Supplement to "Statistical Perspective of Top-K Sparse Softmax Gating Mixture of Experts"

In this supplementary material, we first provide rigorous proofs for all results under the exact-specified settings in Appendix A, while those for the over-specified settings are then presented in Appendix B. Next, we study the identifiability of the top-K sparse softmax gating Gaussian mixture of experts (MoE) in Appendix C. We then carry out several numerical experiments in Appendix D to empirically justify our theoretical results. Finally, we establish the theories for parameter and density estimation beyond the settings of top-K sparse softmax Gaussian MoE in Appendix E.

A Proof for Results under the Exact-specified Settings

In this appendix, we present the proofs for Theorem 1 in Appendix A.1, while that for Theorem 2 is then given in Appendix A.2. Lastly, the proof of Lemma 1 is provided in Appendix A.3.

A.1 Proof of Theorem 1

In this appendix, we conduct a convergence analysis for density estimation in the top-K sparse softmax gating Gaussian MoE using proof techniques in (van de Geer, 2000). For that purpose, it is necessary to introduce some essential notations and key results first.

Let $\mathcal{P}_{k_*}(\Theta) := \{g_G(Y|X) : G \in \mathcal{E}_{k_*}(\Omega)\}$ be the set of all conditional density functions of mixing measures in $\mathcal{E}_{k_*}(\Omega)$. Next, we denote by $N(\varepsilon, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1)$ the covering number of metric space $(\mathcal{P}_{k_*}(\Omega), \|\cdot\|_1)$. Meanwhile, $H_B(\varepsilon, \mathcal{P}_{k_*}(\Omega), h)$ represents for the bracketing entropy of $\mathcal{P}_{k_*}(\Omega)$ under the Hellinger distance. Then, we provide in the following lemma the upper bounds of those terms.

Lemma 3. *If Ω is a bounded set, then the following inequalities hold for any $0 < \eta < 1/2$:*

(i) $\log N(\eta, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1) \lesssim \log(1/\eta)$*;*

(ii) $H_B(\eta, \mathcal{P}_{k_*}(\Omega), h) \lesssim \log(1/\eta)$.

Proof of Lemma 3 is in Appendix A.1.2. Subsequently, we denote

$$\widetilde{\mathcal{P}}_{k_*}(\Omega) := \{g_{(G+G_*)/2}(Y|X) : G \in \mathcal{E}_{k_*}(\Omega)\};$$
$$\widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega) := \{g^{1/2}_{(G+G_*)/2}(Y|X) : G \in \mathcal{E}_{k_*}(\Omega)\}.$$

In addition, for each $\delta > 0$, we define a Hellinger ball centered around the conditional density function $g_{G_*}(Y|X)$ and met with the set $\widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega)$ as

$$\widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega, \delta) := \{g^{1/2} \in \widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega) : h(g, g_{G_*}) \leq \delta\}.$$

To capture the size of the above Hellinger ball, van de Geer (2000) suggest using the following quantity:

$$\mathcal{J}_B(\delta, \widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega, \delta)) := \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(t, \widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega, t), \|\cdot\|)\mathrm{d}t \vee \delta, \tag{9}$$

where $t \vee \delta := \max\{t, \delta\}$. Given those notations, let us recall a standard result for density estimation in van de Geer (2000).

Lemma 4 (Theorem 7.4, van de Geer (2000))**.** *Take $\Psi(\delta) \geq \mathcal{J}_B(\delta, \widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega, \delta))$ such that $\Psi(\delta)/\delta^2$ is a non-increasing function of δ. Then, for some sequence (δ_n) and universal constant c which satisfy $\sqrt{n}\delta_n^2 \geq c\Psi(\delta)$, we obtain that*

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

for any $\delta \geq \delta_n$

Proof of Lemma 4 can be found in van de Geer (2000). Now, we are ready to provide the proof for convergence rate of density estimation in Theorem 1 in Appendix A.1.1.

A.1.1 Main Proof

It is worth noting that for any $t > 0$, we have

$$H_B(t, \widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega, t), \|\cdot\|) \leq H_B(t, \mathcal{P}_{k_*}(\Omega, t), h).$$

Then, the integral in equation (9) is upper bounded as follows:

$$\mathcal{J}_B(\delta, \widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega, \delta)) \leq \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(t, \mathcal{P}_{k_*}(\Omega, t), h)\mathrm{d}t \vee \delta \lesssim \int_{\delta^2/2^{13}}^{\delta} \log(1/t)\mathrm{d}t \vee \delta, \tag{10}$$

where the second inequality follows from part (ii) of Lemma 3.

As a result, by choosing $\Psi(\delta) = \delta \cdot \sqrt{\log(1/\delta)}$, we can verify that $\Psi(\delta)/\delta^2$ is a non-increasing function of δ. Furthermore, the inequality in equation (10) indicates that $\Psi(\delta) \geq \mathcal{J}_B(\delta, \widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega, \delta))$. Next, let us consider a sequence (δ_n) defined as $\delta_n := \sqrt{\log(n)/n}$. This sequence can be validated to satisfy the condition $\sqrt{n}\delta_n^2 \geq c\Psi(\delta)$ for some universal constant c. Therefore, by Lemma 4, we reach the conclusion of Theorem 1:

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

for some universal constant C depending only on Ω.

A.1.2 Proof of Lemma 3

Part (i). In this part, we will derive the following upper bound for the covering number of metric space $(\mathcal{P}_{k_*}(\Omega), \|\cdot\|_1)$ for any $0 < \eta < 1/2$ given the bounded set Ω:

$$\log N(\eta, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1) \lesssim \log(1/\eta).$$

To begin with, we define $\Theta := \{(a, b, \sigma) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+ : (\beta_0, \beta_1, a, b, \sigma) \in \Omega\}$. Note that Ω is a bounded set, then Θ also admits this property. Thus, there exists an η-cover of Θ, denoted by $\overline{\Theta}_\eta$. Additionally, we also define $\Delta := \{(\beta_0, \beta_1) \in \mathbb{R} \times \mathbb{R}^d : (\beta_0, \beta_1, a, b, \sigma) \in \Omega\}$, and $\overline{\Delta}_\eta$ be an η-cover of Δ. Then, it can be validated that $|\overline{\Theta}_\eta| \leq \mathcal{O}(\eta^{-(d+1)k_*})$ and $|\overline{\Delta}_\eta| \leq \mathcal{O}(\eta^{-(d+3)k_*})$.

Subsequently, for each $G = \sum_{i=1}^{k_*} \exp(\beta_{0i})\delta_{(\beta_{1i}, a_i, b_i, \sigma_i)} \in \mathcal{E}_{k_*}(\Omega)$, we take into account two other mixing measures. The first measure is $G' = \sum_{i=1}^{k_*} \exp(\beta_{0i})\delta_{(\beta_{1i}, \overline{a}_i, \overline{b}_i, \overline{\sigma}_i)}$, where $(\overline{a}_i, \overline{b}_i, \overline{\sigma}_i) \in \overline{\Theta}_\eta$ is the closest points to (a_i, b_i, σ_i) in this set for all $i \in [k_*]$. The second one is $\overline{G} := \sum_{i=1}^{k_*} \exp(\overline{\beta}_{0i})\delta_{(\overline{\beta}_{1i}, \overline{a}_i, \overline{b}_i, \overline{\sigma}_i)}$ in which $(\overline{\beta}_{0i}, \overline{\beta}_{1i}) \in \overline{\Delta}_\eta$ for any $i \in [k_*]$. Next, let us define

$$\mathcal{T} := \{g_{\overline{G}} \in \mathcal{P}_{k_*}(\Omega) : (\overline{\beta}_{0i}, \overline{\beta}_{1i}) \in \overline{\Delta}_\eta,\ (\overline{a}_i, \overline{b}_i, \overline{\sigma}_i) \in \overline{\Theta}_\eta, \forall i \in [k_*]\},$$

then it is obvious that $g_{\overline{G}} \in \mathcal{T}$. Now, we will show that $\mathcal{T}$ is an η-cover of metric space $(\mathcal{P}_{k_*}(\Omega), \|\cdot\|_1)$ with a note that it is not necessarily the smallest cover. Indeed, according to the triangle inequality,

$$\|g_G - g_{\overline{G}}\|_1 \leq \|g_G - g_{G'}\|_1 + \|g_{G'} - g_{\overline{G}}\|_1. \tag{11}$$

The first term in the right hand side can be upper bounded as follows:

$$\begin{aligned}
\|g_G - g_{G'}\|_1 &\leq \sum_{i=1}^{k_*} \int_{\mathcal{X}\times\mathcal{Y}} \Big| f(Y|a_i^\top X + b_i, \sigma_i) - f(Y|\overline{a}_i^\top X + \overline{b}_i, \overline{\sigma}_i) \Big| \mathrm{d}(X, Y) \\
&\lesssim \sum_{i=1}^{k_*} \int_{\mathcal{X}\times\mathcal{Y}} \Big(\|a_i - \overline{a}_i\| + \|b_i - \overline{b}_i\| + \|\sigma_i - \overline{\sigma}_i\| \Big) \mathrm{d}(X, Y) \\
&= \sum_{i=1}^{k_*} \Big(\|a_i - \overline{a}_i\| + \|b_i - \overline{b}_i\| + \|\sigma_i - \overline{\sigma}_i\| \Big) \\
&\lesssim \eta.
\end{aligned} \tag{12}$$

Next, we will also demonstrate that $\|g_{G'} - g_{\overline{G}}\|_1 \lesssim \eta$. For that purpose, let us consider $q := \binom{k}{K}$ K-element subsets of $\{1, \ldots, k\}$, which are assumed to take the form $\{\ell_1, \ell_2, \ldots, \ell_K\}$ for any $\ell \in [q]$.

Additionally, we also denote $\{\ell_{K+1},\dots,\ell_k\} := \{1,\dots,k\} \setminus \{\ell_1,\dots,\ell_K\}$ for any $\ell \in [q]$. Then, we define

$$\mathcal{X}_\ell := \{x \in \mathcal{X} : \beta_{1i}^\top x \geq \beta_{1i'}^\top x : i \in \{\ell_1,\dots,\ell_K\}, i' \in \{\ell_{K+1},\dots,\ell_{k_*}\}\},$$
$$\widetilde{\mathcal{X}}_\ell := \{x \in \mathcal{X} : \overline{\beta}_{1i}^\top x \geq \overline{\beta}_{1i'}^\top x : i \in \{\ell_1,\dots,\ell_K\}, i' \in \{\ell_{K+1},\dots,\ell_{k_*}\}\}.$$

By using the same arguments as in the proof of Lemma 1 in Appendix A.3, we achieve that either $\mathcal{X}_\ell = \widetilde{\mathcal{X}}_\ell$ or $\mathcal{X}_\ell$ has measure zero for any $\ell \in [q]$. As the $\mathrm{Softmax}$ function is differentiable, it is a Lipschitz function with some Lipschitz constant $L \geq 0$. Since $\mathcal{X}$ is a bounded set, we may assume that $\|X\| \leq B$ for any $X \in \mathcal{X}$. Next, we denote

$$\pi_\ell(X) := \Big(\beta_{1\ell_i}^\top x + \beta_{0\ell_i}^\top\Big)_{i=1}^K; \qquad \overline{\pi}_\ell(X) := \Big(\overline{\beta}_{1\ell_i}^\top x + \overline{\beta}_{0\ell_i}^\top\Big)_{i=1}^K,$$

for any K-element subset $\{\ell_1,\dots\ell_K\}$ of $\{1,\dots,k_*\}$. Then, we get

$$\begin{aligned}\|\mathrm{Softmax}(\pi_\ell(X)) - \mathrm{Softmax}(\overline{\pi}_\ell(X))\| &\leq L \cdot \|\pi_\ell(X) - \overline{\pi}_\ell(X)\| \\ &\leq L \cdot \sum_{i=1}^K \Big(\|\beta_{1\ell_i} - \overline{\beta}_{1\ell_i}\| \cdot \|X\| + |\beta_{0\ell_i} - \overline{\beta}_{0\ell_i}|\Big) \\ &\leq L \cdot \sum_{i=1}^K \Big(\eta B + \eta\Big) \\ &\lesssim \eta.\end{aligned}$$

Back to the proof for $\|g_{G'} - g_{\overline{G}}\|_1 \lesssim \eta$, it follows from the above results that

$$\begin{aligned}&\|g_{G'} - g_{\overline{G}}\|_1 = \int_{\mathcal{X}\times\mathcal{Y}} |g_{G'}(Y|X) - g_{\overline{G}}(Y|X)| \,\mathrm{d}(X,Y) \\ &\leq \sum_{\ell=1}^q \int_{\mathcal{X}_\ell\times\mathcal{Y}} |g_{G'}(Y|X) - g_{\overline{G}}(Y|X)| \,\mathrm{d}(X,Y) \\ &\leq \sum_{\ell=1}^q \int_{\mathcal{X}_\ell\times\mathcal{Y}} \sum_{i=1}^K \Big|\mathrm{Softmax}(\pi_\ell(X)_i) - \mathrm{Softmax}(\overline{\pi}_\ell(X)_i)\Big| \cdot \Big|f(Y|\overline{a}_{\ell_i}^\top X + \overline{b}_{\ell_i}, \overline{\sigma}_{\ell_i})\Big| \,\mathrm{d}(X,Y) \\ &\lesssim \eta, \qquad (13)\end{aligned}$$

From the results in equations (11), (12) and (13), we deduce that $\|g_G - g_{\overline{G}}\|_1 \lesssim \eta$. This implies that $\mathcal{T}$ is an η-cover of the metric space $(\mathcal{P}_{k_*}(\Omega), \|\cdot\|_1)$. Consequently, we achieve that

$$N(\eta, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1) \lesssim |\overline{\Delta}_\eta| \times |\overline{\Theta}_\eta| \leq \mathcal{O}(1/\eta^{(d+3)k}),$$

which induces the conclusion of this part

$$\log N(\eta, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1) \lesssim \log(1/\eta).$$

Part (ii). Moving to this part, we will provide an upper bound for the bracketing entropy of $\mathcal{P}_{k_*}(\Omega)$ under the Hellinger distance:

$$H_B(\eta, \mathcal{P}_{k_*}(\Omega), h) \lesssim \log(1/\eta).$$

Recall that Θ and $\mathcal{X}$ are bounded sets, we can find positive constants $-\gamma \leq a^\top X + b \leq \gamma$ and $u_1 \leq \sigma \leq u_2$. Let us define

$$Q(Y|X) := \begin{cases} \frac{1}{\sqrt{2\pi u_1}} \exp\Big(-\frac{Y^2}{8u_2}\Big), & \text{for } |Y| \geq 2\gamma \\ \frac{1}{\sqrt{2\pi u_1}}, & \text{for } |Y| < 2\gamma \end{cases}$$

Then, it can be validated that $f(Y|a^\top X + b, \sigma) \leq Q(X,Y)$ for any $(X,Y) \in \mathcal{X}\times\mathcal{Y}$.

Next, let $\tau \leq \eta$ which will be chosen later and $\{g_1,\dots,g_N\}$ be an τ-cover of metric space $(\mathcal{P}_{k_*}(\Omega), \|\cdot\|_1)$ with the covering number $N := N(\tau, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1)$. Additionally, we also consider brackets of the form $[\Psi_i^L(Y|X), \Psi_i^U(Y|X)]$ where

$$\Psi_i^L(Y|X) := \max\{g_i(Y|X) - \tau, 0\}$$
$$\Psi_i^U(Y|X) := \max\{g_i(Y|X) + \tau, Q(Y|X)\}.$$

Then, we can check that $\mathcal{P}_{k_*}(\Omega) \subseteq \bigcup_{i=1}^{N}[\Psi_i^L(Y|X), \Psi_i^U(Y|X)]$ and $\Psi_i^U(Y|X) - \Psi_i^L(Y|X) \leq \min\{2\eta, Q(Y|X)\}$. Let $S := \max\{2\gamma, \sqrt{8u_2}\}\log(1/\tau)$, we have for any $i \in [N]$ that

$$\|\Psi_i^U - \Psi_i^L\|_1 = \int_{|Y|<2\gamma}[\Psi_i^U(Y|X) - \Psi_i^L(Y|X)]\,\mathrm{d}X\mathrm{d}Y + \int_{|Y|\geq 2\gamma}[\Psi_i^U(Y|X) - \Psi_i^L(Y|X)]\,\mathrm{d}X\mathrm{d}Y$$

$$\leq S\tau + \exp\Big(-\frac{S^2}{2u_2}\Big) \leq S'\tau,$$

where S' is some positive constant. This inequality indicates that

$$H_B(S'\tau, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1) \leq \log N(\tau, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1) \leq \log(1/\tau).$$

By setting $\tau = \eta/S'$, we obtain that $H_B(\eta, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1) \lesssim \log(1/\eta)$. Finally, since the norm $\|\cdot\|_1$ is upper bounded by the Hellinger distance, we reach the conclusion of this part:

$$H_B(\eta, \mathcal{P}_{k_*}(\Omega), h) \lesssim \log(1/\eta).$$

Hence, the proof is completed.

A.2 Proof of Theorem 2

Since the Hellinger distance is lower bounded by the Total Variation distance, that is $h \geq V$, we will prove the following Total Variation lower bound:

$$\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_1(G, G_*),$$

which is then respectively broken into local part and global part as follows:

$$\inf_{G\in\mathcal{E}_{k_*}(\Omega):\mathcal{D}_1(G,G_*)\leq\varepsilon'} \frac{\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_1(G, G_*)} > 0, \tag{14}$$

$$\inf_{G\in\mathcal{E}_{k_*}(\Omega):\mathcal{D}_1(G,G_*)>\varepsilon'} \frac{\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_1(G, G_*)} > 0, \tag{15}$$

for some constant $\varepsilon' > 0$.

Proof of claim (14): It is sufficient to show that

$$\lim_{\varepsilon\to 0}\inf_{G\in\mathcal{E}_{k_*}(\Omega):\mathcal{D}_1(G,G_*)\leq\varepsilon} \frac{\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_1(G, G_*)} > 0.$$

Assume that this inequality does not hold, then since the number of experts k_* is known in this case, there exists a sequence of mixing measure $G_n := \sum_{i=1}^{k_*}\exp(\beta_{0i}^n)\delta_{(\beta_{1i}^n, a_i^n, b_i^n, \sigma_i^n)} \in \mathcal{E}_{k_*}(\Omega)$ such that both $\mathcal{D}_1(G_n, G_*)$ and $\mathbb{E}_X[V(g_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_1(G_n, G_*)$ approach zero as n tends to infinity. Now, we define

$$\mathcal{C}_j^n = \mathcal{C}_j(G_n) := \{i \in [k_*] : \|\omega_i^n - \omega_j^*\| \leq \|\omega_i^n - \omega_s^*\|, \ \forall s \neq j\},$$

for any $j \in [k_*]$ as k_* Voronoi cells with respect to the mixing measure G_n, where we denote $\omega_i^n := (\beta_{1i}^n, a_i^n, b_i^n, \sigma_i^n)$ and $\omega_j^* := (\beta_{1j}^*, a_j^*, b_j^*, \sigma_j^*)$. As we use asymptotic arguments in this proof, we can assume without loss of generality (WLOG) that these Voronoi cells does not depend on n, that is, $\mathcal{C}_j = \mathcal{C}_j^n$. Next, it follows from the hypothesis $\mathcal{D}_1(G_n, G_*) \to 0$ as $n \to \infty$ that each Voronoi cell contains only one element. Thus, we continue to assume WLOG that $\mathcal{C}_j = \{j\}$ for any $j \in [k_*]$, which implies that $(\beta_{1j}^n, a_j^n, b_j^n, \sigma_j^n) \to (\beta_{1j}^*, a_j^*, b_j^*, \sigma_j^*)$ and $\exp(\beta_{0j}^n) \to \exp(\beta_{0j}^*)$ as $n \to \infty$.

Subsequently, to specify the top K selection in the formulations of $g_{G_n}(Y|X)$ and $g_{G_*}(Y|X)$, we divide the covariate space $\mathcal{X}$ into some subsets in two ways. In particular, we first consider $q := \binom{k_*}{K}$ different K-element subsets of $[k_*]$, which are assumed to take the form $\{\ell_1, \ldots, \ell_K\}$, for $\ell \in [q]$. Additionally, we denote $\{\ell_{K+1}, \ldots, \ell_{k_*}\} := [k_*] \setminus \{\ell_1, \ldots, \ell_K\}$. Then, we define for each $\ell \in [q]$ two following subsets of $\mathcal{X}$:

$$\mathcal{X}_\ell^n := \Big\{x \in \mathcal{X} : (\beta_{1j}^n)^\top x \geq (\beta_{1j'}^n)^\top x : \forall j \in \{\ell_1, \ldots, \ell_K\}, j' \in \{\ell_{K+1}, \ldots, \ell_{k_*}\}\Big\},$$

$$\mathcal{X}_\ell^* := \Big\{x \in \mathcal{X} : (\beta_{1j}^*)^\top x \geq (\beta_{1j'}^*)^\top x : \forall j \in \{\ell_1, \ldots, \ell_K\}, j' \in \{\ell_{K+1}, \ldots, \ell_{k_*}\}\Big\}.$$

Since $(\beta_{0j}^n, \beta_{1j}^n) \to (\beta_{0j}^*, \beta_{1j}^*)$ as $n \to \infty$ for any $j \in [k_*]$, we have for any arbitrarily small $\eta_j > 0$ that $\|\beta_{1j}^n - \beta_{1j}^*\| \leq \eta_j$ and $|\beta_{0j}^n - \beta_{0j}^*| \leq \eta_j$ for sufficiently large n. By applying Lemma 1, we obtain that $\mathcal{X}_\ell^n = \mathcal{X}_\ell^*$ for any $\ell \in [q]$ for sufficiently large n. WLOG, we assume that

$$\mathcal{D}_1(G_n, G_*) = \sum_{i=1}^{K} \Big[\exp(\beta_{0i}^n)\Big(\|\Delta\beta_{1i}^n\| + \|\Delta a_i^n\| + \|\Delta b_i^n\| + \|\Delta\sigma_i^n\|\Big) + \Big|\exp(\beta_{0i}^n) - \exp(\beta_{0i}^*)\Big|\Big],$$

where we denote $\Delta\beta_{1i}^n := \beta_{1i}^n - \beta_{1i}^*$, $\Delta a_i^n := a_i^n - a_i^*$, $\Delta b_i^n := b_i^n - b_i^*$ and $\Delta\sigma_i^n := \sigma_i^n - \sigma_i^*$.

Let $\ell \in [q]$ such that $\{\ell_1, \ldots, \ell_K\} = \{1, \ldots, K\}$. Then, for almost surely $(X, Y) \in \mathcal{X}_\ell^* \times \mathcal{Y}$, we can rewrite the conditional densities $g_{G_n}(Y|X)$ and $g_{G_*}(Y|X)$ as

$$g_{G_n}(Y|X) = \sum_{i=1}^{K} \frac{\exp((\beta_{1i}^n)^\top X + \beta_{0i}^n)}{\sum_{j=1}^{K} \exp((\beta_{1j}^n)^\top X + \beta_{0j}^n)} \cdot f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n),$$

$$g_{G_*}(Y|X) = \sum_{i=1}^{K} \frac{\exp((\beta_{1i}^*)^\top X + \beta_{0i}^*)}{\sum_{j=1}^{K} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)} \cdot f(Y|(a_i^*)^\top X + b_i^*, \sigma_i^*).$$

Now, we break the rest of our arguments into three steps:

Step 1 - Taylor expansion:

In this step, we take into account $H_n := \Big[\sum_{i=1}^{K} \exp((\beta_{1i}^*)^\top X + \beta_{0i}^*)\Big] \cdot [g_{G_n}(Y|X) - g_{G_*}(Y|X)]$. Then, H_n can be represented as follows:

$$H_n = \sum_{i=1}^{K} \exp(\beta_{0i}^n)\Big[\exp((\beta_{1i}^n)^\top X) f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n) - \exp((\beta_{1i}^*)^\top X) f(Y|(a_i^*)^\top X + b_i^*, \sigma_i^*)\Big]$$

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

$$+ \sum_{i=1}^{K} \Big[\exp(\beta_{0i}^n) - \exp(\beta_{0i}^*)\Big]\Big[\exp((\beta_{1i}^*)^\top X) f(Y|(a_i^*)^\top X + b_i^*, \sigma_i^*) - \exp((\beta_{1i}^*)^\top X) g_{G_n}(Y|X)\Big].$$

By applying the first-order Taylor expansion to the first term in the above representation, which is denoted by A_n, we get that

$$A_n = \sum_{i=1}^{K} \sum_{|\alpha|=1} \frac{\exp(\beta_{0i}^n)}{\alpha!} \cdot (\Delta\beta_{1i}^n)^{\alpha_1} (\Delta a_i^n)^{\alpha_2} (\Delta b_i^n)^{\alpha_3} (\Delta\sigma_i^n)^{\alpha_4}$$
$$\times X^{\alpha_1+\alpha_2} \exp((\beta_{1i}^*)^\top X) \cdot \frac{\partial^{|\alpha_2|+\alpha_3+\alpha_4} f}{\partial h_1^{|\alpha_2|+\alpha_3} \partial\sigma^{\alpha_4}}(Y|(a_i^*)^\top X + b_i^*, \sigma_i^*) + R_1(X, Y),$$

where $R_1(X, Y)$ is a Taylor remainder that satisfies $R_1(X, Y)/\mathcal{D}_1'(X, Y) \to 0$ as $n \to \infty$. Recall that f is the univariate Gaussian density, then we have

$$\frac{\partial^{\alpha_4} f}{\partial\sigma^{\alpha_4}}(Y|(a_i^*)^\top X + b_i^*, \sigma_i^*) = \frac{1}{2^{\alpha_4}} \cdot \frac{\partial^{2\alpha_4} f}{\partial h_1^{2\alpha_4}}(Y|(a_i^*)^\top X + b_i^*, \sigma_i^*),$$

which leads to

$$A_n = \sum_{i=1}^{K} \sum_{|\alpha|=1} \frac{\exp(\beta_{0i}^n)}{2^{\alpha_4}\alpha!} \cdot (\Delta\beta_{1i}^n)^{\alpha_1} (\Delta a_i^n)^{\alpha_2} (\Delta b_i^n)^{\alpha_3} (\Delta\sigma_i^n)^{\alpha_4}$$
$$\times X^{\alpha_1+\alpha_2} \exp((\beta_{1i}^*)^\top X) \cdot \frac{\partial^{|\alpha_2|+\alpha_3+2\alpha_4} f}{\partial h_1^{|\alpha_2|+\alpha_3+2\alpha_4}}(Y|(a_i^*)^\top X + b_i^*, \sigma_i^*) + R_1(X, Y)$$
$$= \sum_{i=1}^{K} \sum_{|\eta_1|+\eta_2=1}^{2} \sum_{\alpha \in \mathcal{J}_{\eta_1,\eta_2}} \frac{\exp(\beta_{0i}^n)}{2^{\alpha_4}\alpha!} \cdot (\Delta\beta_{1i}^n)^{\alpha_1} (\Delta a_i^n)^{\alpha_2} (\Delta b_i^n)^{\alpha_3} (\Delta\sigma_i^n)^{\alpha_4}$$
$$\times X^{\eta_1} \exp((\beta_{1i}^*)^\top X) \cdot \frac{\partial^{\eta_2} f}{\partial h_1^{\eta_2}}(Y|(a_i^*)^\top X + b_i^*, \sigma_i^*) + R_1(X, Y), \quad (16)$$

where we denote $\eta_1 = \alpha_1 + \alpha_2 \in \mathbb{N}^d$, $\eta_2 = |\alpha_2| + \alpha_3 + 2\alpha_4 \in \mathbb{N}$ and an index set

$$\mathcal{J}_{\eta_1,\eta_2} := \{(\alpha_i)_{i=1}^4 \in \mathbb{N}^d \times \mathbb{N}^d \times \mathbb{N} \times \mathbb{N} : \alpha_1 + \alpha_2 = \eta_1,\ \alpha_3 + 2\alpha_4 = \eta_2 - |\alpha_2|\}. \tag{17}$$

By arguing in a similar fashion for the second term in the representation of H_n, we also get that

$$B_n := -\sum_{i=1}^{K} \sum_{|\gamma|=1} \frac{\exp(\beta_{0i}^n)}{\gamma!} (\Delta\beta_{1i}^n) \cdot X^{\gamma} \exp((\beta_{1i}^n)^\top X) g_{G_n}(Y|X) + R_2(X,Y),$$

where $R_2(X,Y)$ is a Taylor remainder such that $R_2(X,Y)/\mathcal{D}_1(G_n,G_*) \to 0$ as $n \to \infty$. Putting the above results together, we rewrite the quantity H_n as follows:

$$\begin{aligned} H_n = &\sum_{i=1}^{K} \sum_{0 \leq |\eta_1| + \eta_2 \leq 2} U_{i,\eta_1,\eta_2}^n \cdot X^{\eta_1} \exp((\beta_{1i}^*)^\top X) \frac{\partial^{\eta_2} f}{\partial h_1^{\eta_2}} (Y|(a_i^*)^\top X + b_i^*, \sigma_i^*) \\ &+ \sum_{i=1}^{K} \sum_{0 \leq |\gamma| \leq 1} W_{i,\gamma}^n \cdot X^{\gamma} \exp((\beta_{1i}^*)^\top X) g_{G_n}(Y|X) + R_1(X,Y) + R_2(X,Y), \end{aligned} \tag{18}$$

in which we respectively define for each $i \in [K]$ that

$$U_{i,\eta_1,\eta_2}^n := \sum_{\alpha \in \mathcal{J}_{\eta_1,\eta_2}} \frac{\exp(\beta_{0i}^n)}{2^{\alpha_4} \alpha!} \cdot (\Delta\beta_{1i}^n)^{\alpha_1} (\Delta a_i^n)^{\alpha_2} (\Delta b_i^n)^{\alpha_3} (\Delta\sigma_i^n)^{\alpha_4},$$

$$W_{i,\gamma}^n := -\frac{\exp(\beta_{0i}^n)}{\gamma!} (\Delta\beta_{1i}^n)^{\gamma},$$

for any $(\eta_1,\eta_2) \neq (\mathbf{0}_d, 0)$ and $|\gamma| \neq \mathbf{0}_d$. Otherwise, $U_{i,\mathbf{0}_d,0}^n = -W_{i,\mathbf{0}_d}^n := \exp(\beta_{0i}^n) - \exp(\beta_{0i}^*)$.

Step 2 - Non-vanishing coefficients:

Moving to the second step, we will show that not all the ratios $U_{i,\eta_1,\eta_2}^n/\mathcal{D}_1(G_n,G_*)$ tend to zero as n goes to infinity. Assume by contrary that all of them approach zero when $n \to \infty$, then for $(\eta_1,\eta_2) = (\mathbf{0}_d, 0)$, it follows that

$$\frac{1}{\mathcal{D}_1(G_n,G_*)} \cdot \sum_{i=1}^{K} \Big| \exp(\beta_{0i}^n) - \exp(\beta_{0i}^*) \Big| = \sum_{i=1}^{K} \frac{|U_{j,\eta_1,\eta_2}^n|}{\mathcal{D}_1(G_n,G_*)} \to 0. \tag{19}$$

Additionally, for tuples (η_1,η_2) where $\eta_1 \in \{e_1, e_2, \ldots, e_d\}$ with $e_j := (0,\ldots,0,\underbrace{1}_{j-th},0,\ldots,0)$ and $\eta_2 = 0$, we get

$$\frac{1}{\mathcal{D}_1(G_n,G_*)} \cdot \sum_{i=1}^{K} \exp(\beta_{0i}^n) \|\Delta\beta_{1i}^n\|_1 = \sum_{i=1}^{K} \frac{|U_{j,\eta_1,\eta_2}^n|}{\mathcal{D}_1(G_n,G_*)} \to 0.$$

By using similar arguments, we end up having

$$\frac{1}{\mathcal{D}_1(G_n,G_*)} \cdot \sum_{i=1}^{K} \exp(\beta_{0i}^n) \Big[\|\Delta\beta_{1i}^n\|_1 + \|\Delta a_i^n\|_1 + |\Delta b_i^n| + |\Delta\sigma_i^n| \Big] \to 0.$$

Due to the topological equivalence between norm-1 and norm-2, the above limit implies that

$$\frac{1}{\mathcal{D}_1(G_n,G_*)} \cdot \sum_{i=1}^{K} \exp(\beta_{0i}^n) \Big[\|\Delta\beta_{1i}^n\| + \|\Delta a_i^n\| + |\Delta b_i^n| + |\Delta\sigma_i^n| \Big] \to 0. \tag{20}$$

Combine equation (19) with equation (20), we deduce that $\mathcal{D}_1(G_n,G_*)/\mathcal{D}_1(G_n,G_*) \to 0$, which is a contradiction. Consequently, at least one among the ratios $U_{i,\eta_1,\eta_2}^n/\mathcal{D}_1(G_n,G_*)$ does not vanish as n tends to infinity.

Step 3 - Fatou's contradiction:

Let us denote by m_n the maximum of the absolute values of $U_{i,\eta_1,\eta_2}^n/\mathcal{D}_1(G_n,G_*)$ and $W_{i,\gamma}^n/\mathcal{D}_1(G_n,G_*)$. It follows from the result achieved in Step 2 that $1/m_n \not\to \infty$.

Recall from the hypothesis that $\mathbb{E}_X[V(g_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_1(G_n, G_*) \to 0$ as $n \to \infty$. Thus, by the Fatou's lemma, we have

$$0 = \lim_{n\to\infty} \frac{\mathbb{E}_X[V(g_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_1(G_n, G_*)} \geq \frac{1}{2} \cdot \int \liminf_{n\to\infty} \frac{|g_{G_n}(Y|X) - g_{G_*}(Y|X)|}{\mathcal{D}_1(G_n, G_*)} \mathrm{d}X\mathrm{d}Y.$$

This result indicates that $|g_{G_n}(Y|X) - g_{G_*}(Y|X)|/\mathcal{D}_1(G_n, G_*)$ tends to zero as n goes to infinity for almost surely (X, Y). As a result, it follows that

$$\lim_{n\to\infty} \frac{H_n}{m_n\mathcal{D}_1(G_n, G_*)} = \lim_{n\to\infty} \frac{|g_{G_n}(Y|X) - g_{G_*}(Y|X)|}{m_n\mathcal{D}_1(G_n, G_*)} = 0.$$

Next, let us denote $U^n_{i,\eta_1,\eta_2}/[m_n\mathcal{D}_1(G_n, G_*)] \to \tau_{i,\eta_1,\eta_2}$ and $W^n_{i,\gamma}/[m_n\mathcal{D}_1(G_n, G_*)] \to \kappa_{i,\gamma}$ with a note that at least one among them is non-zero. From the formulation of H_n in equation (18), we deduce that

$$\sum_{i=1}^{K} \sum_{0\leq|\eta_1|+\eta_2\leq 2} \tau_{i,\eta_1,\eta_2} \cdot X^{\eta_1} \exp((\beta^*_{1i})^\top X) \frac{\partial^{\eta_2} f}{\partial h_1^{\eta_2}}(Y|(a^*_i)^\top X + b^*_i, \sigma^*_i) + \sum_{i=1}^{K} \sum_{0\leq|\gamma|\leq 1} \kappa_{i,\gamma} \cdot X^{\gamma} \exp((\beta^*_{1i})^\top X) g_{G_*}(Y|X) = 0, \tag{21}$$

for almost surely (X, Y). This equation is equivalent to

$$\sum_{i=1}^{K} \sum_{0\leq|\eta_1|\leq 1} \left[\sum_{0\leq\eta_2\leq 2-|\gamma|} \tau_{i,\eta_1,\eta_2} \frac{\partial^{\eta_2} f}{\partial h_1^{\eta_2}}(Y|(a^*_i)^\top X + b^*_i, \sigma^*_i) + \kappa_{i,\eta_1} g_{G_*}(Y|X)\right] \times X^{\eta_1} \exp((\beta^*_{1i})^\top X) = 0,$$

for almost surely (X, Y). Note that $\beta^*_{11}, \ldots, \beta^*_{1K}$ admits pair-wise different values, then $\{\exp((\beta^*_{1i})^\top X) : i \in [K]\}$ is a linearly independent set, which leads to

$$\sum_{0\leq|\eta_1|\leq 1} \left[\sum_{0\leq\eta_2\leq 2-|\gamma|} \tau_{i,\eta_1,\eta_2} \frac{\partial^{\eta_2} f}{\partial h_1^{\eta_2}}(Y|(a^*_i)^\top X + b^*_i, \sigma^*_i) + \kappa_{i,\eta_1} g_{G_*}(Y|X)\right] X^{\eta_1} = 0,$$

for any $i \in [K]$ for almost surely (X, Y). It is clear that the left hand side of the above equation is a polynomial of X belonging to the compact set $\mathcal{X}$. As a result, we get that

$$\sum_{0\leq\eta_2\leq 2-|\gamma|} \tau_{i,\eta_1,\eta_2} \frac{\partial^{\eta_2} f}{\partial h_1^{\eta_2}}(Y|(a^*_i)^\top X + b^*_i, \sigma^*_i) + \kappa_{i,\eta_1} g_{G_*}(Y|X) = 0,$$

for any $i \in [K]$, $0 \leq |\eta_1| \leq 1$ and almost surely (X, Y). Since $(a^*_1, b^*_1, \sigma^*_1), \ldots, (a^*_K, b^*_K, \sigma^*_K)$ have pair-wise distinct values, those of $((a^*_1)^\top X + b^*_1, \sigma^*_1), \ldots, ((a^*_K)^\top X + b^*_K, \sigma^*_K)$ are also pair-wise different. Thus, the set $\left\{\frac{\partial^{\eta_2} f}{\partial h_1^{\eta_2}}(Y|(a^*_i)^\top X + b^*_i, \sigma^*_i),\ g_{G_*}(Y|X) : i \in [K]\right\}$ is linearly independent. Consequently, we obtain that $\tau_{i,\eta_1,\eta_2} = \kappa_{i,\gamma} = 0$ for any $i \in [K]$, $0 \leq |\eta_1| + \eta_2 \leq 2$ and $0 \leq |\gamma| \leq 1$, which contradicts the fact that at least one among these terms is different from zero.

Hence, we can find some constant $\varepsilon' > 0$ such that

$$\inf_{G\in\mathcal{E}_{k_*}(\Omega):\mathcal{D}_1(G,G_*)\leq\varepsilon'} \frac{\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_1(G, G_*)} > 0.$$

Proof of claim (15): Assume by contrary that this claim is not true, then we can seek a sequence $G'_n \in \mathcal{E}_{k_*}(\Omega)$ such that $\mathcal{D}_1(G'_n, G_*) > \varepsilon'$ and

$$\lim_{n\to\infty} \frac{\mathbb{E}_X[V(g_{G'_n}(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_1(G'_n, G_*)} = 0,$$

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 Ω 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 $\mathcal{D}_1(G'_n, G_*) > \varepsilon'$, this result induces that $\mathcal{D}_1(G', G_*) > \varepsilon'$.

Subsequently, by means of the Fatou's lemma, we achieve that

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

It follows that $g_{G'}(Y|X) = g_{G_*}(Y|X)$ for almost surely (X, Y). From Proposition 2, we know that the top-K sparse softmax gating Gaussian mixture of experts is identifiable, thus, we obtain that $G' \equiv G_*$. As a consequence, $\mathcal{D}_1(G', G_*) = 0$, contradicting the fact that $\mathcal{D}_1(G', G_*) > \varepsilon' > 0$.

Hence, the proof is completed.

A.3 Proof of Lemma 1

Let $\eta_i = M_i\varepsilon$, where ε is some fixed positive constant and M_i will be chosen later. For an arbitrary $\ell \in [q]$, since $\mathcal{X}$ and Ω are bounded sets, there exists some constant $c^*_\ell \geq 0$ such that

$$\min_{x,i,i'} \Big[(\beta^*_{1i})^\top x - (\beta^*_{1i'})^\top x \Big] = c^*_\ell \varepsilon, \tag{22}$$

where the minimum is subject to $x \in \mathcal{X}^*_\ell, i \in \{\ell_1, \ldots, \ell_K\}$ and $i' \in \{\ell_{K+1}, \ldots, \ell_{k_*}\}$. We will point out that $c^*_\ell > 0$. Assume by contrary that $c^*_\ell = 0$. For $x \in \mathcal{X}^*_\ell$, we may assume for any $1 \leq i < j \leq k_*$ that

$$(\beta^*_{1\ell_i})^\top x \geq (\beta^*_{1\ell_j})^\top x.$$

Since $c^*_\ell = 0$, it follows from equation (22) that $(\beta^*_{1\ell_K})^\top x - (\beta^*_{1\ell_{K+1}})^\top x = 0$, or equivalently

$$(\beta^*_{1\ell_K} - \beta^*_{1\ell_{K+1}})^\top x = 0.$$

In other words, $\mathcal{X}^*_\ell$ is a subset of

$$\mathcal{Z} := \{x \in \mathcal{X} : (\beta^*_{1\ell_K} - \beta^*_{1\ell_{K+1}})^\top x = 0\}.$$

Since $\beta_{1\ell_K} - \beta_{1\ell_{K+1}} \neq \mathbf{0}_d$ and the distribution of X is continuous, it follows that the set $\mathcal{Z}$ has measure zero. Since $\mathcal{X}^*_\ell \subseteq \mathcal{Z}$, we can conclude that $\mathcal{X}^*_\ell$ also has measure zero, which contradicts the hypothesis of Lemma 1. Therefore, we must have $c^*_\ell > 0$.

As $\mathcal{X}$ is a bounded set, we assume that $\|x\| \leq B$ for any $x \in \mathcal{X}$. Let $x \in \mathcal{X}^*_\ell$, then we have for any $i \in \{\ell_1, \ldots, \ell_K\}$ and $i' \in \{\ell_{K+1}, \ldots, \ell_{k_*}\}$ that

$$\begin{aligned}
\beta_{1i}^\top x &= (\beta_{1i} - \beta^*_{1i})^\top x + (\beta^*_{1i})^\top x \\
&\geq -M_i\varepsilon B + (\beta^*_{1i'})^\top x + c^*_\ell \varepsilon \\
&= -M_i\varepsilon B + c^*_\ell\varepsilon + (\beta^*_{1i'} - \beta_{1i'})^\top x + \beta_{1i'}^\top x \\
&\geq -2M_i\varepsilon B + +c^*_\ell\varepsilon + \beta_{1i'}^\top x.
\end{aligned}$$

By setting $M_i \leq \dfrac{c^*_\ell}{2B}$, we get that $x \in \mathcal{X}_\ell$, which means that $\mathcal{X}^*_\ell \subseteq \mathcal{X}_\ell$. Similarly, assume that there exists some constant $c_\ell \geq 0$ that satisfies

$$\min_{x,i,i'} \Big[(\beta^*_{1i})^\top x - (\beta^*_{1i'})^\top x \Big] = c^*_\ell \varepsilon.$$

Here, the above minimum is subject to $x \in \mathcal{X}_\ell$, $i \in \{\ell_1, \ldots, \ell_K\}$ and $i' \in \{\ell_{K+1}, \ldots, \ell_{k_*}\}$. If $M_i \leq \dfrac{c_\ell}{2B}$, then we also receive that $\mathcal{X}_\ell \subseteq \mathcal{X}^*_\ell$.

Hence, if we set $M_i = \dfrac{1}{2B} \min\{c^*_\ell, c_\ell\}$, we reach the conclusion that $\mathcal{X}^*_\ell = \mathcal{X}_\ell$.

B Proof for Results under Over-specified Settings

In this appendix, we first provide the proofs of Theorem B.1 and Theorem 4 in Appendix B.1 and Appendix B.2, respectively. Subsequently, we present the proof for Proposition 1 in Appendix B.3, while that for Lemma 2 is put in Appendix B.4.

B.1 Proof of Theorem 3

In this appendix, we follow proof techniques presented in Appendix A.1 to demonstrate the result of Theorem 3. Recall that under the over-specified settings, the MLE $\widehat{G}_n$ belongs to the set of all mixing measures with at most $k > k_*$ components, i.e. $\mathcal{O}_k(\Omega)$. Interestingly, if we can adapt the result of part (i) of Lemma 3 to the over-specified settings, then other results presented in Appendix A.1 will also hold true. Therefore, our main goal is to derive following bound for any $0 < \eta < 1/2$ under the over-specified settings:

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

where $\mathcal{P}_k(\Omega) := \{\overline{g}_G(Y|X) : G \in \mathcal{O}_k(\Omega)\}$. For ease of presentation, we will reuse the notations defined in Appendix A.1 with $\mathcal{E}_{k_*}(\Omega)$ being replaced by $\mathcal{O}_k(\Omega)$. Now, let us recall necessary notations for this proof.

Firstly, we define $\Theta = \{(a, b, \sigma) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+ : (\beta_0, \beta_1, a, b, \sigma) \in \Omega\}$, and $\overline{\Theta}_\eta$ is an η-cover of Θ. Additionally, we also denote $\Delta := \{(\beta_0, \beta_1) \in \mathbb{R}^d \times \mathbb{R} : (\beta_0, \beta_1, a, b, \sigma) \in \Omega\}$, and $\overline{\Delta}_\eta$ be an η-cover of Δ. Next, for each mixing measure $G = \sum_{i=1}^{k} \exp(\beta_{0i})\delta_{(\beta_{1i}, a_i, b_i, \sigma_i)} \in \mathcal{O}_k(\Omega)$, we denote $G' = \sum_{i=1}^{k} \exp(\beta_{0i})\delta_{(\beta_{1i}, \overline{a}_i, \overline{b}_i, \overline{\sigma}_i)}$ in which $(\overline{a}_i, \overline{b}_i, \overline{\sigma}_i) \in \overline{\Theta}_\eta$ is the closest point to (a_i, b_i, σ_i) in this set for any $i \in [k]$. We also consider another mixing measure $\overline{G} := \sum_{i=1}^{k} \exp(\overline{\beta}_{0i})\delta_{(\overline{\beta}_{1i}, \overline{a}_i, \overline{b}_i, \overline{\sigma}_i)} \in \mathcal{O}_k(\Omega)$ where $(\overline{\beta}_{0i}, \overline{\beta}_{1i}) \in \overline{\Delta}_\eta$ is the closest point to (β_{0i}, β_{1i}) in this set for any $i \in [k]$.

Subsequently, we define

$$\mathcal{L} := \{\overline{g}_{\overline{G}} \in \mathcal{P}_k(\Omega) : (\overline{\beta}_{0i}, \overline{\beta}_{1i}) \in \overline{\Delta}_\eta,\ (\overline{a}_i, \overline{b}_i, \overline{\sigma}_i) \in \overline{\Theta}_\eta\}.$$

We demonstrate that $\mathcal{L}$ is an η-cover of the metric space $(\mathcal{P}_k(\Omega), \|\cdot\|_1)$, that is, for any $\overline{g}_G \in \mathcal{P}_k(\Omega)$, there exists a density $\overline{g}_{\overline{G}} \in \mathcal{L}$ such that $\|\overline{g}_G - \overline{g}_{\overline{G}}\|_1 \leq \eta$. By the triangle inequality, we have

$$\|\overline{g}_G - \overline{g}_{\overline{G}}\|_1 \leq \|\overline{g}_G - \overline{g}_{G'}\|_1 + \|\overline{g}_{G'} - \overline{g}_{\overline{G}}\|_1. \tag{23}$$

From the formulation of G', we get that

$$\begin{aligned}
\|\overline{g}_G - \overline{g}_{G'}\|_1 &\leq \sum_{i=1}^{k} \int_{\mathcal{X}\times\mathcal{Y}} \Big| f(Y|a_i^\top X + b_i, \sigma_i) - f(Y|\overline{a}_i^\top X + \overline{b}_i, \overline{\sigma}_i) \Big| \mathrm{d}(X, Y) \\
&\lesssim \sum_{i=1}^{k} \int_{\mathcal{X}\times\mathcal{Y}} \Big(\|a_i - \overline{a}_i\| + |b_i - \overline{b}_i| + |\sigma_i - \overline{\sigma}_i| \Big) \mathrm{d}(X, Y) \\
&\lesssim \eta
\end{aligned} \tag{24}$$

Based on inequalities in equations (23) and (24), it is sufficient to show that $\|\overline{g}_{G'} - \overline{g}_{\overline{G}}\|_1 \lesssim \eta$. For any $\overline{\ell} \in [\overline{q}]$, let us define

$$\begin{aligned}
\overline{\mathcal{X}}_{\overline{\ell}} &:= \{x \in \mathcal{X} : (\beta_{1i})^\top x \geq (\beta_{1i'})^\top x,\ \forall i \in \{\overline{\ell}_1, \ldots, \overline{\ell}_{\overline{K}}\}, i' \in \{\overline{\ell}_{\overline{K}+1}, \ldots, \overline{\ell}_k\}\}, \\
\mathcal{X}'_{\overline{\ell}} &:= \{x \in \mathcal{X} : (\overline{\beta}_{1i})^\top x \geq (\overline{\beta}_{1i'})^\top x,\ \forall i \in \{\overline{\ell}_1, \ldots, \overline{\ell}_{\overline{K}}\}, i' \in \{\overline{\ell}_{\overline{K}+1}, \ldots, \overline{\ell}_k\}\}.
\end{aligned}$$

Since the $\mathrm{Softmax}$ function is differentiable, it is a Lipschitz function with some Lipschitz constant $L \geq 0$. Assume that $\|X\| \leq B$ for any $X \in \mathcal{X}$ and denote

$$\pi_{\overline{\ell}}(X) := \Big(\beta_{1\overline{\ell}_i}^\top x + \beta_{0\overline{\ell}_i}^\top\Big)_{i=1}^{\overline{K}}; \qquad \overline{\pi}_{\overline{\ell}}(X) := \Big(\overline{\beta}_{1\overline{\ell}_i}^\top x + \overline{\beta}_{0\overline{\ell}_i}^\top\Big)_{i=1}^{\overline{K}},$$

for any $\overline{K}$-element subset $\{\overline{\ell}_1, \ldots \overline{\ell}_K\}$ of $\{1, \ldots, k\}$. Then, we have

$$\begin{aligned}
\|\mathrm{Softmax}(\pi_{\overline{\ell}}(X)) - \mathrm{Softmax}(\overline{\pi}_{\overline{\ell}}(X))\| &\leq L \cdot \|\pi_{\overline{\ell}}(X) - \overline{\pi}_{\overline{\ell}}(X)\| \\
&\leq L \cdot \sum_{i=1}^{\overline{K}} \Big(\|\beta_{1\overline{\ell}_i} - \overline{\beta}_{1\overline{\ell}_i}\| \cdot \|X\| + |\beta_{0\overline{\ell}_i} - \overline{\beta}_{0\overline{\ell}_i}| \Big) \\
&\leq L \cdot \sum_{i=1}^{\overline{K}} \Big(\eta B + \eta \Big) \\
&\lesssim \eta.
\end{aligned}$$

By arguing similarly to the proof of Lemma 2 in Appendix B.4, we receive that either $\overline{\mathcal{X}}_{\overline{\ell}} = \mathcal{X}'_{\overline{\ell}}$ or $\overline{\mathcal{X}}_{\overline{\ell}}$ has measure zero for any $\overline{\ell} \in [\overline{q}]$. As a result, we deduce that

$$\begin{aligned}
\|\overline{g}_{G'} - \overline{g}_{G_*}\|_1 &\leq \sum_{\overline{\ell}=1}^{\overline{q}} \int_{\overline{\mathcal{X}}_{\overline{\ell}} \times \mathcal{Y}} |\overline{g}_{G'}(Y|X) - \overline{g}_{\overline{G}}(Y|X)| \mathrm{d}(X,Y) \\
&\leq \sum_{\overline{\ell}=1}^{\overline{q}} \int_{\overline{\mathcal{X}}_{\overline{\ell}} \times \mathcal{Y}} \sum_{i=1}^{\overline{K}} \Big|\mathrm{Softmax}(\pi_{\overline{\ell}}(X)_i) - \mathrm{Softmax}(\overline{\pi}_{\overline{\ell}}(X)_i)\Big| \cdot \Big|f(Y|\overline{a}_{\overline{\ell}_i}^{\top} X + \overline{b}_{\overline{\ell}_i}, \overline{\sigma}_{\overline{\ell}_i})\Big| \mathrm{d}(X,Y) \\
&\lesssim \eta.
\end{aligned}$$

Thus, $\mathcal{L}$ is an η-cover of the metric space $(\mathcal{P}_k(\Omega), \|\cdot\|_1)$, which implies that

$$N(\eta, \mathcal{P}_k(\Omega), \|\cdot\|_1) \lesssim |\overline{\Delta}_\eta| \times |\overline{\Theta}_\eta| \leq \mathcal{O}(\eta^{-(d+1)k}) \times \mathcal{O}(\eta^{-(d+3)k}) = \mathcal{O}(\eta^{-(2d+4)k}). \tag{25}$$

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

B.2 Proof of Theorem 4

Similar to the proof of Theorem 2 in Appendix A, our objective here is also to derive the Total Variation lower bound adapted to the over-fitted settings:

$$\mathbb{E}_X[V(\overline{g}_G(\cdot|X), g_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_2(G, G_*).$$

Since the global part of the above inequality can be argued in the same fashion as in Appendix A, we will focus only on demonstrating the following local part via the proof by contradiction method:

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{O}_k(\Theta): \mathcal{D}_2(G, G_*) \leq \varepsilon} \frac{\mathbb{E}_X[V(\overline{g}_G(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_2(G, G_*)} > 0. \tag{26}$$

Assume that the above claim does not hold true, then we can find a sequence of mixing measures $G_n := \sum_{i=1}^{k_n} \exp(\beta_{0i}^n) \delta_{(\beta_{1i}^n, a_i^n, b_i^n, \sigma_i^n)} \in \mathcal{O}_k(\Omega)$ such that both $\mathcal{D}_2(G_n, G_*)$ and $\mathbb{E}_X[V(\overline{g}_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_2(G_n, G_*)$ vanish when n goes to infinity. Additionally, by abuse of notation, we reuse the set of Voronoi cells $\mathcal{C}_j$, for $j \in [k_*]$, defined in Appendix A. Due to the limit $\mathcal{D}_2(G_n, G_*) \to 0$ as $n \to \infty$, it follows that for any $j \in [k_*]$, we have $\sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n) \to \exp(\beta_{0j}^*)$ and $(\beta_{1i}^n, a_i^n, b_i^n, \sigma_i^n) \to (\beta_{1j}^*, a_j^*, b_j^*, \sigma_j^*)$ for all $i \in \mathcal{C}_j$. WLOG, we may assume that

$$\begin{aligned}
\mathcal{D}_2(G_n, G_*) = &\sum_{\substack{j \in [K], \\ |\mathcal{C}_j| > 1}} \sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n) \Big[\|\Delta \beta_{1ij}^n\|^{\bar{r}(|\mathcal{C}_j|)} + \|\Delta a_{ij}^n\|^{\frac{\bar{r}(|\mathcal{C}_j|)}{2}} + |\Delta b_{ij}^n|^{\bar{r}(|\mathcal{C}_j|)} + |\Delta \sigma_{ij}^n|^{\frac{\bar{r}(|\mathcal{C}_j|)}{2}}\Big] \\
&+ \sum_{\substack{j \in [K], \\ |\mathcal{C}_j| = 1}} \sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n) \Big[\|\Delta \beta_{1ij}^n\| + \|\Delta a_{ij}^n\| + |\Delta b_{ij}^n| + |\Delta \sigma_{ij}^n|\Big] + \sum_{j=1}^{K} \Big|\sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^*)\Big|.
\end{aligned}$$

Regarding the top-K selection in the conditional density g_{G_*}, we partition the covariate space $\mathcal{X}$ in a similar fashion to Appendix A. More specifically, we consider $q = \binom{k_*}{K}$ subsets $\{\ell_1, \ldots, \ell_K\}$ of $\{1, \ldots, k_*\}$ for any $\ell \in [q]$, and denote $\{\ell_{K+1}, \ldots, \ell_{k_*}\} := [k_*] \setminus \{\ell_1, \ldots, \ell_K\}$. Then, we define

$$\mathcal{X}_\ell^* := \Big\{x \in \mathcal{X} : (\beta_{1j}^*)^\top x \geq (\beta_{1j'}^*)^\top x, \forall j \in \{\ell_1, \ldots, \ell_K\}, j' \in \{\ell_{K+1}, \ldots, \ell_{k_*}\}\Big\},$$

for any $\ell \in [q]$. On the other hand, we need to introduce a new partition method of the covariate space for the weight selection in the conditional density g_{G_n}. In particular, let $\overline{K} \in \mathbb{N}$ such that $\max_{\{\ell_j\}_{j=1}^K \subset [k_*]} \sum_{j=1}^K |\mathcal{C}_{\ell_j}| \leq \overline{K} \leq k$ and $\overline{q} := \binom{k}{\overline{K}}$. Then, for any $\overline{\ell} \in [\overline{q}]$, we denote $(\overline{\ell}_1, \ldots, \overline{\ell}_k)$ as a subset of $[k]$ and $\{\overline{\ell}_{\overline{K}+1}, \ldots, \overline{\ell}_k\} := [k] \setminus \{\overline{\ell}_1, \ldots, \overline{\ell}_{\overline{K}}\}$. Additionally, we define

$$\mathcal{X}_{\overline{\ell}}^n := \Big\{x \in \mathcal{X} : (\beta_{1i}^n)^\top x \geq (\beta_{1i'}^n)^\top x, \forall i \in \{\overline{\ell}_1, \ldots, \overline{\ell}_{\overline{K}}\}, i' \in \{\overline{\ell}_{\overline{K}+1}, \ldots, \overline{\ell}_k\}\Big\}.$$

Let $X \in \mathcal{X}^*_\ell$ for some $\ell \in [q]$ such that $\{\ell_1, \ldots, \ell_K\} = \{1, \ldots, K\}$. If $\{\overline{\ell}_1, \ldots \overline{\ell}_{\overline{K}}\} \neq \mathcal{C}_1 \cup \ldots \cup \mathcal{C}_K$ for any $\overline{\ell} \in [\overline{q}]$, then $\mathbb{E}_X[V(\overline{g}_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_2(G_n, G_*) \not\to 0$ as n tends to infinity. This contradicts the fact that this term must approach zero. Therefore, we only need to consider the scenario when there exists $\overline{\ell} \in [\overline{q}]$ such that $\{\overline{\ell}_1, \ldots \overline{\ell}_{\overline{K}}\} = \mathcal{C}_1 \cup \ldots \cup \mathcal{C}_K$. Recall that we have $(\beta^n_{0i}, \beta^n_{1i}) \to (\beta^*_{0j}, \beta^*_{1j})$ as $n \to \infty$ for any $j \in [k_*]$ and $i \in \mathcal{C}_j$. Thus, for any arbitrarily small $\eta_j > 0$, we have that $\|\beta^n_{1i} - \beta^*_{1j}\| \leq \eta_j$ and $|\beta^n_{0i} - \beta^*_{0j}| \leq \eta_j$ for sufficiently large n. Then, it follows from Lemma 2 that $\mathcal{X}^*_\ell = \mathcal{X}^n_{\overline{\ell}}$ for sufficiently large n. This result indicates that $X \in \mathcal{X}^n_{\overline{\ell}}$.

Then, we can represent the conditional densities $g_{G_*}(Y|X)$ and $g_{G_n}(Y|X)$ for any sufficiently large n as follows:

$$g_{G_*}(Y|X) = \sum_{j=1}^{K} \frac{\exp((\beta^*_{1j})^\top X + \beta^*_{0j})}{\sum_{j'=1}^{K} \exp((\beta^*_{1j'})^\top X + \beta^*_{0j'})} \cdot f(Y|(a^*_j)^\top X + b^*_j, \sigma^*_j),$$

$$\overline{g}_{G_n}(Y|X) = \sum_{j=1}^{K} \sum_{i \in \mathcal{C}_j} \frac{\exp((\beta^n_{1i})^\top X + \beta^n_{0i})}{\sum_{j'=1}^{K} \sum_{i' \in \mathcal{C}_{j'}} \exp((\beta^n_{1i'})^\top X + \beta^n_{0i'})} \cdot f(Y|(a^n_i)^\top X + b^n_i, \sigma^n_i).$$

Now, we reuse the three-step framework in Appendix A.

Step 1 - Taylor expansion:

Firstly, by abuse of notations, let us consider the quantity

$$H_n := \Big[\sum_{j=1}^{K} \exp((\beta^*_{1j})^\top X + \beta^*_{0j})\Big] \cdot [\overline{g}_{G_n}(Y|X) - g_{G_*}(Y|X)].$$

Similar to Step 1 in Appendix A, we can express this term as

$$\begin{aligned}
H_n &= \sum_{j=1}^{K} \sum_{i \in \mathcal{C}_j} \exp(\beta^n_{0i}) \Big[\exp((\beta^n_{1i})^\top X) f(Y|(a^n_i)^\top X + b^n_i, \sigma^n_i) - \exp((\beta^*_{1j})^\top X) f(Y|(a^*_j)^\top X + b^*_j, \sigma^*_j)\Big] \\
&- \sum_{j=1}^{K} \sum_{i \in \mathcal{C}_j} \exp(\beta^n_{0i}) \Big[\exp((\beta^n_{1i})^\top X) g_{G_n}(Y|X) - \exp((\beta^*_{1j})^\top X) g_{G_n}(Y|X)\Big] \\
&+ \sum_{j=1}^{K} \Big[\sum_{i \in \mathcal{C}_j} \exp(\beta^n_{0i}) - \exp(\beta^*_{0j})\Big] \Big[\exp((\beta^*_{1j})^\top X) f(Y|(a^*_i)^\top X + b^*_i, \sigma^*_i) - \exp((\beta^*_{1j})^\top X) g_{G_n}(Y|X)\Big] \\
&:= A_n + B_n + E_n.
\end{aligned}$$

Next, we proceed to decompose A_n based on the cardinality of the Voronoi cells as follows:

$$\begin{aligned}
A_n &= \sum_{j:|\mathcal{C}_j|=1} \sum_{i \in \mathcal{C}_j} \exp(\beta^n_{0i}) \Big[\exp((\beta^n_{1i})^\top X) f(Y|(a^n_i)^\top X + b^n_i, \sigma^n_i) - \exp((\beta^*_{1j})^\top X) f(Y|(a^*_j)^\top X + b^*_j, \sigma^*_j)\Big] \\
&+ \sum_{j:|\mathcal{C}_j|>1} \sum_{i \in \mathcal{C}_j} \exp(\beta^n_{0i}) \Big[\exp((\beta^n_{1i})^\top X) f(Y|(a^n_i)^\top X + b^n_i, \sigma^n_i) - \exp((\beta^*_{1j})^\top X) f(Y|(a^*_j)^\top X + b^*_j, \sigma^*_j)\Big].
\end{aligned}$$

By applying the Taylor expansions of order 1 and $\bar{r}(|\mathcal{C}_j|)$ to the first and second terms of A_n, respectively, and following the derivation in equation (16), we arrive at

$$\begin{aligned}
A_n = &\sum_{j:|\mathcal{C}_j|=1} \sum_{i \in \mathcal{C}_j} \sum_{1 \leq |\eta_1| + \eta_2 \leq 2} \sum_{\alpha \in \mathcal{J}_{\eta_1,\eta_2}} \frac{\exp(\beta^n_{0i})}{2^{\alpha_4} \alpha!} \cdot (\Delta\beta^n_{1i})^{\alpha_1} (\Delta a^n_i)^{\alpha_2} (\Delta b^n_i)^{\alpha_3} (\Delta\sigma^n_i)^{\alpha_4} \\
&\qquad \times X^{\eta_1} \exp((\beta^*_{1i})^\top X) \cdot \frac{\partial^{\eta_2} f}{\partial h_1^{\eta_2}}(Y|(a^*_i)^\top X + b^*_i, \sigma^*_i) + R_3(X,Y) \\
&+ \sum_{j:|\mathcal{C}_j|>1} \sum_{i \in \mathcal{C}_j} \sum_{1 \leq |\eta_1| + \eta_2 \leq 2\bar{r}(|\mathcal{C}_j|)} \sum_{\alpha \in \mathcal{J}_{\eta_1,\eta_2}} \frac{\exp(\beta^n_{0i})}{2^{\alpha_4} \alpha!} \cdot (\Delta\beta^n_{1i})^{\alpha_1} (\Delta a^n_i)^{\alpha_2} (\Delta b^n_i)^{\alpha_3} (\Delta\sigma^n_i)^{\alpha_4} \\
&\qquad \times X^{\eta_1} \exp((\beta^*_{1i})^\top X) \cdot \frac{\partial^{\eta_2} f}{\partial h_1^{\eta_2}}(Y|(a^*_i)^\top X + b^*_i, \sigma^*_i) + R_4(X,Y),
\end{aligned}$$

where the set $\mathcal{J}_{\eta_1,\eta_2}$ is defined in equation (17) while $R_i(X,Y)$ is a Taylor remainder such that $R_i(X,Y)/\mathcal{D}_2(G_n,G_*)\to 0$ as $n\to\infty$ for $i\in\{3,4\}$. Similarly, we also decompose B_n according to the Voronoi cells as A_n but then invoke the Taylor expansions of order 1 and 2 to the first term and the second term, respectively. In particular, we get

$$\begin{aligned}B_n=&-\sum_{j:|\mathcal{C}_j|=1}\sum_{i\in\mathcal{C}_j}\sum_{|\gamma|=1}\frac{\exp(\beta^n_{0i})}{\gamma!}(\Delta\beta^n_{1i})\cdot X^{\gamma}\exp((\beta^n_{1i})^{\top}X)g_{G_n}(Y|X)+R_5(X,Y)\\&-\sum_{j:|\mathcal{C}_j|>1}\sum_{i\in\mathcal{C}_j}\sum_{1\leq|\gamma|\leq 2}\frac{\exp(\beta^n_{0i})}{\gamma!}(\Delta\beta^n_{1i})\cdot X^{\gamma}\exp((\beta^n_{1i})^{\top}X)g_{G_n}(Y|X)+R_6(X,Y),\end{aligned}$$

where $R_5(X,Y)$ and $R_6(X,Y)$ are Taylor remainders such that their ratios over $\mathcal{D}_2(G_n,G_*)$ approach zero as $n\to\infty$. Subsequently, let us define

$$\begin{aligned}S^n_{j,\eta_1,\eta_2}&:=\sum_{i\in\mathcal{C}_j}\sum_{\alpha\in\mathcal{J}_{\eta_1,\eta_2}}\frac{\exp(\beta^n_{0i})}{2^{\alpha_4}\alpha!}\cdot(\Delta\beta^n_{1i})^{\alpha_1}(\Delta a^n_i)^{\alpha_2}(\Delta b^n_i)^{\alpha_3}(\Delta\sigma^n_i)^{\alpha_4},\\T^n_{j,\gamma}&:=\sum_{i\in\mathcal{C}_j}\frac{\exp(\beta^n_{0i})}{\gamma!}(\Delta\beta^n_{1i})^{\gamma},\end{aligned}$$

for any $(\eta_1,\eta_2)\neq(\mathbf{0}_d,0)$ and $|\gamma|\neq\mathbf{0}_d$, while for $(\eta_1,\eta_2)=(\mathbf{0}_d,0)$ we set

$$S^n_{i,\mathbf{0}_d,0}=-T^n_{i,\mathbf{0}_d}:=\sum_{i\in\mathcal{C}_j}\exp(\beta^n_{0i})-\exp(\beta^*_{0i}).$$

As a consequence, it follows that

$$\begin{aligned}H_n=&\sum_{j=1}^{K}\sum_{|\eta_1|+\eta_2=0}^{2\bar{r}(|\mathcal{C}_j|)}S^n_{j,\eta_1,\eta_2}\cdot X^{\eta_1}\exp((\beta^*_{1i})^{\top}X)\cdot\frac{\partial^{\eta_2}f}{\partial h_1^{\eta_2}}(Y|(a^*_i)^{\top}X+b^*_i,\sigma^*_i)\\&+\sum_{j=1}^{K}\sum_{|\gamma|=0}^{1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}}}T^n_{j,\gamma}\cdot X^{\gamma}\exp((\beta^n_{1i})^{\top}X)g_{G_n}(Y|X)+R_5(X,Y)+R_6(X,Y).\end{aligned}\tag{27}$$

Step 2 - Non-vanishing coefficients:

In this step, we will prove by contradiction that at least one among the ratios $S^n_{j,\eta_1,\eta_2}/\mathcal{D}_2(G_n,G_*)$ does not converge to zero as $n\to\infty$. Assume that all these terms go to zero, then by employing arguments for deriving equations (19) and (20), we get that

$$\begin{aligned}\frac{1}{\mathcal{D}_2(G_n,G_*)}\cdot\Big[\sum_{j=1}^{K}\Big|\sum_{i\in\mathcal{C}_j}\exp(\beta^n_{0i})-\exp(\beta^*_{0j})\Big|\\+\sum_{j:|\mathcal{C}_j|=1}\sum_{i\in\mathcal{C}_j}\exp(\beta^n_{0i})\Big(\|\Delta\beta^n_{1ij}\|+\|\Delta a^n_{ij}\|+|\Delta b^n_{ij}|+|\Delta\sigma^n_{ij}|\Big)\Big]\to 0.\end{aligned}$$

Combine this limit with the representation of $\mathcal{D}_2(G_n,G_*)$, we have that

$$\frac{1}{\mathcal{D}_2(G_n,G_*)}\cdot\sum_{j:|\mathcal{C}_j|>1}\sum_{i\in\mathcal{C}_j}\exp(\beta^n_{0i})\Big(\|\Delta\beta^n_{1ij}\|^{\bar{r}(|\mathcal{C}_j|)}+\|\Delta a^n_{ij}\|^{\frac{\bar{r}(|\mathcal{C}_j|)}{2}}+|\Delta b^n_{ij}|^{\bar{r}(|\mathcal{C}_j|)}+|\Delta\sigma^n_{ij}|^{\frac{\bar{r}(|\mathcal{C}_j|)}{2}}\Big)\not\to 0.$$

This result implies that we can find some index $j'\in[K]:|\mathcal{C}_{j'}|>1$ that satisfies

$$\frac{1}{\mathcal{D}_2(G_n,G_*)}\cdot\sum_{i\in\mathcal{C}_{j'}}\exp(\beta^n_{0i})\Big(\|\Delta\beta^n_{1ij'}\|^{\bar{r}(|\mathcal{C}_{j'}|)}+\|\Delta a^n_{ij'}\|^{\frac{\bar{r}(|\mathcal{C}_{j'}|)}{2}}+|\Delta b^n_{ij'}|^{\bar{r}(|\mathcal{C}_{j'}|)}+|\Delta\sigma^n_{ij'}|^{\frac{\bar{r}(|\mathcal{C}_{j'}|)}{2}}\Big)\not\to 0.$$

For simplicity, we may assume that $j'=1$. Since $S^n_{1,\eta_1,\eta_2}/\mathcal{D}_2(G_n,G_*)$ vanishes as $n\to\infty$ for any $(\eta_1,\eta_2)\in\mathbb{N}^d\times\mathbb{N}$ such that $1\leq|\eta_1|+\eta_2\leq\bar{r}(|\mathcal{C}_j|)$, we divide this term by the left hand side of the above equation and achieve that

$$\frac{\sum_{i\in\mathcal{C}_1}\sum_{\alpha\in\mathcal{J}_{\eta_1,\eta_2}}\frac{\exp(\beta^n_{0i})}{2^{\alpha_4}\alpha!}(\Delta\beta^n_{1i1})^{\alpha_1}(\Delta a^n_{i1})^{\alpha_2}(\Delta b^n_{i1})^{\alpha_3}(\Delta\sigma^n_{i1})^{\alpha_4}}{\sum_{i\in\mathcal{C}_1}\exp(\beta^n_{0i})\Big(\|\Delta b^n_{i1}\|^{\bar{r}(|\mathcal{C}_1|)}+\|\Delta a^n_{i1}\|^{\frac{\bar{r}(|\mathcal{C}_1|)}{2}}+|\Delta b^n_{i1}|^{\bar{r}(|\mathcal{C}_1|)}+|\Delta\sigma^n_{i1}|^{\frac{\bar{r}(|\mathcal{C}_1|)}{2}}\Big)}\to 0,\tag{28}$$

for any $(\eta_1, \eta_2) \in \mathbb{N}^d \times \mathbb{N}$ such that $1 \leq |\eta_1| + \eta_2 \leq \bar{r}(|\mathcal{C}_1|)$.

Subsequently, we define $M_n := \max\{\|\Delta\beta^n_{1i1}\|, \|\Delta a^n_{i1}\|^{1/2}, |\Delta b^n_{i1}|, |\Delta\sigma^n_{i1}|^{1/2} : i \in \mathcal{C}_1\}$ and $p_n := \max\{\exp(\beta^n_{0i}) : i \in \mathcal{C}_1\}$. As a result, the sequence $\exp(\beta^n_{0i})/p_n$ is bounded, which indicates that we can substitute it with its subsequence that admits a positive limit $z^2_{5i} := \lim_{n\to\infty} \exp(\beta^n_{0i})/p_n$. Therefore, at least one among the limits z^2_{5i} equals to one. Furthermore, we also denote

$$(\Delta\beta^n_{1i1})/M_n \to z_{1i},\ (\Delta a^n_{i1})/M_n \to z_{2i},\ (\Delta b^n_{i1})/M_n \to z_{3i},\ (\Delta\sigma^n_{i1})/(2M_n) \to z_{4i}.$$

From the above definition, it follows that at least one among the limits z_{1i}, z_{2i}, z_{3i} and z_{4i} equals to either 1 or -1. By dividing both the numerator and the denominator of the term in equation (28) by $p_n M_n^{|\eta_1|+\eta_2}$, we arrive at the following system of polynomial equations:

$$\sum_{i\in\mathcal{C}_1} \sum_{\alpha\in\mathcal{J}_{\eta_1,\eta_2}} \frac{z^2_{5i}\ z^{\alpha_1}_{1i}\ z^{\alpha_2}_{2i}\ z^{\alpha_3}_{3i}\ z^{\alpha_4}_{4i}}{\alpha_1!\ \alpha_2!\ \alpha_3!\ \alpha_4!} = 0,$$

for all $(\eta_1, \eta_2) \in \mathbb{N}^d \times \mathbb{N} : 1 \leq |\eta_1| + \eta_2 \leq \bar{r}(|\mathcal{C}_1|)$. Nevertheless, from the definition of $\bar{r}(|\mathcal{C}_1|)$, we know that the above system does not admit any non-trivial solutions, which is a contradiction. Consequently, not all the ratios $S^n_{j,\eta_1,\eta_2}/\mathcal{D}_2(G_n, G_*)$ tend to zero as n goes to infinity.

Step 3 - Fatou's contradiction:

It follows from the hypothesis that $\mathbb{E}_X[V(\overline{g}_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_2(G_n, G_*) \to 0$ as $n \to \infty$. Then, by applying the Fatou's lemma, we get

$$0 = \lim_{n\to\infty} \frac{\mathbb{E}_X[V(\overline{g}_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_2(G_n, G_*)} = \frac{1}{2} \cdot \int \liminf_{n\to\infty} \frac{|\overline{g}_{G_n}(Y|X) - g_{G_*}(Y|X)|}{\mathcal{D}_2(G_n, G_*)} \mathrm{d}X\mathrm{d}Y,$$

which implies that $|\overline{g}_{G_n}(Y|X) - g_{G_*}(Y|X)|/\mathcal{D}_2(G_n, G_*) \to 0$ as $n \to \infty$ for almost surely (X, Y).

Next, we define m_n as the maximum of the absolute values of $S^n_{j,\eta_1,\eta_2/\mathcal{D}_2(G_n,G_*)}$. It follows from Step 2 that $1/m_n \not\to \infty$. Moreover, by arguing in the same way as in Step 3 in Appendix A, we receive that

$$H_n/[m_n\mathcal{D}_2(G_n, G_*)] \to 0 \tag{29}$$

as $n \to \infty$. By abuse of notations, let us denote

$$S^n_{j,\eta_1,\eta_2}/[m_n\mathcal{D}_2(G_n, G_*)] \to \tau_{j,\eta_1,\eta_2},$$
$$T^n_{j,\gamma}/[m_n\mathcal{D}_2(G_n, G_*)] \to \kappa_{j,\gamma}.$$

Here, at least one among $\tau_{j,\eta_1,\eta_2}, \kappa_{j,\gamma}$ is non-zero. Then, by putting the results in equations (27) and (29) together, we get

$$\sum_{i=1}^{K} \sum_{|\eta_1|+\eta_2=0}^{2\bar{r}(|\mathcal{C}_j|)} \tau_{i,\eta_1,\eta_2} \cdot X^{\eta_1} \exp((\beta^*_{1i})^\top X) \frac{\partial^{\eta_2} f}{\partial h_1^{\eta_2}}(Y|(a^*_i)^\top X + b^*_i, \sigma^*_i)$$
$$+ \sum_{i=1}^{K} \sum_{|\gamma|=0}^{1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}}} \kappa_{i,\gamma} \cdot X^{\gamma} \exp((\beta^*_{1i})^\top X) g_{G_*}(Y|X) = 0.$$

Arguing in a similar fashion as in Step 3 of Appendix A, we obtain that $\tau_{j,\eta_1,\eta_2} = \kappa_{j,\gamma} = 0$ for any $j \in [K]$, $0 \leq |\eta_1| + \eta_2 \leq 2\bar{r}(|\mathcal{C}_j|)$ and $0 \leq |\gamma| \leq 1 + \mathbf{1}_{\{|\mathcal{C}_j|>1\}}$. This contradicts the fact that at least one among them is non-zero. Hence, the proof is completed.

B.3 Proof of Proposition 1

Since the Hellinger distance is lower bounded by the Total Variation distance, i.e. $h \geq V$, it is sufficient to show that

$$\inf_{G\in\mathcal{O}_k(\Omega)} \mathbb{E}_X[V(\overline{g}_G(\cdot|X), g_{G_*}(\cdot|X))] > 0.$$

For that purpose, we first demonstrate that

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{O}_k(\Omega): \mathcal{D}_2(G, G_*) \leq \varepsilon} \mathbb{E}_X[V(\overline{g}(\cdot|X), g_{G_*}(\cdot|X))] > 0. \tag{30}$$

Assume by contrary that the above claim is not true, then we can find a sequence $G_n = \sum_{i=1}^{k_n} \exp(\beta_{0i}^n) \delta_{(\beta_{1i}^n, a_i^n, b_i^n, \sigma_i^n)} \in \mathcal{O}_k(\Omega)$ that satisfies $\mathcal{D}_2(G_n, G_*) \to 0$ and

$$\mathbb{E}_X[V(\overline{g}_{G_n}(\cdot|X), g_{G_*}(\cdot|X))] \to 0$$

when n tends to infinity. By applying the Fatou's lemma, we have

$$\begin{aligned} 0 &= \lim_{n \to \infty} \mathbb{E}_X[V(\overline{g}_{G_n}(\cdot|X), g_{G_*}(\cdot|X))] \\ &\geq \frac{1}{2} \int_{\mathcal{X} \times \mathcal{Y}} \liminf_{n \to \infty} |\overline{g}_{G_n}(Y|X) - g_{G_*}(Y|X)| \mathrm{d}(X, Y). \end{aligned} \tag{31}$$

The above results indicates that $\overline{g}_{G_n}(Y|X) - g_{G_*}(Y|X) \to 0$ as $n \to \infty$ for almost surely (X, Y). WLOG, we may assume that

$$\max_{\{\ell_1, \ldots, \ell_K\}} \sum_{j=1}^{K} |\mathcal{C}_{\ell_j}| = |\mathcal{C}_1| + |\mathcal{C}_2| + \ldots + |\mathcal{C}_K|.$$

Let us consider $X \in \mathcal{X}_\ell^*$, where $\ell \in [q]$ such that $\{\ell_1, \ldots, \ell_K\} = \{1, \ldots, K\}$. Since $\mathcal{D}_2(G_n, G_*)$ converges to zero, it follows that $(\beta_{1i}^n, a_i^n, b_i^n, \sigma_i^n) \to (\beta_{1j}^*, a_j^*, b_j^*, \sigma_j^*)$ and $\sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n) \to \exp(\beta_{0j}^*)$ for any $i \in \mathcal{C}_j$ and $j \in [k_*]$. Thus, we must have that $X \in \mathcal{X}_{\overline{\ell}}$ for some $\overline{\ell} \in [\overline{q}]$ such that $\{\overline{\ell}_1, \ldots, \overline{\ell}_{\overline{K}}\} = \mathcal{C}_1 \cup \ldots \cup \mathcal{C}_K$. Otherwise, $\overline{g}_{G_n}(Y|X) - g_{G_*}(Y|X) \not\to 0$, which is a contradiction. However, as $\overline{K} < \sum_{j=1}^{K} |\mathcal{C}_j|$, the fact that $\{\overline{\ell}_1, \ldots, \overline{\ell}_{\overline{K}}\} = \mathcal{C}_1 \cup \ldots \cup \mathcal{C}_K$ cannot occur. Therefore, we reach the claim in equation (30). Consequently, there exists some positive constant ε' such that

$$\inf_{G \in \mathcal{O}_k(\Omega): \mathcal{D}_2(G, G_*) \leq \varepsilon'} \mathbb{E}_X[V(\overline{g}_G(\cdot|X), g_{G_*}(\cdot|X))] > 0.$$

Given the above result, it suffices to point out that

$$\inf_{G \in \mathcal{O}_k(\Omega): \mathcal{D}_2(G, G_*) > \varepsilon'} \mathbb{E}_X[V(\overline{g}_G(\cdot|X), g_{G_*}(\cdot|X))] > 0. \tag{32}$$

We continue to use the proof by contradiction method here. In particular, assume that the inequality (32) does not hold, then there exists a sequence of mixing measures $G'_n \in \mathcal{O}_k(\Omega)$ such that $\mathcal{D}_2(G'_n, G_*) > \varepsilon'$ and

$$\mathbb{E}_X[V(\overline{g}_{G'_n}(\cdot|X), g_{G_*}(\cdot|X))] \to 0.$$

By invoking the Fatou's lemma as in equation (31), we get that $\overline{g}_{G'_n}(Y|X) - g_{G_*}(Y|X) \to 0$ as $n \to \infty$ for almost surely (X, Y). Since Ω is a compact set, we can substitute (G_n) with its subsequence which converges to some mixing measure $G' \in \mathcal{O}_k(\Omega)$. Then, the previous limit implies that $\overline{g}_{G'}(Y|X) = g_{G_*}(Y|X)$ for almost surely (X, Y). From the result of Proposition 2 in Appendix C, we know that the top-K sparse softmax gating Gaussian MoE is identifiable. Therefore, we obtain that $G' \equiv G_*$, or equivalently, $\mathcal{D}_2(G', G_*) = 0$

On the other hand, due to the hypothesis $\mathcal{D}_2(G'_n, G_*) > \varepsilon'$ for any $n \in \mathbb{N}$, we also get that $\mathcal{D}_2(G', G_*) > \varepsilon' > 0$, which contradicts the previous result. Hence we reach the claim in equation (32) and totally completes the proof.

B.4 Proof of Lemma 2

Let $\eta_j = M_j \varepsilon$, where ε is some fixed positive constant and M_i will be chosen later. As $\mathcal{X}$ and Ω are bounded sets, we can find some constant $c_\ell^* \geq 0$ such that

$$\min_{x, j, j'} \left[(\beta_{1j}^*)^\top x - (\beta_{1j'}^*)^\top x \right] = c_\ell^* \varepsilon,$$

where the above minimum is subject to $x \in \mathcal{X}_\ell^*, j \in \{\ell_1, \ldots, \ell_K\}$ and $j' \in \{\ell_{K+1}, \ldots, \ell_{k_*}\}$. By arguing similarly to the proof of Lemma 1 in Appendix A.3, we deduce that $c_\ell^* > 0$.

Since $\mathcal{X}$ is a bounded set, we may assume that $\|x\| \leq B$ for any $x \in \mathcal{X}$. Let $x \in \mathcal{X}^*_{\ell}$ and $\overline{\ell} \in [\overline{q}]$ such that $\{\overline{\ell}_1, \ldots, \overline{\ell}_{\overline{K}}\} = \mathcal{C}_{\ell_1} \cup \ldots \cup \mathcal{C}_{\ell_K}$. Then, for any $i \in \{\overline{\ell}_1, \ldots, \overline{\ell}_{\overline{K}}\}$ and $i' \in \{\overline{\ell}_{\overline{K}+1}, \ldots, \overline{\ell}_k\}$, we have that

$$\begin{aligned}
\beta_{1i}^{\top} x &= (\beta_{1i} - \beta^*_{1j})^{\top} x + (\beta^*_{1j})^{\top} x \\
&\geq -M_i \varepsilon B + (\beta^*_{1j'})^{\top} x + c^*_{\ell} \varepsilon \\
&= -M_i \varepsilon B + c^*_{\ell} \varepsilon + (\beta^*_{1j'} - \beta_{1i'})^{\top} x + \beta_{1i'}^{\top} x \\
&\geq -2M_i \varepsilon B + c^*_{\ell} \varepsilon + \beta_{1i'}^{\top} x,
\end{aligned}$$

where $j \in \{\ell_1, \ldots, \ell_K\}$ and $j' \in \{\ell_{K+1}, \ldots, \ell_{k_*}\}$ such that $i \in \mathcal{C}_j$ and $i' \in \mathcal{C}_{j'}$. If $M_j \leq \dfrac{c^*_{\ell}}{2B}$, then we get that $x \in \mathcal{X}_{\overline{\ell}}$, which leads to $\mathcal{X}^*_{\ell} \subseteq \overline{\mathcal{X}}_{\overline{\ell}}$.

Analogously, assume that there exists some constant $c_{\ell} \geq 0$ such that

$$\min_{x,j,j'} \Big[(\beta^*_{1j})^{\top} x - (\beta^*_{1j'})^{\top} x \Big] = c^*_{\ell} \varepsilon,$$

where the minimum is subject to $x \in \overline{\mathcal{X}}_{\overline{\ell}}$, $i \in \{\overline{\ell}_1, \ldots, \overline{\ell}_{\overline{K}}\}$ and $i' \in \{\overline{\ell}_{\overline{K}+1}, \ldots, \overline{\ell}_k\}$. Then, if $M_j \leq \dfrac{c_{\ell}}{2B}$, then we receive that $\overline{\mathcal{X}}_{\overline{\ell}} \subseteq \mathcal{X}^*_{\ell}$.

As a consequence, by setting $M_j = \dfrac{1}{2B} \min\{c^*_{\ell}, c_{\ell}\}$, we achieve the conclusion that $\overline{\mathcal{X}}_{\overline{\ell}} = \mathcal{X}^*_{\ell}$.

C Identifiability of the Top-K Sparse Softmax Gating Gaussian Mixture of Experts

In this appendix, we study the identifiability of the top-K sparse softmax gating Gaussian MoE, which plays an essential role in ensuring the convergence of the MLE $\widehat{G}_n$ to the true mixing measure G_* under Voronoi loss functions.

Proposition 2 (Identifiability). *Let G and G' be two arbitrary mixing measures in $\mathcal{O}_k(\Theta)$. Suppose that the equation $g_G(Y|X) = g_{G'}(Y|X)$ holds for almost surely $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, then it follows that $G \equiv G'$.*

Proof of Proposition 2. First, we assume that two mixing measures G and G' take the following forms: $G = \sum_{i=1}^{k} \exp(\beta_{0i}) \delta_{(\beta_{1i}, a_i, b_i, \sigma_i)}$ and $G' = \sum_{i=1}^{k'} \exp(\beta'_{0i}) \delta_{(\beta'_{1i}, a'_i, b'_i, \sigma'_i)}$. Recall that $g_G(Y|X) = g_{G'}(Y|X)$ for almost surely (X, Y), then we have

$$\begin{aligned}
&\sum_{i=1}^{k} \mathrm{Softmax}(\mathrm{TopK}((\beta_{1i})^{\top} X, K; \beta_{0i})) \cdot f(Y | a_i^{\top} X + b_i, \sigma_i) \\
&\qquad = \sum_{i=1}^{k'} \mathrm{Softmax}(\mathrm{TopK}((\beta'_{1i})^{\top} X, K; \beta'_{0i})) \cdot f(Y | (a'_i)^{\top} + b'_i, \sigma'_i). \qquad (33)
\end{aligned}$$

Due to the identifiability of the location-scale Gaussian mixtures Teicher (1960; 1961; 1963), we get that $k = k'$ and

$$\Big\{ \mathrm{Softmax}(\mathrm{TopK}((\beta_{1i})^{\top} X, K; \beta_{0i})) : i \in [k] \Big\} \equiv \Big\{ \mathrm{Softmax}(\mathrm{TopK}((\beta'_{1i})^{\top} X, K; \beta'_{0i})) : i \in [k] \Big\},$$

for almost surely X. WLOG, we may assume that

$$\mathrm{Softmax}(\mathrm{TopK}((\beta_{1i})^{\top} X, K; \beta_{0i})) = \mathrm{Softmax}(\mathrm{TopK}((\beta'_{1i})^{\top} X, K; \beta'_{0i})), \qquad (34)$$

for almost surely X for any $i \in [k]$. Since the Softmax function is invariant to translations, it follows from equation (34) that $\beta_{1i} = \beta'_{1i} + v_1$ and $\beta_{0i} = \beta'_{0i} + v_0$ for some $v_1 \in \mathbb{R}^d$ and $v_0 \in \mathbb{R}$. Notably, from the assumption of the model, we have $\beta_{1k} = \beta'_{1k} = \mathbf{0}_d$ and $\beta_{0k} = \beta'_{0k} = 0$, which implies that $v_1 = \mathbf{0}_d$ and $v_0 = 0$. As a result, we obtain that $\beta_{1i} = \beta'_{1i}$ and $\beta_{0i} = \beta'_{0i}$ for any $i \in [k]$.

Let us consider $X \in \mathcal{X}_\ell$ where $\ell \in [q]$ such that $\{\ell_1, \ldots, \ell_K\} = \{1, \ldots, K\}$. Then, equation (33) can be rewritten as

$$\sum_{i=1}^{K} \exp(\beta_{0i}) \exp(\beta_{1i}^\top X) f(Y|a_i^\top X + b_i, \sigma_i) = \sum_{i=1}^{K} \exp(\beta_{0i}) \exp(\beta_{1i}^\top X) f(Y|(a_i')^\top X + b_i', \sigma_i'), \tag{35}$$

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 $\exp(\beta_{0i}) = \exp(\beta_{0i'})$ for any $i, i' \in J_j$ and $j \in [m]$. On the other hand, when i and i' do not belong to the same set J_j, we let $\exp(\beta_{0i}) \neq \exp(\beta_{0i'})$. Thus, we can reformulate equation (35) as

$$\sum_{j=1}^{m} \sum_{i \in J_j} \exp(\beta_{0i}) \exp(\beta_{1i}^\top X) f(Y|a_i^\top X + b_i, \sigma_i)$$
$$= \sum_{j=1}^{m} \sum_{i \in J_j} \exp(\beta_{0i}) \exp(\beta_{1i}^\top X) f(Y|(a_i')^\top X + b_i', \sigma_i'),$$

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

$$\{(a_i, b_i, \sigma_i) : i \in J_j\} \equiv \{(a_i', b_i', \sigma_i') : i \in J_j\},$$

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

$$G = \sum_{j=1}^{m} \sum_{i \in J_j} \exp(\beta_{0i}) \delta_{(\beta_{1i}, a_i, b_i, \sigma_i)} = \sum_{j=1}^{m} \sum_{i \in J_j} \exp(\beta_{0i}') \delta_{(\beta_{1i}', a_i', b_i', \sigma_i')} = G'.$$

Hence, we reach the conclusion of this proposition. □

D NUMERICAL EXPERIMENTS

In this appendix, we conduct a few numerical experiments to illustrate the theoretical convergence rates of the MLE $\widehat{G}_n$ to the true mixing measure G_* under both the exact-specified and the over-specified settings.

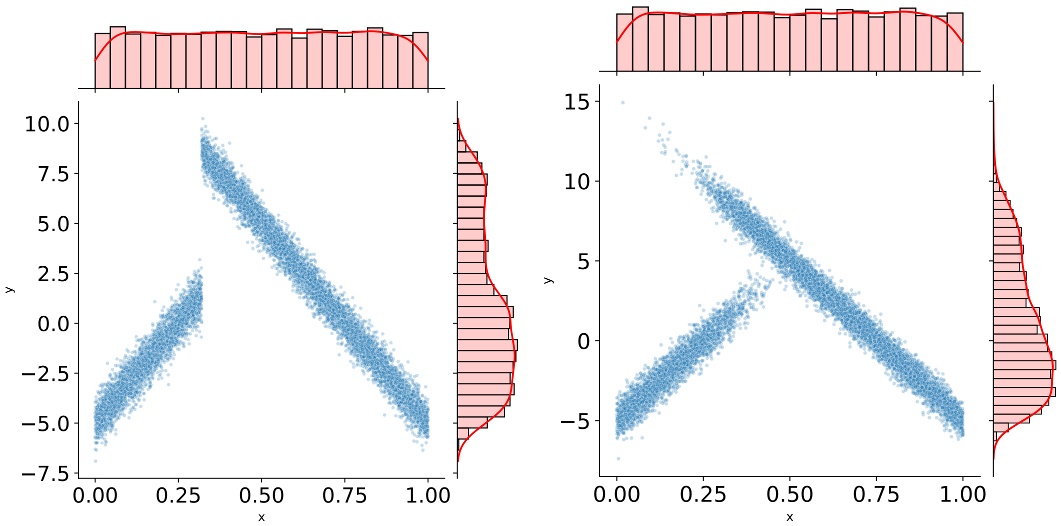
D.1 EXPERIMENTAL SETUP

Synthetic Data. First, we assume that the true mixing measure $G_* = \sum_{i=1}^{k_*} \exp(\beta_{0i}^*) \delta_{(\beta_{1i}^*, a_i^*, b_i^*, \sigma_i^*)}$ is of order $k_* = 2$ and associated with the following ground-truth parameters:

$$\begin{array}{lllll} \beta_{01}^* = -8, & \beta_{11}^* = 25, & a_1^* = -20, & b_1^* = 15, & \sigma_1^* = 0.3, \\ \beta_{02}^* = 0, & \beta_{12}^* = 0, & a_2^* = 20, & b_2^* = -5, & \sigma_2^* = 0.4. \end{array}$$

Then, we generate i.i.d samples $\{(X_i, Y_i)\}_{i=1}^n$ by first sampling X_i's from the uniform distribution $\mathrm{Uniform}[0, 1]$ and then sampling Y_i's from the true conditional density $g_{G_*}(Y|X)$ of top-K sparse softmax gating Gaussian mixture of experts (MoE) given in equation (1). In Figure 2, we visualize the relationship between X and Y when the numbers of experts chosen from $g_{G_*}(Y|X)$ are $K = 1$ (Figure 2a) and $K = 2$ (Figure 2b), respectively. However, throughout the following experiments, we will consider only the scenario when $K = 1$, that is, we choose the best expert from the true conditional density $g_{G_*}(Y|X)$.

Maximum Likelihood Estimation (MLE). A popular approach to determining the MLE $\widehat{G}_n$ for each set of samples is to use the EM algorithm Dempster et al. (1977). However, since there are not any closed-form expressions for updating the gating parameters β_{0i}, β_{1i} in the maximization steps, we have to leverage an EM-based numerical scheme, which was previously used in Chamroukhi et al. (2009). In particular, we utilize a simple coordinate gradient descent algorithm in the maximization steps. Additionally, we select the convergence criterion of $\epsilon = 10^{-6}$ and run a maximum of 2000 EM iterations.



(a) $K = 1$

(b) $K = 2$

Figure 2: A visual representation showcasing the relationship between X and Y, along with their respective marginal distributions when $K = 1$ and $K = 2$.

Initialization. For each $k \in \{k_*, k_* + 1\}$, we randomly distribute elements of the set $\{1, 2, ..., k\}$ into k_* different Voronoi cells $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_{k_*}$, each contains at least one element. Moreover, we repeat this process for each replication. Subsequently, for each $j \in [k_*]$, we initialize parameters β_{1i} by sampling from a Gaussian distribution centered around its true counterpart β^*_{1j} with a small variance, where $i \in \mathcal{C}_j$. Other parameters $\beta_{0i}, a_i, b_i, \sigma_i$ are also initialized in a similar fashion.

D.2 Exact-specified Settings

Under the exact-specified settings, we conduct 40 sample generations for each configuration, across a spectrum of 200 different sample sizes n ranging from 10^2 to 10^5. It can be seen from Figure 3a that the MLE $\widehat{G}_n$ empirically converges to the true mixing measure G_* under the Voronoi metric $\mathcal{D}_1$ at the rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$, which perfectly matches the theoretical parametric convergence rate established in Theorem 2.

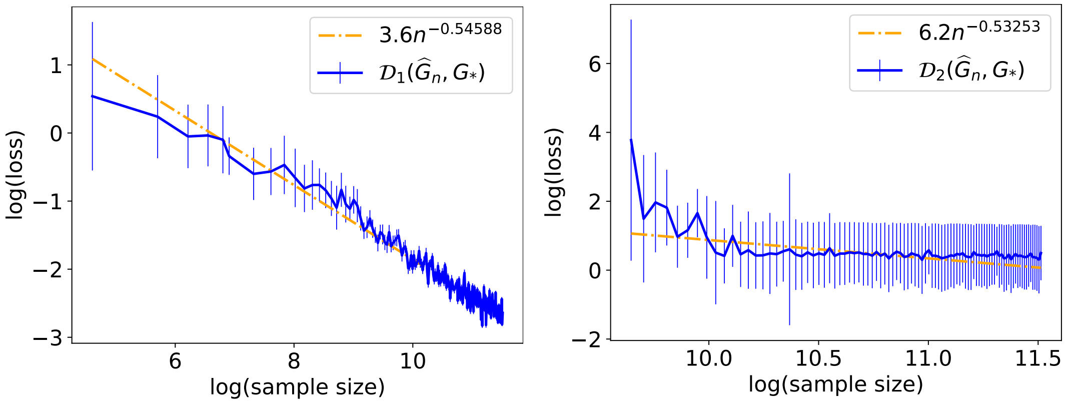
D.3 Over-specified Settings

Under the over-specified settings, we continue to generate 40 samples of size n for each setting, given 100 different choices of sample size $n \in [10^2, 10^5]$. As discussed in Section 3, to guarantee the convergence of density estimation to the true density, we need to select $\overline{K} = 2$ experts from the density estimation. As far as we know, existing works, namely Kwon et al. (2019); Kwon & Caramanis (2020); Kwon et al. (2021), only focus on the global convergence of the EM algorithm for parameter estimation under the input-free gating MoE, while that under the top-K sparse softmax gating MoE has remained poorly understood. Additionally, it is worth noting that the sample size must be sufficiently large so that the empirical convergence rate of the MLE returned by the EM algorithm aligns with the theoretical rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$ derived in Theorem 4.

E Additional Results

In this appendix, we study the convergence rates of parameter estimation under the model (1) when f is a probability density function of an arbitrary location-scale distribution. For that purpose, we first characterize the family of probability density functions of location-scale distributions

$$\mathcal{F} := \{f(Y|h_1(X, a, b), \sigma) : (a, b, \sigma) \in \Theta\}, \tag{36}$$



(a) Exact-specified settings with $K = 1$ (b) Over-specified settings with $K = 1$ and $\overline{K} = 2$

Figure 3: Log-log scaled plots illustrating simulation results under the exact-specified and the over-specified settings. We analyze the MLE $\widehat{G}_n$ across 40 independent samples, spanning sample sizes from 10^2 to 10^5. The blue curves depict the mean discrepancy between the MLE $\widehat{G}_n$ and the true mixing measure G_*, accompanied by error bars signifying two empirical standard deviations under the exact-specified settings. Additionally, an orange dash-dotted line represents the least-squares fitted linear regression line for these data points.

where $h_1(X, a, b) := a^\top X + b$ stands for the location, σ denotes the scale and Θ is a compact subset of $\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+$, based on the following notion of strong identifiability, which was previously studied in Manole & Ho (2022) and Ho & Nguyen (2016):

Definition 1 (Strong Identifiability)**.** *We say that the family $\mathcal{F}$ is strongly identifiable if the probability density function $f(Y|h_1(X, a, b), \sigma)$ is twice differentiable w.r.t its parameters and the following assumption holds true:*

For any $k \geq 1$ and k pairwise different tuples $(a_1, b_1, \sigma_1), \ldots, (a_k, b_k, \sigma_k) \in \Theta$, if there exist real coefficients $\alpha^{(i)}_{\ell_1,\ell_2}$, for $i \in [k_*]$ and $0 \leq \ell_1 + \ell_2 \leq 2$, such that

$$\sum_{i=1}^{k} \sum_{\ell_1+\ell_2=0}^{2} \alpha^{(i)}_{\ell_1,\ell_2} \cdot \frac{\partial^{\ell_1+\ell_2} f}{\partial h_1^{\ell_1} \partial \sigma^{\ell_2}}(Y|h_1(X, a_i, b_i), \sigma(X, \sigma_i)) = 0,$$

for almost surely (X, Y), then we obtain that $\alpha^{(i)}_{\ell_1,\ell_2} = 0$ for any $i \in [k_*]$ and $0 \leq \ell_1 + \ell_2 \leq 2$.

Example 1. *The families of Student's t-distributions and Laplace distributions are strongly identifiable, while the family of location-scale Gaussian distributions is not.*

In high level, we need to establish the Total Variation lower bound $\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))] \gtrsim \mathcal{D}(G, G_*)$ for any $G \in \mathcal{O}_k(\Omega)$. Then, this bound together with the density estimation rate in Theorem 1 (resp. Theorem 3) leads to the parameter estimation rates in Theorem 2 (resp. Theorem 4). Here, the key step is to decompose the difference $g_{\widehat{G}_n}(Y|X) - g_{G_*}(Y|X)$ into a combination of linearly independent terms using Taylor expansions. Therefore, we have to involve the above notion of strong identifiability, and separate our convergence analysis based on that notion.

Subsequently, since the convergence rates of maximum likelihood estimation when $\mathcal{F}$ is a family of location-scale Gaussian distributions, which is not strongly identifiable, have already been studied in Section 2 and Section 3, we will focus on the scenario when the family $\mathcal{F}$ is strongly identifiable in the sequel. Under that assumption, the density $f(Y|h_1, \sigma)$ is twice differentiable in (h_1, σ), therefore, it is also Lipschitz continuous. As a consequence, the density estimation rates under both the exact-specified and over-specified in Theorem 1 and Theorem 3 still hold true. Therefore, we aim to establish the parameter estimation rates under those settings in Appendix E.1 and Appendix E.2, respectively.

E.1 Exact-specified Settings

In this appendix, by using the Voronoi loss function $\mathcal{D}_1(G, G_*)$ defined in equation (5), we demonstrate in the following theorem that the rates for estimating true parameters $\exp(\beta_{0i}^*), \beta_{1i}^*, a_i^*, b_i^*, \sigma_i^*$ are of parametric order $\widetilde{\mathcal{O}}(n^{-1/2})$, which totally match those in Theorem 2.

Theorem 5. *Under the exact-specified settings, if the family $\mathcal{F}$ is strongly identifiable, then the Hellinger lower bound $\mathbb{E}_X[h(g_G(\cdot|X), g_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_1(G, G_*)$ holds for any mixing measure $G \in \mathcal{E}_{k_*}(\Omega)$. Consequently, we can find a universal constant $C_3 > 0$ depending only on G_*, Ω and K such that*

$$\mathbb{P}\Big(\mathcal{D}_1(\widehat{G}_n, G_*) > C_3\sqrt{\log(n)/n}\Big) \lesssim n^{-c_3},$$

where $c_3 > 0$ is a universal constant that depends only on Ω.

Proof of Theorem 5 is in Appendix E.3.1.

E.2 Over-specified Settings

In this appendix, we capture the convergence rates of parameter estimation under the over-specified settings when the family $\mathcal{F}$ is strongly identifiable.

Voronoi metric. It is worth noting that when $\mathcal{F}$ is strongly identifiable, the interaction among expert parameters in the second PDE (6) no longer holds true. As a result, it is not necessary to involve the solvability of the system (7) in the formulation of the Voronoi loss as in equation (8). Instead, let us consider the Voronoi loss function $\mathcal{D}_3(G, G_*)$ defined as

$$\begin{aligned}\mathcal{D}_3(G, G_*) := \max_{\{\ell_j\}_{j=1}^K \subset [k_*]} \Bigg\{ &\sum_{\substack{j\in[K],\\ |\mathcal{C}_{\ell_j}|=1}} \sum_{i\in\mathcal{C}_{\ell_j}} \exp(\beta_{0i})\Big[\|\Delta\beta_{1i\ell_j}\| + \|\Delta a_{i\ell_j}\| + |\Delta b_{i\ell_j}| + |\Delta\sigma_{i\ell_j}|\Big] \\ &+ \sum_{\substack{j\in[K],\\ |\mathcal{C}_{\ell_j}|>1}} \sum_{i\in\mathcal{C}_{\ell_j}} \exp(\beta_{0i})\Big[\|\Delta\beta_{1i\ell_j}\|^2 + \|\Delta a_{i\ell_j}\|^2 + |\Delta b_{i\ell_j}|^2 + |\Delta\sigma_{i\ell_j}|^2\Big] \\ &+ \sum_{j=1}^K \Big| \sum_{i\in\mathcal{C}_{\ell_j}} \exp(\beta_{0i}) - \exp(\beta_{0\ell_j}^*)\Big| \Bigg\}, \end{aligned} \quad (37)$$

for any mixing measure $G \in \mathcal{O}_k(\Omega)$. Equipped with loss function, we are ready to illustrate the convergence behavior of maximum likelihood estimation in the following theorem:

Theorem 6. *Under the over-specified settings, if the family $\mathcal{F}$ is strongly identifiable, then the Hellinger lower bound $\mathbb{E}_X[h(g_G(\cdot|X), g_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_2(G, G_*)$ holds for any mixing measure $G \in \mathcal{O}_k(\Omega)$. As a consequence, we can find a universal constant $C_4 > 0$ depending only on G_*, Ω and K such that*

$$\mathbb{P}\Big(\mathcal{D}_3(\widehat{G}_n, G_*) > C_4\sqrt{\log(n)/n}\Big) \lesssim n^{-c_4},$$

where $c_4 > 0$ is a universal constant that depends only on Ω.

Proof of Theorem 6 is in Appendix E.3.2. Theorem 6 indicates that true parameters $\beta_{1i}^*, a_i^*, b_i^*, \sigma_i^*$, which are fitted by a single component, share the same estimation rates of order $\widetilde{\mathcal{O}}(n^{-1/2})$ as those in Theorem 4. By contrast, the estimation rates for true parameters $\beta_{1i}^*, a_i^*, b_i^*, \sigma_i^*$, which are fitted by more than one component, are of order $\widetilde{\mathcal{O}}(n^{-1/4})$. Notably, these rates are no longer determined by the solvability of the system (7). Thus, they are significantly faster than those in Theorem 4. The main reason for this improvement is when $\mathcal{F}$ is strongly identifiable, the interaction among expert parameters via the second PDE in equation (6) does not occur.

E.3 Proofs of Additional Results

E.3.1 Proof of Theorem 5

Following from the result of Theorem 1 and since the Hellinger distance is lower bounded by the Total Variation distance, i.e. $h \geq V$, it is sufficient to demonstrate for any mixing measure $G \in \mathcal{O}_k(\Theta)$ that

$$\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_1(G, G_*),$$

which is then respectively broken into local part and global part as follows:

$$\inf_{G \in \mathcal{E}_{k_*}(\Omega): \mathcal{D}_1(G,G_*) \leq \varepsilon'} \frac{\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_1(G, G_*)} > 0, \tag{38}$$

$$\inf_{G \in \mathcal{E}_{k_*}(\Omega): \mathcal{D}_1(G,G_*) > \varepsilon'} \frac{\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_1(G, G_*)} > 0, \tag{39}$$

for some constant $\varepsilon' > 0$. Subsequently, we will prove only the local part (38), while the proof of the global part (39) can be done similarly to that in Appendix A.2.

Proof of claim (38): It is sufficient to show that

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{E}_{k_*}(\Omega): \mathcal{D}_1(G,G_*) \leq \varepsilon} \frac{\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_1(G, G_*)} > 0.$$

Assume that this inequality does not hold, then since the number of experts k_* is known in this case, there exists a sequence of mixing measure $G_n := \sum_{i=1}^{k_*} \exp(\beta_{0i}^n) \delta_{(\beta_{1i}^n, a_i^n, b_i^n, \sigma_i^n)} \in \mathcal{E}_{k_*}(\Omega)$ such that both $\mathcal{D}_1(G_n, G_*)$ and $\mathbb{E}_X[V(g_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_1(G_n, G_*)$ approach zero as n tends to infinity. Since $\mathcal{D}_1(G_n, G_*) \to 0$ as $n \to \infty$, each Voronoi cell contains only one element. Thus, we may assume WLOG that $\mathcal{C}_j = \{j\}$ for any $j \in [k_*]$, which implies that $(\beta_{1j}^n, a_j^n, b_j^n, \sigma_j^n) \to (\beta_{1j}^*, a_j^*, b_j^*, \sigma_j^*)$ and $\exp(\beta_{0j}^n) \to \exp(\beta_{0j}^*)$ as $n \to \infty$. WLOG, we assume that

$$\mathcal{D}_1(G_n, G_*) = \sum_{i=1}^{K} \Big[\exp(\beta_{0i}^n) \Big(\|\Delta \beta_{1i}^n\| + \|\Delta a_i^n\| + \|\Delta b_i^n\| + \|\Delta \sigma_i^n\| \Big) + \Big| \exp(\beta_{0i}^n) - \exp(\beta_{0i}^*) \Big| \Big],$$

where we denote $\Delta \beta_{1i}^n := \beta_{1i}^n - \beta_{1i}^*$, $\Delta a_i^n := a_i^n - a_i^*$, $\Delta b_i^n := b_i^n - b_i^*$ and $\Delta \sigma_i^n := \sigma_i^n - \sigma_i^*$.

Subsequently, by arguing in the same fashion as in Appendix A.2, we obtain that $\mathcal{X}_\ell^n = \mathcal{X}_\ell^*$, where

$$\mathcal{X}_\ell^n := \Big\{ x \in \mathcal{X} : (\beta_{1j}^n)^\top x \geq (\beta_{1j'}^n)^\top x : \forall j \in \{\ell_1, \ldots, \ell_K\}, j' \in \{\ell_{K+1}, \ldots, \ell_{k_*}\} \Big\},$$
$$\mathcal{X}_\ell^* := \Big\{ x \in \mathcal{X} : (\beta_{1j}^*)^\top x \geq (\beta_{1j'}^*)^\top x : \forall j \in \{\ell_1, \ldots, \ell_K\}, j' \in \{\ell_{K+1}, \ldots, \ell_{k_*}\} \Big\},$$

for any $\ell \in [q]$ for sufficiently large n.

Let $\ell \in [q]$ such that $\{\ell_1, \ldots, \ell_K\} = \{1, \ldots, K\}$. Then, for almost surely $(X, Y) \in \mathcal{X}_\ell^* \times \mathcal{Y}$, we can rewrite the conditional densities $g_{G_n}(Y|X)$ and $g_{G_*}(Y|X)$ as

$$g_{G_n}(Y|X) = \sum_{i=1}^{K} \frac{\exp((\beta_{1i}^n)^\top X + \beta_{0i}^n)}{\sum_{j=1}^{K} \exp((\beta_{1j}^n)^\top X + \beta_{0j}^n)} \cdot f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n),$$
$$g_{G_*}(Y|X) = \sum_{i=1}^{K} \frac{\exp((\beta_{1i}^*)^\top X + \beta_{0i}^*)}{\sum_{j=1}^{K} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)} \cdot f(Y|(a_i^*)^\top X + b_i^*, \sigma_i^*).$$

Now, we break the rest of our arguments into three steps:

Step 1 - Taylor expansion:

In this step, we take into account $H_n := \Big[\sum_{i=1}^{K}\exp((\beta_{1i}^*)^\top X+\beta_{0i}^*)\Big]\cdot[g_{G_n}(Y|X)-g_{G_*}(Y|X)]$. Note that the quantity H_n can be represented as follows:

$$
\begin{aligned}
H_n &= \sum_{i=1}^{K}\exp(\beta_{0i}^n)\Big[\exp((\beta_{1i}^n)^\top X)f(Y|(a_i^n)^\top X+b_i^n,\sigma_i^n)-\exp((\beta_{1i}^*)^\top X)f(Y|(a_i^*)^\top X+b_i^*,\sigma_i^*)\Big]\\
&-\sum_{i=1}^{K}\exp(\beta_{0i}^n)\Big[\exp((\beta_{1i}^n)^\top X)g_{G_n}(Y|X)-\exp((\beta_{1i}^*)^\top X)g_{G_n}(Y|X)\Big]\\
&+\sum_{i=1}^{K}\Big[\exp(\beta_{0i}^n)-\exp(\beta_{0i}^*)\Big]\Big[\exp((\beta_{1i}^*)^\top X+\beta_{0i}^*)f(Y|(a_i^*)^\top X+b_i^*,\sigma_i^*)-\exp((\beta_{1i}^*)^\top X)g_{G_n}(Y|X)\Big].
\end{aligned}
$$

By applying the first-order Taylor expansion to the first term in the above representation, which is denoted by A_n, we get that

$$
\begin{aligned}
A_n = \sum_{i=1}^{K}\sum_{|\alpha|=1}&\frac{\exp(\beta_{0i}^n)}{\alpha!}\cdot(\Delta\beta_{1i}^n)^{\alpha_1}(\Delta a_i^n)^{\alpha_2}(\Delta b_i^n)^{\alpha_3}(\Delta\sigma_i^n)^{\alpha_4}\\
&\times X^{\alpha_1+\alpha_2}\exp((\beta_{1i}^*)^\top X)\cdot\frac{\partial^{|\alpha_2|+\alpha_3+\alpha_4}f}{\partial h_1^{|\alpha_2|+\alpha_3}\partial\sigma^{\alpha_4}}(Y|(a_i^*)^\top X+b_i^*,\sigma_i^*)+R_1(X,Y),
\end{aligned}
$$

where $R_1(X,Y)$ is a Taylor remainder that satisfies $R_1(X,Y)/\mathcal{D}_1(X,Y)\to 0$ as $n\to\infty$. Let $\eta_1=\alpha_1+\alpha_2\in\mathbb{N}^d$, $\eta_2=|\alpha_2|+\alpha_3\in\mathbb{N}$ and $\eta_3=\alpha_4\in\mathbb{N}$, then we can rewrite A_n as follows:

$$
\begin{aligned}
A_n = \sum_{i=1}^{K}\sum_{\eta_3=0}^{1}\sum_{|\eta_1|+\eta_2=1-\eta_3}^{2-\eta_3}\sum_{\alpha\in\mathcal{I}_{\eta_1,\eta_2,\eta_3}}&\frac{\exp(\beta_{0i}^n)}{\alpha!}\cdot(\Delta\beta_{1i}^n)^{\alpha_1}(\Delta a_i^n)^{\alpha_2}(\Delta b_i^n)^{\alpha_3}(\Delta\sigma_i^n)^{\alpha_4}\\
&\times X^{\eta_1}\exp((\beta_{1i}^*)^\top X)\cdot\frac{\partial^{\eta_2+\eta_3}f}{\partial h_1^{\eta_2}\partial\sigma^{\eta_3}}(Y|(a_i^*)^\top X+b_i^*,\sigma_i^*)+R_1(X,Y),
\end{aligned}
\tag{40}
$$

where we define

$$
\mathcal{I}_{\eta_1,\eta_2,\eta_3} := \{(\alpha_i)_{i=1}^4\in\mathbb{N}^d\times\mathbb{N}^d\times\mathbb{N}\times\mathbb{N}:\alpha_1+\alpha_2=\eta_1,\ |\alpha_2|+\alpha_3=\eta_2,\ \alpha_4=\eta_3\}.
\tag{41}
$$

By arguing in a similar fashion for the second term in the representation of H_n, we also get that

$$
B_n := -\sum_{i=1}^{K}\sum_{|\gamma|=1}\frac{\exp(\beta_{0i}^n)}{\gamma!}(\Delta\beta_{1i}^n)\cdot X^{\gamma}\exp((\beta_{1i}^n)^\top X)g_{G_n}(Y|X)+R_2(X,Y),
$$

where $R_2(X,Y)$ is a Taylor remainder such that $R_2(X,Y)/\mathcal{D}_1(G_n,G_*)\to 0$ as $n\to\infty$. Putting the above results together, we rewrite the quantity H_n as follows:

$$
\begin{aligned}
H_n &= \sum_{i=1}^{K}\sum_{\eta_3=0}^{1}\sum_{|\eta_1|+\eta_2=0}^{2-\eta_3}U^n_{i,\eta_1,\eta_2,\eta_3}\cdot X^{\eta_1}\exp((\beta_{1i}^*)^\top X)\frac{\partial^{\eta_2+\eta_3}f}{\partial h_1^{\eta_2}\partial\sigma^{\eta_3}}(Y|(a_i^*)^\top X+b_i^*,\sigma_i^*)\\
&+\sum_{i=1}^{K}\sum_{0\leq|\gamma|\leq 1}W^n_{i,\gamma}\cdot X^{\gamma}\exp((\beta_{1i}^*)^\top X)g_{G_n}(Y|X)+R_1(X,Y)+R_2(X,Y),
\end{aligned}
\tag{42}
$$

in which we respectively define for each $i\in[K]$ that

$$
\begin{aligned}
U^n_{i,\eta_1,\eta_2,\eta_3} &:= \sum_{\alpha\in\mathcal{I}_{\eta_1,\eta_2,\eta_3}}\frac{\exp(\beta_{0i}^n)}{\alpha!}\cdot(\Delta\beta_{1i}^n)^{\alpha_1}(\Delta a_i^n)^{\alpha_2}(\Delta b_i^n)^{\alpha_3}(\Delta\sigma_i^n)^{\alpha_4},\\
W^n_{i,\gamma} &:= -\frac{\exp(\beta_{0i}^n)}{\gamma!}(\Delta\beta_{1i}^n)^{\gamma},
\end{aligned}
$$

for any $(\eta_1,\eta_2,\eta_3)\neq(\mathbf{0}_d,0,0)$ and $|\gamma|\neq\mathbf{0}_d$. Additionally, $U^n_{i,\mathbf{0}_d,0,0}=-W^n_{i,\mathbf{0}_d}:=\exp(\beta_{0i}^n)-\exp(\beta_{0i}^*)$.

Step 2 - Non-vanishing coefficients:

Moving to the second step, we will show that not all the ratios $U^n_{i,\eta_1,\eta_2,\eta_3}/\mathcal{D}_1(G_n,G_*)$ tend to zero as n goes to infinity. Assume by contrary that all of them approach zero when $n\to\infty$, then for $(\eta_1,\eta_2,\eta_3)=(\mathbf{0}_d,0,0)$, it follows that

$$\frac{1}{\mathcal{D}_1(G_n,G_*)}\cdot\sum_{i=1}^{K}\Big|\exp(\beta^n_{0i})-\exp(\beta^*_{0i})\Big|=\sum_{i=1}^{K}\frac{|U^n_{j,\eta_1,\eta_2,\eta_3}|}{\mathcal{D}_1(G_n,G_*)}\to 0. \tag{43}$$

Additionally, for tuples (η_1,η_2,η_3) where $\eta_1\in\{e_1,e_2,\ldots,e_d\}$ with $e_j:=(0,\ldots,0,\underbrace{1}_{j-th},0,\ldots,0)$ and $\eta_2=\eta_3=0$, we get

$$\frac{1}{\mathcal{D}_1(G_n,G_*)}\cdot\sum_{i=1}^{K}\exp(\beta^n_{0i})\|\Delta\beta^n_{1i}\|_1=\sum_{i=1}^{K}\sum_{\eta_1\in\{e_1,\ldots,e_d\}}\frac{|U^n_{j,\eta_1,0,0}|}{\mathcal{D}_1(G_n,G_*)}\to 0.$$

By using similar arguments, we end up with

$$\frac{1}{\mathcal{D}_1(G_n,G_*)}\cdot\sum_{i=1}^{K}\exp(\beta^n_{0i})\Big[\|\Delta\beta^n_{1i}\|_1+\|\Delta a^n_i\|_1+|\Delta b^n_i|+|\Delta\sigma^n_i|\Big]\to 0.$$

Due to the topological equivalence between norm-1 and norm-2, the above limit implies that

$$\frac{1}{\mathcal{D}_1(G_n,G_*)}\cdot\sum_{i=1}^{K}\exp(\beta^n_{0i})\Big[\|\Delta\beta^n_{1i}\|+\|\Delta a^n_i\|+|\Delta b^n_i|+|\Delta\sigma^n_i|\Big]\to 0. \tag{44}$$

Combine equation (43) with equation (44), we deduce that $\mathcal{D}_1(G_n,G_*)/\mathcal{D}_1(G_n,G_*)\to 0$, which is a contradiction. Consequently, at least one among the ratios $U^n_{i,\eta_1,\eta_2,\eta_3}/\mathcal{D}_1(G_n,G_*)$ does not vanish as n tends to infinity.

Step 3 - Fatou's contradiction:

Let us denote by m_n the maximum of the absolute values of $U^n_{i,\eta_1,\eta_2,\eta_3}/\mathcal{D}_1(G_n,G_*)$ and $W^n_{i,\gamma}/\mathcal{D}_1(G_n,G_*)$. It follows from the result achieved in Step 2 that $1/m_n\not\to\infty$.

Recall from the hypothesis that $\mathbb{E}_X[V(g_{G_n}(\cdot|X),g_{G_*}(\cdot|X))]/\mathcal{D}_1(G_n,G_*)\to 0$ as $n\to\infty$. Thus, by the Fatou's lemma, we have

$$0=\lim_{n\to\infty}\frac{\mathbb{E}_X[V(g_{G_n}(\cdot|X),g_{G_*}(\cdot|X))]}{\mathcal{D}_1(G_n,G_*)}=\frac{1}{2}\cdot\int\liminf_{n\to\infty}\frac{|g_{G_n}(Y|X)-g_{G_*}(Y|X)|}{\mathcal{D}_1(G_n,G_*)}\mathrm{d}X\mathrm{d}Y.$$

This result indicates that $|g_{G_n}(Y|X)-g_{G_*}(Y|X)|/\mathcal{D}_1(G_n,G_*)$ tends to zero as n goes to infinity for almost surely (X,Y). As a result, it follows that

$$\lim_{n\to\infty}\frac{H_n}{m_n\mathcal{D}(G_n,G_*)}=\Big[\sum_{i=1}^{K}\exp((\beta^*_{1i})^\top X+\beta^*_{0i})\Big]\cdot\lim_{n\to\infty}\frac{|g_{G_n}(Y|X)-g_{G_*}(Y|X)|}{m_n\mathcal{D}_1(G_n,G_*)}=0.$$

Next, let us denote $U^n_{i,\eta_1,\eta_2,\eta_3}/[m_n\mathcal{D}_1(G_n,G_*)]\to\tau_{i,\eta_1,\eta_2,\eta_3}$ and $W^n_{i,\gamma}/[m_n\mathcal{D}_1(G_n,G_*)]\to\kappa_{i,\gamma}$ with a note that at least one among them is non-zero. From the formulation of H_n in equation (42), we deduce that

$$\sum_{i=1}^{K}\sum_{\eta_3=0}^{1}\sum_{|\eta_1|+\eta_2=0}^{2-\eta_3}\tau_{i,\eta_1,\eta_2,\eta_3}\cdot X^{\eta_1}\exp((\beta^*_{1i})^\top X)\frac{\partial^{\eta_2+\eta_3}f}{\partial h_1^{\eta_2}\partial\sigma^{\eta_3}}(Y|(a^*_i)^\top X+b^*_i,\sigma^*_i)$$
$$+\sum_{i=1}^{K}\sum_{0\leq|\gamma|\leq 1}\kappa_{i,\gamma}\cdot X^{\gamma}\exp((\beta^*_{1i})^\top X)g_{G_*}(Y|X)=0,$$

for almost surely (X,Y). This equation is equivalent to

$$\sum_{|\eta_1|=0}^{1}\Big[\sum_{i=1}^{K}\sum_{\eta_2+\eta_3=0}^{2-|\eta_1|}\tau_{i,\eta_1,\eta_2,\eta_3}\exp((\beta^*_{1i})^\top X)\frac{\partial^{\eta_2+\eta_3}f}{\partial h_1^{\eta_2}\partial\sigma^{\eta_3}}(Y|(a^*_i)^\top X+b^*_i,\sigma^*_i)$$
$$+\kappa_{i,\eta_1}\exp((\beta^*_{1i})^\top X)g_{G_*}(Y|X)\Big]\times X^{\eta_1}=0,$$

for almost surely (X, Y). It is clear that the left hand side of the above equation is a polynomial of X belonging to the compact set $\mathcal{X}$. As a result, we get that

$$\sum_{i=1}^{K}\sum_{\eta_2+\eta_3=0}^{2-|\eta_1|}\tau_{i,\eta_1,\eta_2,\eta_3}\exp((\beta_{1i}^*)^\top X)\frac{\partial^{\eta_2+\eta_3}f}{\partial h_1^{\eta_2}\partial\sigma^{\eta_3}}(Y|(a_i^*)^\top X+b_i^*,\sigma_i^*)$$
$$+\kappa_{i,\eta_1}g_{G_*}(Y|X)\exp((\beta_{1i}^*)^\top X)=0,$$

for any $i\in[K]$, $0\leq|\eta_1|\leq 1$ and almost surely (X,Y). Since $(a_1^*,b_1^*,\sigma_1^*),\ldots,(a_K^*,b_K^*,\sigma_K^*)$ have pair-wise distinct values and the family $\mathcal{F}$ is strongly identifiable, the set

$$\Big\{\frac{\partial^{\eta_2+\eta_3}f}{\partial h_1^{\eta_2}\partial\sigma^{\eta_3}}(Y|(a_i^*)^\top X+b_i^*,\sigma_i^*):i\in[K],\ 0\leq\eta_2+\eta_3\leq 2-|\eta_1|\Big\}$$

is linearly independent w.r.t (X,Y). Consequently, we obtain that $\tau_{i,\eta_1,\eta_2,\eta_3}=\kappa_{i,\eta_1}=0$ for any $i\in[K]$, $0\leq|\eta_1|\leq 1$ and $0\leq\eta_2+\eta_3\leq 2-|\eta_1|$, which contradicts the fact that at least one among these terms is different from zero.

Hence, we reach the desired conclusion.

E.3.2 Proof of Theorem 6

Similar to the proof of Theorem 5 in Appendix E.3.1, our objective here is also to derive the local part of the following Total Variation lower bound:

$$\mathbb{E}_X[V(\overline{g}_G(\cdot|X),g_{G_*}(\cdot|X))]\gtrsim\mathcal{D}_3(G,G_*),$$

for any $G\in\mathcal{O}_k(\Theta)$. In particular, we aim to show that

$$\lim_{\varepsilon\to 0}\inf_{G\in\mathcal{O}_k(\Theta):\mathcal{D}_3(G,G_*)\leq\varepsilon}\frac{\mathbb{E}_X[V(\overline{g}_G(\cdot|X),g_{G_*}(\cdot|X))]}{\mathcal{D}_2(G,G_*)}>0. \tag{45}$$

Assume that the above claim does not hold true, then we can find a sequence of mixing measures $G_n:=\sum_{i=1}^{k_n}\exp(\beta_{0i}^n)\delta_{(\beta_{1i}^n,a_i^n,b_i^n,\sigma_i^n)}\in\mathcal{O}_k(\Omega)$ such that both $\mathcal{D}_3(G_n,G_*)$ and $\mathbb{E}_X[V(\overline{g}_{G_n}(\cdot|X),g_{G_*}(\cdot|X))]/\mathcal{D}_3(G_n,G_*)$ vanish when n goes to infinity. Then, it follows that for any $j\in[k_*]$, we have $\sum_{i\in\mathcal{C}_j}\exp(\beta_{0i}^n)\to\exp(\beta_{0j}^*)$ and $(\beta_{1i}^n,a_i^n,b_i^n,\sigma_i^n)\to(\beta_{1j}^*,a_j^*,b_j^*,\sigma_j^*)$ for all $i\in\mathcal{C}_j$. WLOG, we may assume that

$$\mathcal{D}_3(G_n,G_*)=\sum_{\substack{j\in[K],\\|\mathcal{C}_j|>1}}\sum_{i\in\mathcal{C}_j}\exp(\beta_{0i}^n)\Big[\|\Delta\beta_{1ij}^n\|^2+\|\Delta a_{ij}^n\|^2+|\Delta b_{ij}^n|^2+|\Delta\sigma_{ij}^n|^2\Big]$$
$$+\sum_{\substack{j\in[K],\\|\mathcal{C}_j|=1}}\sum_{i\in\mathcal{C}_j}\exp(\beta_{0i}^n)\Big[\|\Delta\beta_{1ij}^n\|+\|\Delta a_{ij}^n\|+|\Delta b_{ij}^n|+|\Delta\sigma_{ij}^n|\Big]+\sum_{j=1}^{K}\Big|\sum_{i\in\mathcal{C}_j}\exp(\beta_{0i}^n)-\exp(\beta_{0j}^*)\Big|.$$

Subsequently, let $X\in\mathcal{X}_\ell^*$ for some $\ell\in[q]$ such that $\{\ell_1,\ldots,\ell_K\}=\{1,\ldots,K\}$, where

$$\mathcal{X}_\ell^*:=\Big\{x\in\mathcal{X}:(\beta_{1j}^*)^\top x\geq(\beta_{1j'}^*)^\top x,\forall j\in\{\ell_1,\ldots,\ell_K\},j'\in\{\ell_{K+1},\ldots,\ell_{k_*}\}\Big\}.$$

Then, for any $\overline{\ell}\in[\overline{q}]$, we denote $(\overline{\ell}_1,\ldots,\overline{\ell}_k)$ as a permutation of $(1,\ldots,k)$ and

$$\mathcal{X}_{\overline{\ell}}^n:=\Big\{x\in\mathcal{X}:(\beta_{1i}^n)^\top x\geq(\beta_{1i'}^n)^\top x,\forall i\in\{\overline{\ell}_1,\ldots,\overline{\ell}_{\overline{K}}\},i'\in\{\overline{\ell}_{\overline{K}+1},\ldots,\overline{\ell}_k\}\Big\}.$$

If $\{\overline{\ell}_1,\ldots\overline{\ell}_{\overline{K}}\}\neq\mathcal{C}_1\cup\ldots\cup\mathcal{C}_K$ for any $\overline{\ell}\in[\overline{q}]$, then $V(\overline{g}_{G_n}(\cdot|X),g_{G_*}(\cdot|X))/\mathcal{D}_3(G_n,G_*)\not\to 0$ as n tends to infinity. This contradicts the fact that this term must approach zero. Therefore, we only need to consider the scenario when there exists $\overline{\ell}\in[\overline{q}]$ such that $\{\overline{\ell}_1,\ldots\overline{\ell}_{\overline{K}}\}=\mathcal{C}_1\cup\ldots\cup\mathcal{C}_K$. By using the same arguments as in Appendix B.2, we obtain that $X\in\mathcal{X}_{\overline{\ell}}^n$.

Then, we can represent the conditional densities $g_{G_*}(Y|X)$ and $g_{G_n}(Y|X)$ for any sufficiently large n as follows:

$$g_{G_*}(Y|X) = \sum_{j=1}^{K} \frac{\exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)}{\sum_{j'=1}^{K} \exp((\beta_{1j'}^*)^\top X + \beta_{0j'}^*)} \cdot f(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*),$$

$$\overline{g}_{G_n}(Y|X) = \sum_{j=1}^{K} \sum_{i \in \mathcal{C}_j} \frac{\exp((\beta_{1i}^n)^\top X + \beta_{0i}^n)}{\sum_{j'=1}^{K} \sum_{i' \in \mathcal{C}_{j'}} \exp((\beta_{1i'}^n)^\top X + \beta_{0i'}^n)} \cdot f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n).$$

Now, we reuse the three-step framework in Appendix E.3.1.

Step 1 - Taylor expansion:

Firstly, by abuse of notations, let us consider the quantity

$$H_n := \Big[\sum_{j=1}^{K} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)\Big] \cdot [\overline{g}_{G_n}(Y|X) - g_{G_*}(Y|X)].$$

Similar to Step 1 in Appendix A, we can express this term as

$$\begin{aligned}
H_n &= \sum_{j=1}^{K} \sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n) \Big[\exp((\beta_{1i}^n)^\top X) f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n) - \exp((\beta_{1j}^*)^\top X) f(Y|(a_j^*)^\top X + b_j^*, \sigma_j^*)\Big] \\
&- \sum_{j=1}^{K} \sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n) \Big[\exp((\beta_{1i}^n)^\top X) g_{G_n}(Y|X) - \exp((\beta_{1j}^*)^\top X) g_{G_n}(Y|X)\Big] \\
&+ \sum_{j=1}^{K} \Big[\sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^*)\Big] \Big[\exp((\beta_{1j}^*)^\top X) f(Y|(a_i^*)^\top X + b_i^*, \sigma_i^*) - \exp((\beta_{1j}^*)^\top X) g_{G_n}(Y|X)\Big] \\
&:= A_n + B_n + E_n.
\end{aligned}$$

Next, we proceed to decompose A_n based on the cardinality of the Voronoi cells as follows:

$$\begin{aligned}
A_n = &\sum_{j:|\mathcal{C}_j|=1} \sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n) \Big[\exp((\beta_{1i}^n)^\top X) f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n) - \exp((\beta_{1i}^*)^\top X) f(Y|(a_i^*)^\top X + b_i^*, \sigma_i^*)\Big] \\
&+ \sum_{j:|\mathcal{C}_j|>1} \sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n) \Big[\exp((\beta_{1i}^n)^\top X) f(Y|(a_i^n)^\top X + b_i^n, \sigma_i^n) - \exp((\beta_{1i}^*)^\top X) f(Y|(a_i^*)^\top X + b_i^*, \sigma_i^*)\Big].
\end{aligned}$$

By applying the Taylor expansions of first and second orders to the first and second terms of A_n, respectively, and following the derivation in equation (40), we arrive at

$$\begin{aligned}
A_n = &\sum_{j:|\mathcal{C}_j|=1} \sum_{i \in \mathcal{C}_j} \sum_{\eta_3=0}^{1} \sum_{|\eta_1|+\eta_2=1-\eta_3}^{2-\eta_3} \sum_{\alpha \in \mathcal{I}_{\eta_1,\eta_2,\eta_3}} \frac{\exp(\beta_{0i}^n)}{\alpha!} \cdot (\Delta \beta_{1ij}^n)^{\alpha_1} (\Delta a_{ij}^n)^{\alpha_2} (\Delta b_{ij}^n)^{\alpha_3} (\Delta \sigma_i^n)^{\alpha_4} \\
&\qquad \times X^{\eta_1} \exp((\beta_{1j}^*)^\top X) \cdot \frac{\partial^{\eta_2+\eta_3} f}{\partial h_1^{\eta_2} \partial \sigma^{\eta_3}} (Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) + R_3(X,Y) \\
&+ \sum_{j:|\mathcal{C}_j|>1} \sum_{i \in \mathcal{C}_j} \sum_{\eta_3=0}^{2} \sum_{|\eta_1|+\eta_2=1-\mathbf{1}_{\{\eta_3>0\}}}^{4-\eta_3} \sum_{\alpha \in \mathcal{I}_{\eta_1,\eta_2,\eta_3}} \frac{\exp(\beta_{0i}^n)}{\alpha!} \cdot (\Delta \beta_{1ij}^n)^{\alpha_1} (\Delta a_{ij}^n)^{\alpha_2} (\Delta b_{ij}^n)^{\alpha_3} (\Delta \sigma_{ij}^n)^{\alpha_4} \\
&\qquad \times X^{\eta_1} \exp((\beta_{1j}^*)^\top X) \cdot \frac{\partial^{\eta_2+\eta_3} f}{\partial h_1^{\eta_2} \partial \sigma^{\eta_3}} (Y|(a_j^*)^\top X + b_j^*, \sigma_j^*) + R_4(X,Y),
\end{aligned}$$

where the set $\mathcal{I}_{\eta_1,\eta_2,\eta_3}$ is defined in equation (41) while $R_i(X,Y)$ is a Taylor remainder such that $R_i(X,Y)/\mathcal{D}_3(G_n, G_*) \to 0$ as $n \to \infty$ for $i \in \{3, 4\}$. Similarly, we also decompose B_n as

$$\begin{aligned}
B_n = &- \sum_{j:|\mathcal{C}_j|=1} \sum_{i \in \mathcal{C}_j} \sum_{|\gamma|=1} \frac{\exp(\beta_{0i}^n)}{\gamma!} (\Delta \beta_{1i}^n) \cdot X^{\gamma} \exp((\beta_{1j}^*)^\top X) g_{G_n}(Y|X) + R_5(X,Y) \\
&- \sum_{j:|\mathcal{C}_j|>1} \sum_{i \in \mathcal{C}_j} \sum_{1 \le |\gamma| \le 2} \frac{\exp(\beta_{0i}^n)}{\gamma!} (\Delta \beta_{1i}^n) \cdot X^{\gamma} \exp((\beta_{1j}^*)^\top X) g_{G_n}(Y|X) + R_6(X,Y),
\end{aligned}$$

where $R_5(X,Y)$ and $R_6(X,Y)$ are Taylor remainders such that their ratios over $\mathcal{D}_3(G_n,G_*)$ approach zero as $n\to\infty$. Subsequently, let us define

$$S^n_{j,\eta_1,\eta_2,\eta_3}:=\sum_{i\in\mathcal{C}_j}\sum_{\alpha\in\mathcal{I}_{\eta_1,\eta_2,\eta_3}}\frac{\exp(\beta^n_{0i})}{\alpha!}\cdot(\Delta\beta^n_{1ij})^{\alpha_1}(\Delta a^n_{ij})^{\alpha_2}(\Delta b^n_{ij})^{\alpha_3}(\Delta\sigma^n_{ij})^{\alpha_4},$$

$$T^n_{j,\gamma}:=-\sum_{i\in\mathcal{C}_j}\frac{\exp(\beta^n_{0i})}{\gamma!}(\Delta\beta^n_{1ij})^{\gamma}=-S^n_{j,\gamma,0,0},$$

for any $(\eta_1,\eta_2,\eta_3)\neq(\mathbf{0}_d,0,0)$ and $|\gamma|\neq\mathbf{0}_d$, while for $(\eta_1,\eta_2,\eta_3)=(\mathbf{0}_d,0)$ we set

$$S^n_{i,\mathbf{0}_d,0}=-T^n_{i,\mathbf{0}_d}:=\sum_{i\in\mathcal{C}_j}\exp(\beta^n_{0i})-\exp(\beta^*_{0i}).$$

As a consequence, it follows that

$$\begin{aligned}H_n=&\sum_{j=1}^{K}\sum_{\eta_3=0}^{1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}}}\sum_{|\eta_1|+\eta_2=0}^{2(1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}})-\eta_3}S^n_{j,\eta_1,\eta_2,\eta_3}\cdot X^{\eta_1}\exp((\beta^*_{1j})^\top X)\cdot\frac{\partial^{\eta_2}f}{\partial h_1^{\eta_2}}(Y|(a^*_j)^\top X+b^*_j,\sigma^*_j)\\&+\sum_{j=1}^{K}\sum_{|\gamma|=0}^{1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}}}T^n_{j,\gamma}\cdot X^{\gamma}\exp((\beta^*_{1j})^\top X)g_{G_n}(Y|X)+R_5(X,Y)+R_6(X,Y).\end{aligned}\tag{46}$$

Step 2 - Non-vanishing coefficients:

In this step, we will prove by contradiction that at least one among the ratios $S^n_{j,\eta_1,\eta_2,\eta_3}/\mathcal{D}_3(G_n,G_*)$ does not converge to zero as $n\to\infty$. Assume that all these terms go to zero, then by employing arguments for deriving equations (43) and (44), we get that

$$\begin{aligned}\frac{1}{\mathcal{D}_3(G_n,G_*)}\cdot\Big[\sum_{j=1}^{K}\Big|\sum_{i\in\mathcal{C}_j}\exp(\beta^n_{0i})-\exp(\beta^*_{0j})\Big|&\\+\sum_{j:|\mathcal{C}_j|=1}\sum_{i\in\mathcal{C}_j}\exp(\beta^n_{0i})\Big(\|\Delta\beta^n_{1ij}\|+\|\Delta a^n_{ij}\|+|\Delta b^n_{ij}|+|\Delta\sigma^n_{ij}|\Big)\Big]&\to 0.\end{aligned}$$

Next, let $e_j:=(0,\ldots,0,\underbrace{1}_{j-th},0,\ldots,0)$ for any $j\in[d]$. Then, we have

$$\frac{1}{\mathcal{D}_3(G_n,G_*)}\cdot\sum_{i=1}^{K}\exp(\beta^n_{0i})\|\Delta\beta^n_{1ij}\|^2=\sum_{i=1}^{K}\sum_{\eta_1\in\{2e_1,\ldots,2e_d\}}\frac{|U^n_{j,\eta_1,0,0}|}{\mathcal{D}_3(G_n,G_*)}\to 0.$$

Similarly, we also get that

$$\frac{1}{\mathcal{D}_3(G_n,G_*)}\cdot\sum_{i=1}^{K}\exp(\beta^n_{0i})\|\Delta b^n_{ij}\|^2\to 0,\quad\frac{1}{\mathcal{D}_3(G_n,G_*)}\cdot\sum_{i=1}^{K}\exp(\beta^n_{0i})\|\Delta\sigma^n_{ij}\|^2\to 0$$

Moreover, note that

$$\frac{1}{\mathcal{D}_3(G_n,G_*)}\cdot\sum_{i=1}^{K}\exp(\beta^n_{0i})\|\Delta a^n_{ij}\|^2=\sum_{i=1}^{K}\sum_{\eta_1\in\{2e_1,\ldots,2e_d\}}\frac{|U^n_{j,\eta_1,2,0}|}{\mathcal{D}_3(G_n,G_*)}\to 0.$$

Gathering all the above limits, we obtain that $1=\mathcal{D}_3(G_n,G_*)/\mathcal{D}_3(G_n,G_*)\to 0$ as $n\to\infty$, which is a contradiction. Thus, at least one among the terms $S^n_{j,\eta_1,\eta_2,\eta_3}/\mathcal{D}_3(G_n,G_*)$ does not converge to zero as $n\to\infty$

Step 3 - Fatou's contradiction:

It follows from the hypothesis that $\mathbb{E}_X[V(\overline{g}_{G_n}(\cdot|X),g_{G_*}(\cdot|X))]/\mathcal{D}_3(G_n,G_*)\to 0$ as $n\to\infty$. Then, by applying the Fatou's lemma, we get

$$0=\lim_{n\to\infty}\frac{\mathbb{E}_X[V(\overline{g}_{G_n}(\cdot|X),g_{G_*}(\cdot|X))]}{\mathcal{D}_3(G_n,G_*)}=\frac{1}{2}\cdot\int\liminf_{n\to\infty}\frac{|\overline{g}_{G_n}(Y|X)-g_{G_*}(Y|X)|}{\mathcal{D}_3(G_n,G_*)}\mathrm{d}X\mathrm{d}Y,$$

which implies that $|\overline{g}_{G_n}(Y|X) - g_{G_*}(Y|X)|/\mathcal{D}_3(G_n, G_*) \to 0$ as $n \to \infty$ for almost surely (X, Y).

Next, we define m_n as the maximum of the absolute values of $S^n_{j,\eta_1,\eta_2/\mathcal{D}_3(G_n,G_*)}$. It follows from Step 2 that $1/m_n \not\to \infty$. Moreover, by arguing in the same way as in Step 3 in Appendix E.1, we receive that

$$H_n/[m_n\mathcal{D}_3(G_n, G_*)] \to 0 \tag{47}$$

as $n \to \infty$. By abuse of notations, let us denote

$$S^n_{j,\eta_1,\eta_2,\eta_3}/[m_n\mathcal{D}_3(G_n, G_*)] \to \tau_{j,\eta_1,\eta_2,\eta_3}$$

Here, at least one among $\tau_{j,\eta_1,\eta_2,\eta_3}$ is non-zero. Then, by putting the results in equations (46) and (47) together, we get

$$\sum_{j=1}^{K} \sum_{\eta_3=0}^{1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}}} \sum_{|\eta_1|+\eta_2=0}^{2(1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}})-\eta_3} \tau_{j,\eta_1,\eta_2,\eta_3} \cdot X^{\eta_1} \exp((\beta^*_{1j})^\top X) \frac{\partial^{\eta_2+\eta_3} f}{\partial h_1^{\eta_2} \partial \sigma^{\eta_3}} (Y|(a^*_j)^\top X + b^*_j, \sigma^*_j)$$
$$+ \sum_{j=1}^{K} \sum_{|\gamma|=0}^{1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}}} -\tau^n_{j,\gamma,0,0} \cdot X^{\gamma} \exp((\beta^*_{1j})^\top X) g_{G_*}(Y|X) = 0,$$

for almost surely (X, Y). Arguing in a similar fashion as in Step 3 of Appendix E.1, we obtain that $\tau_{j,\eta_1,\eta_2,\eta_3} = 0$ for any $j \in [K]$, $0 \leq |\eta_1| + \eta_2 + \eta_3 \leq 2(1 + \mathbf{1}_{\{|\mathcal{C}_j|>1\}})$ and $0 \leq |\gamma| \leq 1 + \mathbf{1}_{\{|\mathcal{C}_j|>1\}}$. This contradicts the fact that at least one among them is non-zero. Hence, the proof is completed.