover, since the coefficients of elements $F^{(|\alpha_3|+\ell)}(Y;X,\omega_j^*)$, for $j = 1$, $\alpha_3 = \mathbf{0}_d$ and $0 \leq \ell \leq 2\bar{r}_j$ in the set $\mathcal{F}_{2,0}$ converges to zero, i.e.

$$\frac{1}{\mathcal{D}_{6n}} \cdot \sum_{i\in\mathcal{A}_1} \sum_{\substack{\alpha_4+2\alpha_5=\ell,\\ 1\leq\alpha_4+\alpha_5\leq\bar{r}_1}} \frac{\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)}{2^{\alpha_5}\alpha_3!\alpha_4!\alpha_5!} (\Delta b_{ij}^n)^{\alpha_4} (\Delta\nu_{ij}^n)^{\alpha_5} \to 0, \tag{82}$$

for any $1 \leq \ell \leq \bar{r}_1$. Then, we divide the left hand side of equation (82) by that of equation (81), and achieve that

$$\frac{\sum_{i\in\mathcal{A}_1} \sum_{\substack{\alpha_4+2\alpha_5=\ell,\\ 1\leq\alpha_4+\alpha_5\leq\bar{r}_1}} \dfrac{\exp\left(\dfrac{\beta_{0i}^n}{\tau^n}\right)}{2^{\alpha_5}\alpha_4!\alpha_5!} (\Delta b_{ij}^n)^{\alpha_4} (\Delta\nu_{ij}^n)^{\alpha_5}}{\sum_{i\in\mathcal{A}_1} \exp\left(\frac{\beta_{0i}^n}{\tau^n}\right) \Big[|\Delta b_{i1}^n|^{\bar{r}_1} + |\Delta\nu_{i1}^n|^{\bar{r}_1/2}\Big]} \to 0, \tag{83}$$

for any $1 \leq \ell \leq \bar{r}_1$.

Let us define $\overline{M}_n := \max\{|\Delta b_{i1}^n|, |\Delta\nu_{i1}^n|^{1/2} : i \in \mathcal{A}_1\}$ and $\overline{\pi}_n := \max_{i\in\mathcal{A}_1} \exp(\frac{\beta_{0i}^n}{\tau^n})$. Since the sequence $\exp\left(\frac{\beta_{0i}^n}{\tau^n}\right)/\overline{\pi}_n$ is bounded, it is possible to replace it by its subsequence that has a positive limit $q_{3i}^2 := \lim_{n\to\infty} \exp(\frac{\beta_{0i}^n}{\tau^n})/\overline{\pi}_n$. Thus, at least one among q_{3i}^2, for $i \in \mathcal{A}_1$, is equal to one.

In addition, we also define

$$(\Delta b_{i1}^n)/\overline{M}_n \to q_{4i}, \quad (\Delta\nu_{i1}^n)/[2\overline{M}_n] \to q_{5i}.$$

It is worth noting that at least one among q_{4i} and q_{5i} for $i \in \mathcal{A}_1$ is equal to either 1 or -1. Subsequently, we divide both the numerator and the denominator of the ratio in equation (83) by $\overline{\pi}_n\overline{M}_n^{\ell}$, and then obtain the following system of polynomial equations:

$$\sum_{i\in\mathcal{A}_1} \sum_{\substack{\alpha_4+2\alpha_5=\ell,\\ 1\leq\alpha_4+\alpha_5\leq\bar{r}_1}} \frac{q_{3i}^2\, q_{4i}^{\alpha_4}\, q_{5i}^{\alpha_5}}{\alpha_4!\,\alpha_5!} = 0,$$

for all $1 \leq \ell \leq \bar{r}_1$. However, from the definition of $\bar{r}(|\mathcal{A}_1|)$, the above system does not have any non-trivial solutions, which contradicts to the fact that at least one among q_{4i} and q_{5i} for $i \in \mathcal{A}_1$ is non-zero. Therefore, not all the coefficients in the representations of $[A_n - R_1(X,Y) - R_2(X,Y)]/\mathcal{D}_{6n}$, $[B_n - R_3(X,Y) - R_4(X,Y)]/\mathcal{D}_{6n}$, $E_{n,1}/\mathcal{D}_{6n}$ and $E_{n,2}/\mathcal{D}_{6n}$ converge to zero as $n \to \infty$.

Step 3. In this step, we use the Fatou's lemma to show that all the coefficients in the representations of $[A_n - R_1(X,Y) - R_2(X,Y)]/\mathcal{D}_{6n}$, $[B_n - R_3(X,Y) - R_4(X,Y)]/\mathcal{D}_{6n}$, $E_{n,1}/\mathcal{D}_{6n}$ and $E_{n,2}/\mathcal{D}_{6n}$ converge to zero as $n \to \infty$, which leads to a contradiction to the results in Step 2. In particular, let us denote m_n as the maximum of the absolute values of those coefficients. It follows from the claim in Step 2 that $1/m_n \not\to \infty$. Next, we denote

$$\frac{1}{m_n\mathcal{D}_{6n}} \cdot \Big[\sum_{i\in\mathcal{A}_j} \exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big) - \exp\Big(\frac{\beta^*_{0j}}{\tau^*}\Big)\Big] \to \phi_{0,j}, \qquad \frac{1}{m_n\mathcal{D}_{6n}} \cdot \sum_{i\in\mathcal{A}_j} \exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)(\Delta\beta^n_{1ij})^{(u)} \to \phi^{(u)}_{1,j},$$
$$\frac{1}{m_n\mathcal{D}_{6n}} \cdot \sum_{i\in\mathcal{A}_j} \exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)(\Delta\tau^n) \to \phi_{2,j}, \qquad \frac{1}{m_n\mathcal{D}_{6n}} \cdot \sum_{i\in\mathcal{A}_j} \exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)(\Delta a^n_{ij})^{(u)} \to \phi^{(u)}_{3,j},$$
$$\frac{1}{m_n\mathcal{D}_{6n}} \cdot \sum_{i\in\mathcal{A}_j} \exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)(\Delta b^n_{ij}) \to \phi_{4,j}, \qquad \frac{1}{m_n\mathcal{D}_{6n}} \cdot \sum_{i\in\mathcal{A}_j} \exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)(\Delta \nu^n_{ij}) \to \phi_{5,j},$$

as $n \to \infty$ for any $u \in [d]$ and $j \in [k_*]$. By means of the Fatou's lemma, we have that

$$\lim_{n\to\infty} \frac{\mathbb{E}_X[V(p_{G_n}(\cdot|X), p_{G_*}(\cdot|X))]}{m_n\mathcal{D}_{6n}} \geq \int \liminf_{n\to\infty} \frac{|p_{G_n}(Y|X) - p_{G_*}(Y|X)|}{2m_n\mathcal{D}_{6n}} \mathrm{d}(X,Y).$$

Recall from equation (68) that the limit the left hand side is equal to zero, which implies that $\frac{|p_{G_n}(Y|X)-p_{G_*}(Y|X)|}{m_n\mathcal{D}_{6n}} \to 0$. as $n\to\infty$ for almost surely (X,Y). Thus, we also get that $\frac{Q_n}{m_n\mathcal{D}_{6n}} \to 0$ as $n\to\infty$, which implies that

$$\lim_{n\to\infty} \frac{1}{m_n\mathcal{D}_{6n}} \cdot [A_{n,1} + A_{n,2} - B_{n,1} - B_{n,2} + E_{n,1} - E_{n,2}] = \lim_{n\to\infty} \frac{Q_n}{m_n\mathcal{D}_{6n}} = 0, \tag{84}$$

for almost surely (X,Y). Now, we derive the limits of terms in the above right hand side. In particular, from the formulations of

- $E_{n,1}$ and $E_{n,2}$ in equation (70), we have

$$\frac{E_{n,1}}{m_n\mathcal{D}_{6n}} \to \sum_{j=1}^{k_*} \phi_{0,j} F(Y;X,\omega^*_j), \qquad \frac{E_{n,2}}{m_n\mathcal{D}_{6n}} \to \sum_{j=1}^{k_*} \phi_{0,j} H(Y;X,\beta^*_{1j},\tau^*). \tag{85}$$

- $A_{n,1}$ in equation (71), we deduce that

$$\frac{A_{n,1}}{m_n\mathcal{D}_{6n}} \to \sum_{j:|\mathcal{A}_j|=1} \sum_{\eta=0}^{2} C^*_{\eta,j}(X) \cdot F^{(\eta)}(Y;X,\omega^*_j), \tag{86}$$

where

$$C^*_{0,j}(X) := \sum_{u=1}^{d} \frac{\phi^{(u)}_{1,j}}{\tau^*} \cdot \frac{\partial\bar{\sigma}}{\partial\beta^{(u)}_1}(X,\beta^*_{1j}) - \phi_{2,j} \cdot \frac{\bar{\sigma}(X,\beta^*_{1j})}{(\tau^*)^2},$$
$$C^*_{1,j}(X) := \sum_{u=1}^{d} \phi^{(u)}_{3,j} \cdot X^{(u)} + \phi_{4,j},$$
$$C^*_{2,j}(X) := \frac{1}{2}\phi_{5,j},$$

for any $j : |\mathcal{A}_j| = 1$.

- $B_{n,1}$ in equation (75), we get

$$\frac{B_{n,1}}{\mathcal{D}_{6n}} \to \sum_{j:|\mathcal{A}_j|=1} C^*_{0,j}(X) H(Y;X,\beta^*_{1j},\tau^*). \tag{87}$$

- $A_{n,2}$ in equation (74), we have

$$\lim_{n\to\infty}\frac{A_{n,2}}{\mathcal{D}_{6n}} = \sum_{j:|\mathcal{A}_j|>1}\sum_{s=0}^{\bar{r}_j}\sum_{|\alpha_3|=0}^{\bar{r}_j-s}\sum_{\ell=0}^{2(\bar{r}_j-s-|\alpha_3|)}\lim_{n\to\infty}\frac{T_{n,s,\alpha_3,\ell,j}(X)}{\mathcal{D}_{6n}}\cdot X^{\alpha_3}F^{(|\alpha_3|+\ell)}(Y;X,\omega_j^*).$$

From the arguments in Step 2, we deduce that the value of m_n is the ratio between one element of the following set and the loss $\mathcal{D}_{6n}$:

$$\Big\{\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)|\Delta\beta_{1ij}^n|, \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)|\Delta\tau^n|, \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)|\Delta a_{ij}^n|, \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)|\Delta b_{ij}^n|, \exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)|\Delta\nu_{ij}^n|, i\in\mathcal{A}_j, j:|\mathcal{A}_j|>1\Big\}$$
$$\cup\Big\{\sum_{i\in\mathcal{A}_j}(\Delta\beta_{1ij}^n)^2, \sum_{i\in\mathcal{A}_j}(\Delta\tau^n)^2, \sum_{i\in\mathcal{A}_j}(\Delta a_{ij}^n)^2, j:|\mathcal{A}_j|>1\Big\}. \tag{88}$$

Thus, the associated coefficients $T_{n,s,\alpha_3,\ell,j}/\mathcal{D}_{6n}$ in the representation of $A_{n,2}/\mathcal{D}_{6n}$ converge to zero as $n\to\infty$ for any $s\geq 3$. Therefore, we consider only the limits of $T_{n,s,\alpha_3,\ell,j}/\mathcal{D}_{6n}$ for $s\in\{0,1,2\}$. In particular, let us denote

$$\frac{1}{m_n\mathcal{D}_{6n}}\cdot\sum_{i\in\mathcal{A}_j}\sum_{\substack{\alpha_4+2\alpha_5=\ell,\\1\leq\alpha_4+\alpha_5\leq\bar{r}_j}}\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\frac{1}{2^{\alpha_5}\alpha_3!\alpha_4!\alpha_5!}(\Delta a_{ij}^n)^{\alpha_3}(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5}\to\psi_{0,\alpha_3,\ell,j},$$

$$\frac{1}{m_n\mathcal{D}_{6n}}\cdot\sum_{i\in\mathcal{A}_j}\sum_{\substack{\alpha_4+2\alpha_5=\ell,\\1\leq\alpha_4+\alpha_5\leq\bar{r}_j}}\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\frac{1}{2^{\alpha_5}\alpha_3!\alpha_4!\alpha_5!}(\Delta a_{ij}^n)^{\alpha_3}(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5}(\Delta\beta_{1ij}^n)^{(u)}\to\psi_{1,\alpha_3,\ell,j}^{(u)},$$

$$\frac{1}{m_n\mathcal{D}_{6n}}\cdot\sum_{i\in\mathcal{A}_j}\sum_{\substack{\alpha_4+2\alpha_5=\ell,\\1\leq\alpha_4+\alpha_5\leq\bar{r}_j}}\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\frac{1}{2^{\alpha_5}\alpha_3!\alpha_4!\alpha_5!}(\Delta a_{ij}^n)^{\alpha_3}(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5}(\Delta\tau^n)\to\psi_{2,\alpha_3,\ell,j},$$

$$\frac{1}{m_n\mathcal{D}_{6n}}\cdot\sum_{i\in\mathcal{A}_j}\sum_{\substack{\alpha_4+2\alpha_5=\ell,\\1\leq\alpha_4+\alpha_5\leq\bar{r}_j}}\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\frac{1}{2^{\alpha_5}\alpha_3!\alpha_4!\alpha_5!}(\Delta a_{ij}^n)^{\alpha_3}(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5}(\Delta\beta_{1ij}^n)^{(u)}(\Delta\beta_{1ij}^n)^{(v)}\to\psi_{3,\alpha_3,\ell,j}^{(u,v)},$$

$$\frac{1}{m_n\mathcal{D}_{6n}}\cdot\sum_{i\in\mathcal{A}_j}\sum_{\substack{\alpha_4+2\alpha_5=\ell,\\1\leq\alpha_4+\alpha_5\leq\bar{r}_j}}\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\frac{1}{2^{\alpha_5}\alpha_3!\alpha_4!\alpha_5!}(\Delta a_{ij}^n)^{\alpha_3}(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5}(\Delta\beta_{1ij}^n)^{(u)}(\Delta\tau^n)\to\psi_{4,\alpha_3,\ell,j}^{(u)},$$

$$\frac{1}{m_n\mathcal{D}_{6n}}\cdot\sum_{i\in\mathcal{A}_j}\sum_{\substack{\alpha_4+2\alpha_5=\ell,\\1\leq\alpha_4+\alpha_5\leq\bar{r}_j}}\exp\Big(\frac{\beta_{0i}^n}{\tau^n}\Big)\frac{1}{2^{\alpha_5}\alpha_3!\alpha_4!\alpha_5!}(\Delta a_{ij}^n)^{\alpha_3}(\Delta b_{ij}^n)^{\alpha_4}(\Delta\nu_{ij}^n)^{\alpha_5}(\Delta\tau^n)^2\to\psi_{5,\alpha_3,\ell,j},$$

for any $j:|\mathcal{A}_j|>1$ and $u,v\in[d]$. Then, we have that

$$\frac{A_{n,2}}{\mathcal{D}_{6n}}\to\sum_{j:|\mathcal{A}_j|>1}\sum_{s=0}^{2}\sum_{|\alpha_3|=0}^{\bar{r}_j-s}\sum_{\ell=0}^{2(\bar{r}_j-s-|\alpha_3|)}T_{s,\alpha_3,\ell,j}^*\cdot X^{\alpha_3}F^{(|\alpha_3|+\ell)}(Y;X,\omega_j^*), \tag{89}$$

where

$$T_{0,\alpha_3,\ell,j}^* := \psi_{0,\alpha_3,\ell,j},$$
$$T_{1,\alpha_3,\ell,j}^* := \sum_{u=1}^{d}\frac{\psi_{1,\alpha_3,\ell,j}^{(u)}}{\tau^*}\cdot\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta_{1j}^*) - \frac{\psi_{2,\alpha_3,\ell,j}}{(\tau^*)^2}\cdot\bar{\sigma}(X,\beta_{1j}^*),$$
$$T_{2,\alpha_3,\ell,j}^* := \sum_{1\leq u,v\leq d}\frac{\psi_{3,\alpha_3,\ell,j}^{(u,v)}}{1+\mathbf{1}_{\{u=v\}}}\Big[\frac{1}{\tau^*}\cdot\frac{\partial^2\sigma}{\partial\beta_1^{(u)}\partial\beta_1^{(v)}}((\beta_{1j}^*)^\top X) + \frac{1}{(\tau^*)^2}\cdot\Big(\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta_{1j}^*)\Big)\Big(\frac{\partial\sigma}{\partial\beta_1^{(v)}}((\beta_{1j}^*)^\top X)\Big)\Big]$$
$$+\frac{1}{2}\psi_{5,\alpha_3,\ell,j}\Big[\frac{2}{(\tau^*)^3}\cdot\bar{\sigma}(X,\beta_{1j}^*) + \frac{1}{(\tau^*)^4}\cdot\bar{\sigma}^2(X,\beta_{1j}^*)\Big]$$
$$-\sum_{u=1}^{d}\psi_{4,\alpha_3,\ell,j}^{(u)}\Big[\frac{1}{(\tau^*)^2}\cdot\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta_{1j}^*) + \frac{1}{(\tau^*)^3}\cdot\bar{\sigma}(X,\beta_{1j}^*)\frac{\partial\bar{\sigma}}{\partial\beta_1^{(u)}}(X,\beta_{1j}^*)\Big].$$

- $B_{n,2}$ in equation (76), by denoting

$$\frac{1}{m_n\mathcal{D}_{6n}}\cdot\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)(\Delta\beta^n_{1ij})^{(u)}(\Delta\beta^n_{1ij})^{(v)}\to\varphi^{(u,v)}_{0,j},$$

$$\frac{1}{m_n\mathcal{D}_{6n}}\cdot\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)(\Delta\beta^n_{1ij})^{(u)}(\Delta\tau^n)\to\varphi^{(u)}_{1,j},$$

$$\frac{1}{m_n\mathcal{D}_{6n}}\cdot\sum_{i\in\mathcal{A}_j}\exp\Big(\frac{\beta^n_{0i}}{\tau^n}\Big)(\Delta\tau^n)^2\to\varphi_{2,j},$$

we have

$$\frac{B_{n,2}}{\mathcal{D}_{6n}}\to\sum_{j:|\mathcal{A}_j|>1}S^*_j(X)\cdot H(Y;X,\beta^*_{1j},\tau^*),\tag{90}$$

where

$$\begin{aligned}S^*_j(X):=\Bigg\{&\sum_{u=1}^{d}\frac{\phi^{(u)}_{1,j}}{\tau^*}\cdot\frac{\partial\bar\sigma}{\partial\beta^{(u)}_1}(X,\beta^*_{1j})-\frac{\phi_{2,j}}{(\tau^*)^2}\cdot\bar\sigma(X,\beta^*_{1j})-\sum_{u=1}^{d}\varphi^{(u)}_{1,j}\Big[\frac{1}{(\tau^*)^2}\cdot\frac{\partial\bar\sigma}{\partial\beta^{(u)}_1}(X,\beta^*_{1j})\\&+\frac{1}{(\tau^*)^3}\cdot\bar\sigma(X,\beta^*_{1j})\frac{\partial\bar\sigma}{\partial\beta^{(u)}_1}(X,\beta^*_{1j})\Big]+\sum_{1\le u,v\le d}\frac{\varphi^{(u,v)}_{0,j}}{1+\mathbf{1}_{\{u=v\}}}\Big[\frac{1}{\tau^*}\cdot\frac{\partial^2\sigma}{\partial\beta^{(u)}_1\partial\beta^{(v)}_1}((\beta^*_{1j})^\top X)\\&+\frac{1}{(\tau^*)^2}\cdot\Big(\frac{\partial\bar\sigma}{\partial\beta^{(u)}_1}(X,\beta^*_{1j})\Big)\Big(\frac{\partial\sigma}{\partial\beta^{(v)}_1}((\beta^*_{1j})^\top X)\Big)\Big]+\frac{1}{2}\varphi_{2,j}\Big[\frac{2}{(\tau^*)^3}\cdot\bar\sigma(X,\beta^*_{1j})+\frac{1}{(\tau^*)^4}\cdot\bar\sigma^2(X,\beta^*_{1j})\Big]\Bigg\}.\end{aligned}$$

Recall that not all the ratios between elements in the set (88) and the loss $\mathcal{D}_{6n}$ converge to zero as $n\to\infty$. Thus, at least one element of the following set union is non-zero:

$$\Big\{\phi_{0,j},\phi^{(u)}_{1,j},\phi_{2,j},\phi^{(u)}_{3,j},\phi_{4,j},\phi_{5,j}:u\in[d],j:|\mathcal{A}_j|=1\Big\}\cup\Big\{\phi_{0,j},\psi_{0,2\mathbf{e}_u,0,j},\varphi^{(u,u)}_{0,j},\varphi_{2,j}:u\in[d],j:|\mathcal{A}_j|>1\Big\}\tag{91}$$

Now, we show that all elements in the union (91) must be zero. Indeed, putting the results in equations (84), (85), (86), (87), (89) and (90), we obtain that

$$\begin{aligned}&\sum_{j=1}^{k_*}\phi_{0,j}F(Y;X,\omega^*_j)+\sum_{j:|\mathcal{A}_j|=1}\sum_{\eta=0}^{2}C^*_{\eta,j}(X)F^{(\eta)}(Y;X,\omega^*_j)-\sum_{j:|\mathcal{A}_j|=1}(C^*_{0,j}(X)+\phi_{0,j})H(Y;X,\beta^*_{1j},\tau^*)\\&+\sum_{j:|\mathcal{A}_j|>1}\sum_{s=0}^{2}\sum_{|\alpha_3|=0}^{\bar r_j-s}\sum_{\ell=0}^{2(\bar r_j-s-|\alpha_3|)}T^*_{s,\alpha_3,\ell,j}(X)\cdot X^{\alpha_3}F^{(|\alpha_3|+\ell)}(Y;X,\omega^*_j)\\&\qquad\qquad-\sum_{j:|\mathcal{A}_j|>1}(S^*_j(X)+\phi_{0,j})\cdot H(Y;X,\beta^*_{1j},\tau^*)=0,\end{aligned}\tag{92}$$

for almost surely (X,Y). It can be verified that the set

$$\begin{aligned}&\Big\{F^{(\eta)}(Y;X,\omega^*_j),\ H(Y;X,\beta^*_{1j},\tau^*):0\le\eta\le2,j:|\mathcal{A}_j|=1\Big\}\\\cup&\Big\{X^{\alpha_3}F^{(|\alpha_3|+\ell)}(Y;X,\omega^*_j),\ H(Y;X,\beta^*_{1j},\tau^*):j:|\mathcal{A}_j|>1,\ 0\le\alpha_3\le\bar r_j,\ 0\le\ell\le2(\bar r_j-|\alpha_3|)\Big\}.\end{aligned}$$

is linearly independent w.r.t Y. Thus, in the left hand side of equation (92), the coefficients associated with the following terms must be zero.

- $F(Y;X,\omega^*_j)$, where $j:|\mathcal{A}_j|=1$: $\phi_{0,j}+C^*_{0,j}(X)=0$. More explicitly, we have

$$\phi_{0,j}+\sum_{u=1}^{d}\frac{\phi^{(u)}_{1,j}}{\tau^*}\cdot\frac{\partial\bar\sigma}{\partial\beta^{(u)}_1}(X,\beta^*_{1j})-\phi_{2,j}\cdot\frac{\bar\sigma(X,\beta^*_{1j})}{(\tau^*)^2}=0,$$

for any $j : |\mathcal{A}_j| = 1$, for almost surely X. Since the function σ satisfies the conditions in Definition 3.4, we deduce that $\phi_{0,j} = \phi_{1,j}^{(u)} = \phi_{2,j} = 0$, for any $j : |\mathcal{A}_j| = 1$.

- $F^{(1)}(Y; X, \omega_j^*)$, where $j : |\mathcal{A}_j| = 1$: $C_{1,j}^*(X) = 0$. More explicitly, we have

$$\sum_{u=1}^{d} \phi_{3,j}^{(u)} \cdot X^{(u)} + \phi_{4,j} = 0,$$

for almost surely X. Since the set $\{X^{(u)}, 1 : u \in [d]\}$ is linearly independent, the above equation implies that $\phi_{3,j}^{(u)} = \phi_{4,j} = 0$ for any $j : |\mathcal{A}_j| = 1$.

- $F^{(2)}(Y; X, \omega_j^*)$, where $j : |\mathcal{A}_j| = 1$: $C_{2,j}^*(X) = 0$, or equivalently, $\phi_{5,j} = 0$ for any $j : |\mathcal{A}_j| = 1$.

- $H(Y; X, \beta_{1j}^*, \tau^*)$, where $j : |\mathcal{A}_j| > 1$: $S_j^*(X) + \phi_{0,j} = 0$. More explicitly, we have

$$\begin{aligned}
\phi_{0,j} + \Bigg\{ &\sum_{u=1}^{d} \frac{\phi_{1,j}^{(u)}}{\tau^*} \cdot \frac{\partial \bar{\sigma}}{\partial \beta_1^{(u)}}(X, \beta_{1j}^*) - \frac{\phi_{2,j}}{(\tau^*)^2} \cdot \bar{\sigma}(X, \beta_{1j}^*) - \sum_{u=1}^{d} \varphi_{1,j}^{(u)} \Big[\frac{1}{(\tau^*)^2} \cdot \frac{\partial \bar{\sigma}}{\partial \beta_1^{(u)}}(X, \beta_{1j}^*) \\
&+ \frac{1}{(\tau^*)^3} \cdot \bar{\sigma}(X, \beta_{1j}^*) \frac{\partial \bar{\sigma}}{\partial \beta_1^{(u)}}(X, \beta_{1j}^*) \Big] + \sum_{1 \le u,v \le d} \frac{\varphi_{0,j}^{(u,v)}}{1 + \mathbf{1}_{\{u=v\}}} \Big[\frac{1}{\tau^*} \cdot \frac{\partial^2 \sigma}{\partial \beta_1^{(u)} \partial \beta_1^{(v)}}((\beta_{1j}^*)^\top X) \\
&+ \frac{1}{(\tau^*)^2} \cdot \Big(\frac{\partial \bar{\sigma}}{\partial \beta_1^{(u)}}(X, \beta_{1j}^*) \Big) \Big(\frac{\partial \sigma}{\partial \beta_1^{(v)}}((\beta_{1j}^*)^\top X) \Big) \Big] + \frac{1}{2} \varphi_{2,j} \Big[\frac{2}{(\tau^*)^3} \cdot \bar{\sigma}(X, \beta_{1j}^*) + \frac{1}{(\tau^*)^4} \cdot \bar{\sigma}^2(X, \beta_{1j}^*) \Big] \Bigg\} = 0,
\end{aligned}$$

for almost surely X. As the function σ meets the conditions in Definition (3.4), i.e. the set

$$\begin{aligned}
\Big\{ 1,\ \bar{\sigma}(X, \beta_{1j}^*),\ \bar{\sigma}^2(X, \beta_{1j}^*),\ \frac{\partial \bar{\sigma}}{\partial \beta_1^{(u)}}(X, \beta_{1j}^*),\ \bar{\sigma}(X, \beta_{1j}^*) \frac{\partial \bar{\sigma}}{\partial \beta_1^{(u)}}(X, \beta_{1j}^*), \\
\frac{\partial^2 \sigma}{\partial \beta_1^{(u)} \partial \beta_1^{(v)}}((\beta_{1j}^*)^\top X),\ \Big(\frac{\partial \bar{\sigma}}{\partial \beta_1^{(u)}}(X, \beta_{1j}^*) \Big) \Big(\frac{\partial \sigma}{\partial \beta_1^{(v)}}((\beta_{1j}^*)^\top X) \Big) : u, v \in [d] \Big\}
\end{aligned}$$

is linearly independent, the coefficients associated with 1, $\bar{\sigma}^2(X, \beta_{1j}^*)$, $\bar{\sigma}(X, \beta_{1j}^*) \frac{\partial \bar{\sigma}}{\partial \beta_1^{(u)}}(X, \beta_{1j}^*)$ and $\frac{\partial^2 \sigma}{\partial \beta_1^{(u)} \partial \beta_1^{(v)}}((\beta_{1j}^*)^\top X)$ must be zero, i.e. $\phi_{0,j} = \varphi_{2,j} = \varphi_{1,j}^{(u)} = \varphi_{0,j}^{(u,v)} = 0$ for any $u, v \in [d]$ and $j : |\mathcal{A}_j| > 1$.

- $X^{\alpha_3} F^{(|\alpha_3| + \ell)}(Y; X, \beta_{1j}^*, \tau^*)$, where $\alpha_3 = 2\mathbf{e}_u$, $\ell = 0$, $j : |\mathcal{A}_j| > 1$: $T_{0,\alpha_3,\ell,j}^*(X) + T_{1,\alpha_3,\ell,j}^*(X) + T_{2,\alpha_3,\ell,j}^*(X) = 0$. More explicitly, we have

$$\begin{aligned}
\psi_{0,\alpha_3,\ell,j} &+ \sum_{u=1}^{d} \frac{\psi_{1,\alpha_3,\ell,j}^{(u)}}{\tau^*} \cdot \frac{\partial \bar{\sigma}}{\partial \beta_1^{(u)}}(X, \beta_{1j}^*) - \frac{\psi_{2,\alpha_3,\ell,j}}{(\tau^*)^2} \cdot \bar{\sigma}(X, \beta_{1j}^*), \\
&+ \sum_{1 \le u,v \le d} \frac{\psi_{3,\alpha_3,\ell,j}^{(u,v)}}{1 + \mathbf{1}_{\{u=v\}}} \Big[\frac{1}{\tau^*} \cdot \frac{\partial^2 \sigma}{\partial \beta_1^{(u)} \partial \beta_1^{(v)}}((\beta_{1j}^*)^\top X) + \frac{1}{(\tau^*)^2} \cdot \Big(\frac{\partial \bar{\sigma}}{\partial \beta_1^{(u)}}(X, \beta_{1j}^*) \Big) \Big(\frac{\partial \sigma}{\partial \beta_1^{(v)}}((\beta_{1j}^*)^\top X) \Big) \Big] \\
&\qquad + \frac{1}{2} \psi_{5,\alpha_3,\ell,j} \Big[\frac{2}{(\tau^*)^3} \cdot \bar{\sigma}(X, \beta_{1j}^*) + \frac{1}{(\tau^*)^4} \cdot \bar{\sigma}^2(X, \beta_{1j}^*) \Big] \\
&\qquad - \sum_{u=1}^{d} \psi_{4,\alpha_3,\ell,j}^{(u)} \Big[\frac{1}{(\tau^*)^2} \cdot \frac{\partial \bar{\sigma}}{\partial \beta_1^{(u)}}(X, \beta_{1j}^*) + \frac{1}{(\tau^*)^3} \cdot \bar{\sigma}(X, \beta_{1j}^*) \frac{\partial \bar{\sigma}}{\partial \beta_1^{(u)}}(X, \beta_{1j}^*) \Big] = 0,
\end{aligned}$$

for almost surely X. Since the function σ satisfies the conditions in Definition 3.4, we deduce that $\psi_{0,\alpha_3,\ell,j} = \psi_{0,2\mathbf{e}_u,0,j} = 0$, for any $u \in [d]$ and $j : |\mathcal{A}_j| > 1$.

Gather the above results, we see that all elements in the set (88) are equal to zero, which is a contradiction. Hence, we reach the conclusion of the theorem.

C. Identifiability of the (Activation) Dense-to-sparse Gating Gaussian Mixture of Experts

Proposition C.1. *Assume that G is a mixing measure in $\mathcal{O}_k(\Theta)$ that satisfy $g_G(Y|X) = g_{G_*}(Y|X)$ for almost surely (X,Y). Then, we obtain that $G \equiv G_*(\lambda)$, where $G_*(\lambda) := \sum_{i=1}^{k_*} \exp(\beta^*_{0i}/\tau^*)\delta_{(\lambda\beta^*_{1i},\lambda\tau^*,a^*_i,b^*_i,\nu^*_i)}$, for some $\lambda \neq 0$.*

Proof of Proposition C.1. Firstly, let us recall that two mixing measures G and G_* admit the following forms:

$$G = \sum_{i=1}^{k'} \exp\left(\frac{\beta_{0i}}{\tau}\right)\delta_{(\beta_{1i},\tau,a_i,b_i,\nu_i)}, \qquad G_* = \sum_{i=1}^{k_*} \exp\left(\frac{\beta^*_{0i}}{\tau^*}\right)\delta_{(\beta^*_{1i},\tau^*,a^*_i,b^*_i,\nu^*_i)},$$

where $k' \leq k$. Since $g_G(Y|X) = g_{G_*}(Y|X)$ for almost surely (X,Y), we have

$$\sum_{i=1}^{k} \mathrm{Softmax}\Big(\frac{(\beta_{1i})^\top X + \beta_{0i}}{\tau}\Big) \cdot f(Y|a_i^\top X + b_i, \nu_i) = \sum_{i=1}^{k'} \mathrm{Softmax}\Big(\frac{(\beta^*_{1i})^\top X + \beta^*_{0i}}{\tau^*}\Big) \cdot f(Y|(a^*_i)^\top + b^*_i, \nu^*_i). \quad (93)$$

As the mixture of location-scale Gaussian distributions is identifiable (Teicher, 1960; 1961; 1963), it follows that $k' = k_*$ and

$$\Big\{\mathrm{Softmax}\Big(\frac{(\beta_{1i})^\top X + \beta_{0i}}{\tau}\Big) : i \in [k']\Big\} = \Big\{\mathrm{Softmax}\Big(\frac{(\beta^*_{1i})^\top X + \beta^*_{0i}}{\tau^*}\Big) : i \in [k_*]\Big\},$$

for almost surely X. WLOG, we may assume that

$$\mathrm{Softmax}\Big(\frac{(\beta_{1i})^\top X + \beta_{0i}}{\tau}\Big) = \mathrm{Softmax}\Big(\frac{(\beta^*_{1i})^\top X + \beta^*_{0i}}{\tau^*}\Big), \quad (94)$$

for almost surely X for any $i \in [k_*]$. It is worth noting that the Softmax function is invariant to translations, then equation (94) indicates that $\beta_{1i}/\tau = \beta^*_{1i}/\tau^* + v_1$ and $\beta_{0i}/\tau = \beta^*_{0i}/\tau^* + v_0$ for some $v_1 \in \mathbb{R}^d$ and $v_0 \in \mathbb{R}$. However, from the assumptions $\beta_{1k} = \beta^*_{1k} = \mathbf{0}_d$ and $\beta_{0k} = \beta^*_{0k} = 0$, we deduce that $v_1 = \mathbf{0}_d$ and $v_0 = 0$. Consequently, we get that $\beta_{1i}/\tau = \beta^*_{1i}/\tau^*$ and $\beta_{0i}/\tau = \beta^*_{0i}/\tau^*$ for any $i \in [k]$. Thus, we deduce that $\beta_{1i} = \lambda\beta^*_{1i}$ and $\tau = \lambda\tau^*$, for some $\lambda \neq 0$.

Then, equation (93) can be rewritten as

$$\sum_{i=1}^{k_*} \exp\left(\frac{\beta_{0i}}{\tau}\right)\exp\left(\frac{(\beta_{1i})^\top X}{\tau}\right) f(Y|(a_i)^\top X + b_i, \nu_i) = \sum_{i=1}^{k_*} \exp\left(\frac{\beta^*_{0i}}{\tau^*}\right)\exp\left(\frac{(\beta^*_{1i})^\top X}{\tau^*}\right) f(Y|(a^*_i)^\top X + b^*_i, \nu^*_i), \quad (95)$$

for almost surely (X,Y). Next, we denote $J_1, J_2, \ldots, J_m$ as a partition of the index set $[k_*]$, where $m \leq k$, such that $\exp(\beta_{0i}/\tau) = \exp(\beta^*_{0i'}/\tau^*)$ for any $i, i' \in J_j$ and $j \in [k_*]$. On the other hand, when i and i' do not belong to the same set J_j, we let $\exp(\beta_{0i}/\tau) \neq \exp(\beta_{0i'}/\tau^*)$. Thus, we can reformulate equation (95) as

$$\sum_{j=1}^{m}\sum_{i\in J_j} \exp\left(\frac{\beta_{0i}}{\tau}\right)\exp\left(\frac{(\beta_{1i})^\top X}{\tau}\right) f(Y|(a_i)^\top X + b_i, \nu_i) = \sum_{j=1}^{m}\sum_{i\in J_j} \exp\left(\frac{\beta^*_{0i}}{\tau^*}\right)\exp\left(\frac{(\beta^*_{1i})^\top X}{\tau^*}\right) f(Y|(a^*_i)^\top X + b^*_i, \nu^*_i),$$

for almost surely (X,Y). This results leads to $\{((a_i)^\top X + b_i, \nu_i) : i \in J_j\} \equiv \{((a^*_i)^\top X + b^*_i, \nu^*_i) : i \in J_j\}$, for almost surely X for any $j \in [m]$. Therefore, we have

$$\{(a_i, b_i, \nu_i) : i \in J_j\} \equiv \{(a^*_i, b^*_i, \nu^*_i) : i \in J_j\},$$

for any $j \in [m]$. As a consequence,

$$G = \sum_{j=1}^{m}\sum_{i\in J_j} \exp\left(\frac{\beta_{0i}}{\tau}\right)\delta_{(\beta_{1i},\tau,a_i,b_i,\nu_i)} = \sum_{j=1}^{m}\sum_{i\in J_j} \exp\left(\frac{\beta_{0i}}{\tau}\right)\delta_{(\lambda\beta^*_{1i},\lambda\tau^*,a^*_i,b^*_i,\nu^*_i)} = G_*(\lambda).$$

Hence, we reach the conclusion of this proposition. □

Proposition C.2. *Assume that G is a mixing measure in $\mathcal{O}_k(\Theta)$ that satisfy $p_G(Y|X) = p_{G_*}(Y|X)$ for almost surely (X,Y). Then, we obtain that $G \equiv G_*$.*

Proof of Proposition C.2. Firstly, let us recall that two mixing measures G and G' admit the following forms:

$$G = \sum_{i=1}^{k'} \exp\left(\frac{\beta_{0i}}{\tau}\right)\delta_{(\beta_{1i},\tau,a_i,b_i,\nu_i)}, \qquad G_* = \sum_{i=1}^{k_*} \exp\left(\frac{\beta_{0i}^*}{\tau^*}\right)\delta_{(\beta_{1i}^*,\tau^*,a_i^*,b_i^*,\nu_i^*)},$$

where $k' \leq k$. Since $p_G(Y|X) = p_{G_*}(Y|X)$ for almost surely (X,Y), we have

$$\sum_{i=1}^{k} \mathrm{Softmax}\Big(\frac{\sigma((\beta_{1i})^\top X) + \beta_{0i}}{\tau}\Big) \cdot f(Y|a_i^\top X + b_i, \nu_i) = \sum_{i=1}^{k'} \mathrm{Softmax}\Big(\frac{\sigma((\beta_{1i}^*)^\top X) + \beta_{0i}^*}{\tau^*}\Big) \cdot f(Y|(a_i^*)^\top + b_i^*, \nu_i^*). \tag{96}$$

As the mixture of location-scale Gaussian distributions is identifiable (Teicher, 1960; 1961; 1963), it follows that $k' = k_*$ and

$$\Big\{\mathrm{Softmax}\Big(\frac{\sigma((\beta_{1i})^\top X) + \beta_{0i}}{\tau}\Big) : i \in [k']\Big\} = \Big\{\mathrm{Softmax}\Big(\frac{\sigma((\beta_{1i}^*)^\top X) + \beta_{0i}^*}{\tau^*}\Big) : i \in [k_*]\Big\},$$

for almost surely X. WLOG, we may assume that

$$\mathrm{Softmax}\Big(\frac{\sigma((\beta_{1i})^\top X) + \beta_{0i}}{\tau}\Big) = \mathrm{Softmax}\Big(\frac{\sigma((\beta_{1i}^*)^\top X) + \beta_{0i}^*}{\tau^*}\Big), \tag{97}$$

for almost surely X for any $i \in [k_*]$. It is worth noting that the $\mathrm{Softmax}$ function is invariant to translations, then equation (97) indicates that $[\sigma((\beta_{1i})^\top X) + \beta_{0i}]/\tau = [\sigma((\beta_{1i}^*)^\top X) + \beta_{0i}^*]/\tau^* + v$ for some $v \in \mathbb{R}$. However, from the assumptions $\sigma((\beta_{1k_*}^*)^\top X) + \beta_{0k_*}^* = 0$, we deduce $v = 0$. Consequently, we get that

$$[\sigma((\beta_{1i})^\top X) + \beta_{0i}]/\tau = [\sigma((\beta_{1i}^*)^\top X) + \beta_{0i}^*]/\tau^*,$$

for almost surely X for any $i \in [k]$. Thus, when $X = \mathbf{0}_d$, we deduce that $\beta_{0i}/\tau = \beta_{0i}^*/\tau^*$, which implies that $\sigma((\beta_{1i})^\top X)/\tau = \sigma((\beta_{1i}^*)^\top X)/\tau^*$ for almost surely X. Again, when $X = \mathbf{0}_d$, since $\sigma(0) \neq 0$, we obtain that $\tau = \tau^*$, and therefore, $\beta_{1i} = \beta_{1i}^*$ for any $i \in [k_*]$.

Then, equation (96) can be rewritten as

$$\sum_{i=1}^{k_*} \exp\left(\frac{\beta_{0i}}{\tau}\right) \exp\left(\frac{\sigma((\beta_{1i})^\top X)}{\tau}\right) f(Y|(a_i)^\top X + b_i, \nu_i) = \sum_{i=1}^{k_*} \exp\left(\frac{\beta_{0i}^*}{\tau^*}\right) \exp\left(\frac{\sigma((\beta_{1i}^*)^\top X)}{\tau^*}\right) f(Y|(a_i^*)^\top X + b_i^*, \nu_i^*), \tag{98}$$

for almost surely (X,Y). Next, we denote $J_1, J_2, \ldots, J_m$ as a partition of the index set $[k_*]$, where $m \leq k$, such that $\exp(\beta_{0i}/\tau) = \exp(\beta_{0i'}^*/\tau^*)$ for any $i, i' \in J_j$ and $j \in [k_*]$. On the other hand, when i and i' do not belong to the same set J_j, we let $\exp(\beta_{0i}/\tau) \neq \exp(\beta_{0i'}/\tau^*)$. Thus, we can reformulate equation (98) as

$$\sum_{j=1}^{m}\sum_{i \in J_j} \exp\left(\frac{\beta_{0i}}{\tau}\right) \exp\left(\frac{\sigma((\beta_{1i})^\top X)}{\tau}\right) f(Y|(a_i)^\top X + b_i, \nu_i)$$
$$= \sum_{j=1}^{m}\sum_{i \in J_j} \exp\left(\frac{\beta_{0i}^*}{\tau^*}\right) \exp\left(\frac{\sigma((\beta_{1i}^*)^\top X)}{\tau^*}\right) f(Y|(a_i^*)^\top X + b_i^*, \nu_i^*),$$

for almost surely (X,Y). This results leads to $\{((a_i)^\top X + b_i, \nu_i) : i \in J_j\} \equiv \{((a_i^*)^\top X + b_i^*, \nu_i^*) : i \in J_j\}$, for almost surely X for any $j \in [m]$. Therefore, we have

$$\{(a_i, b_i, \nu_i) : i \in J_j\} \equiv \{(a_i^*, b_i^*, \nu_i^*) : i \in J_j\},$$

for any $j \in [m]$. As a consequence,

$$G = \sum_{j=1}^{m}\sum_{i \in J_j} \exp\left(\frac{\beta_{0i}}{\tau}\right)\delta_{(\beta_{1i},\tau,a_i,b_i,\nu_i)} = \sum_{j=1}^{m}\sum_{i \in J_j} \exp\left(\frac{\beta_{0i}}{\tau}\right)\delta_{(\beta_{1i}^*,\tau^*,a_i^*,b_i^*,\nu_i^*)} = G_*.$$

Hence, we reach the conclusion of this proposition. □

D. Experimental Details

In this appendix, we provide additional details regarding the experimental setups.

D.1. Expectation-Maximization (EM) Algorithm

Our approach for parameter estimation in the dense-to-sparse gating Gaussian MoE model employs an Expectation-Maximization (EM) algorithm. We present the Expectation-Maximization (EM) algorithm used for parameter estimation in the context of a Mixture of Experts (MoE) model. The derivation and formulation closely follow (Jordan & Xu, 1995).

The EM algorithm is an iterative optimization technique used for finding maximum likelihood estimates of parameters in models with latent variables. In the context of MoE, the latent variables correspond to the assignment of data points to specific experts. Denote the latent variables as Z_{ik}, indicating whether expert k is responsible for the observation i, i.e.,

$$Z_{ij} = \begin{cases} 1 & \text{if } Y_i \text{ is generated from the } j\text{th expert,} \\ 0 & \text{Otherwise.} \end{cases}$$

The algorithm alternates between two key steps: Expectation and Maximization.

D.1.1. Expectation Step

In the E-step, the algorithm estimates the posterior probabilities of the latent variables given the observed data and the current parameters. This involves calculating the responsibility of each expert for each data point. We denote the responsibility of the jth expert for (X_i, Y_i) at iteration t by

$$\begin{aligned} \tau_{ij}^{(t)} &= \mathbb{E}[Z_{i,j} | (X_i, Y_i); \Theta^{(t)}] \\ &= \frac{\text{Softmax}\left(\frac{\beta_{1j}^{(t)\top} X_i + \beta_{0j}^{(t)}}{\tau^{(t)}}\right) \cdot f\left(Y_i \,\middle|\, a_j^{(t)\top} X_i + b_j^{(t)}, \nu_j^{(t)}\right)}{\sum_{l=1}^{k} \text{Softmax}\left(\frac{\beta_{1l}^{(t)\top} X_i + \beta_{0l}^{(t)}}{\tau^{(t)}}\right) \cdot f\left(Y_i \,\middle|\, a_l^{(t)\top} X_i + b_l^{(t)}, \nu_l^{(t)}\right)}. \end{aligned}$$

Here, $\Theta^{(t)}$ represents the set of all parameters in the Mixture of Experts (MoE) model at iteration t.

D.1.2. Maximization Step

In the Maximization step, the model parameters are updated to maximize the expected log-likelihood obtained from the latent variable distribution derived in the Expectation step. This involves updating the parameters of both the expert models and the gating network based on the responsibilities computed in the E-step.

M-step for expert parameters. The update equations for expert parameters are given by:

$$\begin{aligned} \nu_j^{(t+1)} &= \frac{1}{\sum_{i=1}^{n} \tau_{ij}^{(t)}} \sum_{i=1}^{n} \tau_{ij}^{(t)} \left(Y_i - (a_j^{(t)\top} X_i + b_j^{(t)})\right)^2, \\ \theta_j^{(t+1)} &= \left(\sum_{i=1}^{n} \tau_{ij}^{(t)} \cdot \frac{\widetilde{X}_i \widetilde{X}_i^\top}{\nu_j^{(t)}}\right)^{-1} \left(\sum_{i=1}^{n} \tau_{ij}^{(t)} \cdot \frac{Y_i \widetilde{X}_i}{\nu_j^{(t)}}\right), \end{aligned}$$

where $\widetilde{X}_i^\top = (X_i^\top, 1)$, and $\theta_j^{(t)\top} = (a_j^{(t)\top}, b_j^{(t)})$ for $j = 1, 2, ..., k$.

M-step for gating parameters. Finally, the update for gating parameters can be viewed as a specific form of a generalized linear model, specifically a multinomial logit model, as observed by (Jordan & Jacobs, 1994). Efficient fitting of such models is achieved through iteratively reweighted least squares (IRLS), a variant of Newton's method. First, let's denote $\theta_{0j}^{(t)\top} := (\beta_{1j}^{(t)\top}, \beta_{0j}^{(t)}, \tau^{(t)})$. Then, the update rule for the gating network parameters is given by:

$$\theta_{0j}^{(t+1)} = \theta_{0j}^{(t)} + \eta (R_j^{(t)})^{-1} e_j^{(t)},$$

where

$$
\begin{aligned}
e_j^{(t)} &= \sum_{i=1}^{n}\sum_{j=1}^{k}\left(\tau_{ij}^{(t)} - \mathrm{Softmax}\left(\frac{\beta_{1j}^{(t)\top}X_i + \beta_{0j}^{(t)}}{\tau^{(t)}}\right)\right) g_{ij}^{(t)},\\
R_j^{(t)} &= \sum_{i=1}^{n}\sum_{j=1}^{k}\mathrm{Softmax}\left(\frac{\beta_{1j}^{(t)\top}X_i + \beta_{0j}^{(t)}}{\tau^{(t)}}\right)\left(1 - \mathrm{Softmax}\left(\frac{\beta_{1j}^{(t)\top}X_i + \beta_{0j}^{(t)}}{\tau^{(t)}}\right)\right) g_{ij}^{(t)} g_{ij}^{(t)\top},\\
g_{ij}^{(t)} &= \nabla_{\theta_{0j}}\left(\frac{\beta_{1j}^{(t)\top}X_i + \beta_{0j}^{(t)}}{\tau^{(t)}}\right) = \left(X_i^{\top}, 1, -\frac{\beta_{1j}^{(t)\top}X_i + \beta_{0j}^{(t)}}{(\tau^{(t)})^2}\right)^{\top}.
\end{aligned}
$$

These two steps are iteratively repeated until convergence, providing a framework for estimating the model parameters that maximize the likelihood of the observed data. Furthermore, it is noteworthy to highlight that we choose the convergence criterion as $\epsilon = 10^{-6}$ and execute a maximum of 1000 iterations for the EM algorithm, with an 100 iterations for the Iteratively Reweighted Least Squares (IRLS) algorithm at each EM iteration, employing a learning rate of $\eta = 0.01$.

D.2. Experimental Setup

Synthetic Data. We conducted experiments using synthetic datasets generated with the true mixing measure $G_* = \sum_{i=1}^{2}\exp(\beta_{0i}^*/\tau^*)\delta_{(\beta_{1i}^*,\tau^*,a_i^*,b_i^*,\nu_i^*)}$ of order $k^* = 2$. We generated i.i.d samples $\{(X_i, Y_i)\}_{i=1}^{n}$ by initially sampling X_i values from a zero-mean Gaussian distribution with unit variance, followed by sampling Y_i values from the true conditional density $g_{G_*}(Y|X)$ of a Gaussian mixture with softmax gating, comprising $k_* = 2$ experts:

$$
g_{G_*}(Y|X) = \sum_{i=1}^{2}\mathrm{Softmax}\Big(\frac{(\beta_{1i}^*)^{\top}X + \beta_{0i}^*}{\tau^*}\Big)\cdot f(Y|(a_i^*)^{\top}X + b_i^*, \nu_i^*). \tag{99}
$$

The values corresponding to the true parameters are detailed in Table 2.

	Gating parameters	Expert parameters
Expert 1	$(\beta_{01}^*, \beta_{11}^*, \tau^*) = (0, -10, 0.5)$	$(a_1^*, b_1^*, \nu_1^*) = (-1, 2, 0.3)$
Expert 2	$(\beta_{02}^*, \beta_{12}^*, \tau^*) = (0, 0, 0.5)$	$(a_2^*, b_2^*, \nu_2^*) = (1, 2, 0.4)$

Table 2. Parameter values of the true model.

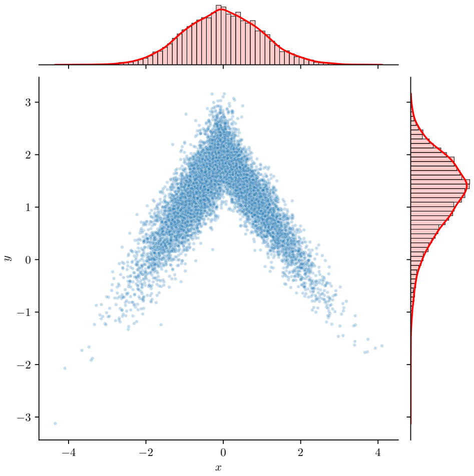


Figure 2. A visual illustration depicting the correlation between variables X and Y, along with their individual marginal distributions.

Initialization. For each $k \in \{k_*, k_* + 1\}$, elements from the set $\{1, 2, ..., k\}$ are randomly distributed among k_* Voronoi cells denoted as $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$, ensuring that each cell contains at least one element. This process is repeated for each replication. Subsequently, for each $j \in [k]$, all parameters are initialized by sampling from a Gaussian distribution centered around their true counterparts with a small variance, where $i \in \mathcal{C}_j$.