On DeepSeekMoE: Statistical Benefits of Shared Experts and Normalized Sigmoid Gating

Huy Nguyen[†] Thong T. Doan[⋄] Quang Pham[‡]
Nghi D. Q. Bui[⋄] Nhat Ho[†,⋆] Alessandro Rinaldo[†,⋆]

[†]The University of Texas at Austin
[⋄]FPT Software AI Center
[‡] Independent Researcher

June 12, 2025

Abstract

Mixture of experts (MoE) methods are a key component in most large language model architectures, including the recent series of DeepSeek models. Compared to other MoE implementations, DeepSeekMoE stands out because of two unique features: the deployment of a shared expert strategy and of the normalized sigmoid gating mechanism. Despite the prominent role of DeepSeekMoE in the success of the DeepSeek series of models, there have been only a few attempts to justify theoretically the value of the shared expert strategy, while its normalized sigmoid gating has remained unexplored. To bridge this gap, we undertake a comprehensive theoretical study of these two features of DeepSeekMoE from a statistical perspective. We perform a convergence analysis of the expert estimation task to highlight the gains in sample efficiency for both the shared expert strategy and the normalized sigmoid gating, offering useful insights into the design of expert and gating structures. To verify empirically our theoretical findings, we carry out several experiments on both synthetic data and real-world datasets for (vision) language modeling tasks. Finally, we conduct an extensive empirical analysis of the router behaviors, ranging from router saturation, router change rate, to expert utilization.

1 Introduction

The recent years have witnessed a dramatic increase in the use and success of of deep learning models, leading to remarkable advances in a variety of fields, namely natural language processing [30, 18, 20, 38], computer vision [62, 41], multimodal learning [23, 79], and reinforcement learning [4, 10]. However, this trend has also introduced several challenges in terms of computational efficiency. One common approach to tackle this challenge is to leverage Mixture-of-Experts (MoE) architecture, which allows to scale up the model capacity without a proportional increase in computation.

Originally proposed by Jacob et al. [28], MoE has been known as a form of ensemble learning that combines the power of several individual models through an adaptive gating network. In particular, these individual models are termed experts and can be formulated as classifiers [8, 52], regression models [19, 34], or feed-forward networks (FFNs) [64, 12]. Meanwhile, the gating network is responsible for dynamically assigning input-dependent softmax weights to experts based on their

[⋆] Co-last authors.

specialization in the input domain. Then, to improve the scalability of MoE, Shazeer et al. [64] have recently introduced a sparse version of MoE which activates only a subset of specialized experts per input, allowing to increase the number of trainable parameters while keeping the computation overhead nearly unchanged. As a result, there has been a surge of interest in employing the sparse MoE architecture in several large-scale applications, particularly language modeling [15, 14, 22, 69, 60].

Despite their widespread use in large language models, the sparse MoE architecture faces the challenge of knowledge redundancy, that is, multiple experts may end up acquiring overlapping knowledge, leading to the redundancy of expert parameters. In response to this issue, Dai et al. [12] have come up with a novel DeepSeekMoE framework that divides the set of experts into two disjoint subsets. Experts in the first subset are referred to as shared experts and are always activated to capture common knowledge across different domains. On the other hand, only few experts in the second subset, called routed experts, are activated, typically via a sparse softmax gating mechanism to learn specialized knowledge. This shared expert strategy helps enhance expert specialization by encouraging experts to specialize in distinctive aspects of the data, thereby alleviating the parameter redundancy problem. The new DeepSeekMoE architecture has been adopted as a vital component in the series of high-performing DeepSeek language models, most notably DeepSeek-V2 [14] which uses sparse softmax gating in the DeepSeekMoE framework, and DeepSeek-V3 [15], which employs a sparse normalized sigmoid gating. Surprisingly, the shared expert strategy has only been briefly investigated in [12] from the perspective of expert specialization, while there have been no studies on the benefits of the normalized sigmoid gating.

Contributions. The primary goal of this paper is to provide a comprehensive theoretical study of these two distinguishing features of DeepSeekMoE. Below we perform a convergence analysis of the task of parameter estimation in order to examine the sample efficiency of the shared expert strategy, that is the rate, as a function of the number of data points, at which each expert to specialize in some aspects of the data. Furthermore, we also compare the sample efficiency of the normalized sigmoid gating used in the DeepSeek-V3 model to that of the softmax gating used in the DeepSeek-V2 model. Our contributions are threefold and can be summarized as follows.

(1) Sample efficiency of the shared expert strategy. Our analysis in Section 2 reveals that shared experts admit significantly faster convergence rates than routed experts and experts in MoE models without the shared expert strategy, whose rates depend in a complicated manner on the solvability of certain systems of polynomial equations as well as the number of fitted experts (see Table 1). As a result, a smaller amount of data are required to approximate shared experts compared to non-shared experts in DeepSeekMoE and standard MoE models to achieve the same level of statistical accuracy.

(2) Sample efficiency of the normalized sigmoid gating. Similarly, when using the normalized sigmoid gating instead of the softmax gating, the convergence rates of routed experts no longer hinge on the solvability of a system of polynomial equations and, therefore, are provably faster than those of shared experts, which remain unchanged in this setting (see also Table 1). Thus, the amount of data required to estimate routed experts within a given error decreases substantially, demonstrating the sample efficiency of the normalized sigmoid gating over the standard softmax gating. Due to space limitations and the technical nature of this analysis, we present these results in Appendix A.

Table 1: Summary of expert estimation rates in DeepSeek-V2's MoE with softmax gating (Section 2) and DeepSeek-V3's MoE with normalized sigmoid gating (Appendix A). Below, the function r_2 stands for the solvability of certain systems of polynimial equations specified in Appendix B, while the notation $\mathcal{V}_{2,j}$ denotes a Voronoi cell defined in equation (3). For the normalized sigmoid gating setting, we consider two complementary parameter settings, namely sparse regime and dense regime (see Appendix A for further details).

DeepSeek-V2's MoE	**ReLU FFN Experts**	**Linear Experts**
Shared Experts	$\widetilde{\mathcal{O}}_P(n^{-1/4})$	$\widetilde{\mathcal{O}}_P(n^{-1/4})$
Routed Experts	$\widetilde{\mathcal{O}}_P(n^{-1/4})$	$\widetilde{\mathcal{O}}_P(n^{-1/r_2(\lvert\mathcal{V}_{2,j}\rvert)})$

DeepSeek-V3's MoE	**ReLU FFN Experts**		**Linear Experts**	
	Sparse Regime	Dense Regime	Sparse Regime	Dense Regime
Shared Experts	$\widetilde{\mathcal{O}}_P(n^{-1/4})$		$\widetilde{\mathcal{O}}_P(n^{-1/4})$	
Routed Experts	$\widetilde{\mathcal{O}}_P(n^{-1/4})$	$\widetilde{\mathcal{O}}_P(n^{-1/2})$	$\widetilde{\mathcal{O}}_P(n^{-1/r_2(\lvert\mathcal{V}_{2,j}\rvert)})$	$\widetilde{\mathcal{O}}_P(n^{-1/2})$

(3) Empirical validation. To validate our theoretical findings, we conduct extensive numerical experiments on simulated and real-world data. The experimental results on synthetic data are in very close agreement with our theoretical findings about the convergence rates of the shared expert strategy and the normalized sigmoid gating; see Appendix G for detailed results. The experiments summarized in Section 3 on language modeling and vision-language modeling further demonstrate the applicability of our theoretical insights in real-world scenarios. Finally, we perform a detailed analysis of router behavior in Section 3.3, providing further insights into the contribution and dynamics of each component of the DeepSeekMoE architecture.

Notation. For any $n \in \mathbb{N}$, we let $[n] = \{1, 2, \ldots, n\}$. For any vectors $v := (v_i)_{i=1}^d \in \mathbb{R}^d$ and $\alpha := (\alpha_i)_{i=1}^d \in \mathbb{N}^d$, we denote $v^\alpha := \prod_{i=1}^d v_i^{\alpha_i}$, $|v| := \sum_{i=1}^d v_i$ and $\alpha! := \prod_{i=1}^d \alpha_i!$, while $\|v\|$ represents the ℓ_2-norm of v. The cardinality of a set S is denoted with $|S|$. Finally, for any two positive sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, we write $a_n = \mathcal{O}(b_n)$ or $a_n \lesssim b_n$ if $a_n \leq C b_n$ for all $n \in \mathbb{N}$, for some constant $C > 0$. For a sequence $(A_n)_{n\geq 1}$ of positive random variables, the notation $A_n = \mathcal{O}_P(b_n)$ signifies A_n/b_n is stochastically bounded, that is, for any $\epsilon > 0$, there exists an $M > 0$ such that $\mathbb{P}(A_n/b_n > M) < \epsilon$ for all n large enough. We further write $A_n = \widetilde{\mathcal{O}}_P(b_n)$ when $A_n = \mathcal{O}_P(b_n \log^c(b_n))$, for some $c > 0$. Finally, for two Lebesgue probability densities on $\mathbb{R}^d$, f_1 and f_2, $V(f_1, f_2) := \frac{1}{2}\int |f_1(y) - f_2(y)| dy$ denotes their total variation distance.

2 On Shared Expert Strategy

Below we derive convergence rates for the shared expert estimation problem in the DeepSeekMoE architecture. For ease of presentation, we will focus here on the dense DeepSeekMoE case, and analyze the less popular sparse DeepSeekMoE settings in Appendix F. After formally introducing the settings, we formulate a *strong identifiability* condition on the expert functions ensuring fast expert convergence rates. We then turn to linear experts, which violate the strong identifiability condition, and prove that, in fact, they exhibit slow rates of convergence.

Problem setting. Assume that $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ are i.i.d. samples drawn from a Gaussian DeepSeekMoE model, whose conditional density function $f_{G_1^*,G_2^*}(y|x)$ is given by

$$f_{G_1^*,G_2^*}(y|x) :=\frac{1}{2}\sum_{i=1}^{k_1^*}\omega_i^*\pi(Y|h_1(x,\kappa_i^*),\tau_i^*) + \frac{1}{2}\sum_{i=1}^{k_2^*}\frac{\exp((\beta_{1i}^*)^\top x+\beta_{0i}^*)}{\sum_{j=1}^{k_2^*}\exp((\beta_{1j}^*)^\top x+\beta_{0j}^*)}\pi(y|h_2(x,\eta_i^*),\nu_i^*). \quad (1)$$

Above, $\pi(\cdot|\mu,\nu)$ denotes the Gaussian density function with mean μ and variance ν, $h_1(\cdot,\kappa_i^*)$ and $h_2(\cdot,\eta_i^*)$ are real-valued functions on $\mathbb{R}^d$ referred to as *shared* and *routed* experts, respectively. The weight parameters $\omega_1^*, \omega_2^*, \ldots, \omega_{k_1^*}^*$ are positive and satisfy $\sum_{i=1}^{k_1^*}\omega_i^* = 1$. We conveniently represent all the model parameters with the *mixing measures* $G_1^* := \sum_{i=1}^{k_1^*}\omega_i^*\delta_{(\kappa_i^*,\tau_i^*)}$ and $G_2^* := \sum_{i=1}^{k_2^*}\exp(\beta_{0i}^*)\delta_{(\beta_{1i}^*,\eta_i^*,\nu_i^*)}$, a combination of Dirac δ-measures with mass on the unknown true parameters $\theta_{1i}^* := (\omega_i^*, \kappa_i^*, \tau_i^*)$ in $\Theta_1 \subseteq \mathbb{R}\times\mathbb{R}^{d_1}\times\mathbb{R}_+$ and $\theta_{2i}^* := (\beta_{0i}^*, \beta_{1i}^*, \eta_i^*, \nu_i^*)$ in $\Theta_2 \subseteq \mathbb{R}\times\mathbb{R}^d\times\mathbb{R}^{d_2}\times\mathbb{R}_+$, respectively. Thus, our goal is to estimate the pair of ground-truth mixing measures (G_1^*, G_2^*).

Maximum likelihood estimation (MLE). As the numbers k_1^* and k_2^* of shared and routed experts are unknown, we consider the ground-truth model (1) with up to $k_1 > k_1^*$ shared experts and $k_2 > k_2^*$ routed experts. Towards that goal, we let $\mathcal{G}_{k_1,k_2}(\Theta) := \mathcal{G}_{k_1}(\Theta_1)\times\mathcal{G}_{k_2}(\Theta)$ stands for the set of mixing measure pairs (G_1, G_2) with at most k_1 and k_2 atoms, respectively; that is $\mathcal{G}_{k_1}(\Theta_1) := \Big\{G_1 = \sum_{i=1}^{k_1'}\omega_i\delta_{(\kappa_i,\tau_i)} : 1 \leq k_1' \leq k_1\Big\}$ and $\mathcal{G}_{k_2}(\Theta_2) := \Big\{G_2 = \sum_{i=1}^{k_2'}\exp(\beta_{0i})\delta_{(\beta_{1i},\eta_i^*,\nu_i^*)} : 1 \leq k_2' \leq k_2\Big\}$. Our final estimator is the MLE over $\mathcal{G}_{k_1,k_2}(\Theta)$, i.e.

$$(\widehat{G}_1^n, \widehat{G}_2^n) \in \underset{(G_1,G_2)\in\mathcal{G}_{k_1,k_2}(\Theta)}{\arg\max}\ \frac{1}{n}\sum_{i=1}^{n}\log(f_{G_1,G_2}(Y_i|X_i)), \quad (2)$$

Universal assumptions. For our theoretical analysis, we impose the following three mild assumptions on the ground-truth parameters throughout the paper.

(A.1) The parameter space Θ is compact with fixed dimension, while the input space $\mathcal{X}$ is bounded.

(A.2) The last pair of gating parameters vanish, that is, $\beta_{1k_2^*}^* = 0_d$ and $\beta_{0k_2^*}^* = 0$ (to avoid non-identifiability due to invariance to translation of the softmax gating function). In addition, at least one among parameters $\{\beta_{1i}^*, i \in [k_2^*]\}$, is non-zero (to maintain the dependence of the gating on the input value).

(A.3) The expert parameters $(\kappa_i^*)_{i=1}^{k_1^*}$ and $(\eta_i^*)_{i=1}^{k_2^*}$ are distinct. Meanwhile, the expert functions $h_1(\cdot,\kappa)$ and $h_2(\cdot,\eta)$ are bounded and Lipschitz continuous w.r.t κ and η.

Equipped with these assumptions, we are now ready to give our first consistency result for the ground-truth conditional density $f_{G_1^*,G_2^*}$.

Proposition 1. *The maximum likelihood density estimator $f_{\widehat{G}_1^n,\widehat{G}_2^n}(Y|X)$ converges to the true density $f_{G_1^*,G_2^*}(Y|X)$ in total variation distance at the rate*

$$\mathbb{E}_X[V(f_{\widehat{G}_1^n,\widehat{G}_2^n}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X))] = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}}).$$

The above result, whose proof can be found in Appendix E.1, shows that the true density function $f_{G_1^*,G_2^*}(y|x)$ can be estimated at a rate that is nearly parametric. Following a strategy used in the analysis of MoE models [56], if one can exhibit an appropriate loss function over the mixing measures, say $\mathcal{D}((G_1,G_2),(G_2^*,G_2^*))$, that, up to constant, is a lower bound on $\mathbb{E}_X[V(f_{\widehat{G}_1^n,\widehat{G}_2^n}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X))]$, Proposition 1 will then imply a near parametric rate also for the parameters and expert functions themselves. However, the derivation of this lower bound is challenging. Specifically, a key step in establishing the aforementioned lower bound is to decompose the difference $f_{\widehat{G}_1^n,\widehat{G}_2^n}(Y|X) - f_{G_1^*,G_2^*}(Y|X)$ through a series of Taylor expansions of the functions $x \mapsto \pi(Y|h_1(x,\kappa),\tau)$ and $x \mapsto F(Y|X;\beta_1,\eta,\nu) := \exp(\beta_1^\top x)\pi(Y|h_2(x,\eta),\nu)$ w.r.t their parameters (κ,τ) and (β_1,η,ν), respectively. When the difference of the densities converges to zero (as ensured by Proposition 1), then one may expect the coefficients of this Taylor expansions, which correspond to the difference between the true and estimated parameters, will also vanish. However, this is true only provided that these functions and their partial derivatives arising from the Taylor expansions remain linearly independent. To ensure that this property holds, we formulate a new, non-trivial condition, called *strong identifiability* for the expert functions h_1 and h_2.

Definition 1 (Strong Identifiability)**.** *We say that the expert functions $x \mapsto h_1(x,\kappa)$ and $x \mapsto h_2(x,\eta)$ are strongly identifiable if they are twice differentiable w.r.t their parameters κ and η, and if for any $k_1, k_2 \geq 1$, $\kappa_1,\ldots,\kappa_{k_1}$ and $\eta_1,\ldots,\eta_{k_2}$, each of the sets*

$$\left\{\frac{\partial h_1}{\partial \kappa^{(u_1)}}(x,\kappa_i) : i \in [k_1],\ u_1 \in [d_1]\right\}, \left\{\frac{\partial h_1}{\partial \kappa^{(u_1)}}(x,\kappa_i)\frac{\partial h_1}{\partial \kappa^{(v_1)}}(x,\kappa_i), 1 : i \in [k_1],\ u_1, v_1 \in [d_1]\right\},$$
$$\left\{\frac{\partial h_2}{\partial \eta^{(u_2)}}(x,\eta_j),\ \frac{\partial^2 h_2}{\partial \eta^{(u_2)}\partial \eta^{(v_2)}}(x,\eta_j),\ x^{(u)}\frac{\partial h_2}{\partial \eta^{(v_2)}}(x,\eta_j) : j \in [k_2],\ u_2, v_2 \in [d_2],\ u \in [d]\right\}$$

consists of linearly independent functions (in x).

Examples. Two-layer FFNs $h_1(x,(\kappa_2,\kappa_1,\kappa_0)) := \kappa_2 \mathrm{ReLU}(\kappa_1^\top x + \kappa_0)$ and $h_2(x,(\eta_2,\eta_1)) := \eta_2 \mathrm{GELU}(\eta_1^\top x)$ are strongly identifiable. The same claim holds when replacing the ReLU function with other activation functions such as sigmoid and tanh. On the other hand, linear experts $h_1(x,(\kappa_1,\kappa_0)) := \kappa_1^\top x + \kappa_0$ and $h_2(x,(\eta_1,\eta_0)) := \eta_1^\top x + \eta_0$ fail to satisfy the strong identifiability condition because $\frac{\partial h_1}{\partial \kappa_0}\frac{\partial h_1}{\partial \kappa_0} = 1$ and $\frac{\partial h_2}{\partial \eta_1} = x\frac{\partial h_2}{\partial \eta_0}$ for all x.

2.1 Strongly Identifiable Experts

Our next task is to construct a loss over pairs of mixing measures (G_1,G_2) and (G_1^*,G_2^*). To this end, let us revisit the concepts of Voronoi cells and Voronoi loss function presented in [49].

Voronoi loss. For any pair of mixing measures (G_1,G_2) with $k_1' \leq k_1$ and $k_2' \leq k_2$ atoms, we distribute their atoms to the Voronoi cells $\mathcal{V}_{1,j_1} \equiv \mathcal{V}_{1,j_1}(G)$ and $\mathcal{V}_{2,j_2} \equiv \mathcal{V}_{2,j_2}(G)$, defined as

$$\begin{aligned}\mathcal{V}_{1,j_1} &:= \{i_1 \in [k_1'] : \|\xi_{i_1} - \xi_{j_1}^*\| \leq \|\xi_{i_1} - \xi_{\ell_1}^*\|,\ \forall \ell_1 \neq j_1\},\\ \mathcal{V}_{2,j_2} &:= \{i_2 \in [k_2'] : \|\zeta_{i_2} - \zeta_{j_2}^*\| \leq \|\zeta_{i_2} - \zeta_{\ell_2}^*\|,\ \forall \ell_2 \neq j_2\},\end{aligned} \tag{3}$$

where we denote $\xi_{i_1} := (\kappa_{i_1}, \tau_{i_1})$, $\xi^*_{j_1} := (\kappa^*_{j_1}, \tau^*_{j_1})$ for all $j_1 \in [k^*_1]$, and $\zeta_{i_2} := (\beta_{1i_2}, \eta_{i_2}, \nu_{i_2})$, $\zeta^*_{j_2} := (\beta^*_{1j_2}, \eta^*_{j_2}, \nu^*_{j_2})$ for all $j_2 \in [k^*_2]$. Then, the proposed Voronoi loss over mixing measures is

$$\begin{aligned}
\mathcal{D}_1((G_1, G_2), (G^*_1, G^*_2)) &:= \sum_{j=1}^{k^*_1} \Big| \sum_{i \in \mathcal{V}_{1,j}} \omega_i - \omega^*_j \Big| + \sum_{j=1}^{k^*_2} \Big| \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}) - \exp(\beta^*_{0j}) \Big| \\
&+ \sum_{\substack{j \in [k^*_1], \\ |\mathcal{V}_{1,j}|=1}} \sum_{i \in \mathcal{V}_{1,j}} \omega_i (\|\Delta\kappa_{ij}\| + |\Delta\tau_{ij}|) + \sum_{\substack{j \in [k^*_2], \\ |\mathcal{V}_{2,j}|=1}} \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}) (\|\Delta\beta_{1ij}\| + \|\Delta\eta_{ij}\| + |\Delta\nu_{ij}|) \\
&+ \sum_{\substack{j \in [k^*_1], \\ |\mathcal{V}_{1,j}|>1}} \omega_i (\|\Delta\kappa_{ij}\|^2 + |\Delta\tau_{ij}|^2) + \sum_{\substack{j \in [k^*_2], \\ |\mathcal{V}_{2,j}|>1}} \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}) (\|\Delta\beta_{1ij}\|^2 + \|\Delta\eta_{ij}\|^2 + |\Delta\nu_{ij}|^2), \qquad (4)
\end{aligned}$$

where we let $\Delta\kappa_{ij} := \kappa_i - \kappa^*_j$, $\Delta\tau_{ij} := \tau_i - \tau^*_j$, $\Delta\beta_{1ij} := \beta_{1i} - \beta^*_{1j}$, $\Delta\eta_{ij} := \eta_i - \eta^*_j$, and $\Delta\nu_{ij} := \nu_i - \nu^*_j$. It is clear that convergence of the mixing measures in the $\mathcal{D}_1$ loss is equivalent to convergence of their respective parameters. Thus, though not a metric over mixing measures, the $\mathcal{D}_1$ loss can be used to characterize parameter and expert estimation rates.

Theorem 1. *Assume that the expert functions h_1 and h_2 are strongly identifiable. Then, the lower bound $\mathbb{E}_X[V(f_{G_1,G_2}(\cdot|X), f_{G^*_1,G^*_2}(\cdot|X))] \gtrsim \mathcal{D}_1((G_1, G_2), (G^*_1, G^*_2))$ holds for all $(G_1, G_2) \in \mathcal{G}_{k_1,k_2}(\Theta)$. As a consequence,*

$$\mathcal{D}_1((\widehat{G}^n_1, \widehat{G}^n_2), (G^*_1, G^*_2)) = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}}). \qquad (5)$$

The combination of Theorem 1, whose proof is in Appendix D.1, and of the form of the loss $\mathcal{D}_1$ leads to various estimation rates. Below we say that a parameter is *exactly-specified* or *over-specified* depending on whether the associated Voronoi cell has one or more elements, respectively.

(i) Shared experts. For shared experts, we see that the estimation rate for exactly-specified parameters κ^*_j, τ^*_j, is nearly parametric, i.e. of order $\widetilde{\mathcal{O}}_P(n^{-1/2})$. On the other hand, over-specified parameters κ^*_j, τ^*_j, admit slightly slower estimation rates, of order $\widetilde{\mathcal{O}}_P(n^{-1/4})$. As for the expert estimation rates, since the shared expert function $h_1(\cdot, \kappa)$ is Lipschitz continuous, we have that $|h_1(x, \hat{\kappa}^n_i) - h_1(x, \kappa^*_j)| \lesssim \|\hat{\kappa}^n_i - \kappa^*_j\|$ for almost every x.It then follows that the estimation rates for exactly-specified and over-specified shared experts $h_1(x, \kappa^*_j)$ are also of order $\widetilde{\mathcal{O}}_P(n^{-1/2})$ and $\widetilde{\mathcal{O}}_P(n^{-1/4})$, respectively. Thus, polynomially many data points $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-4})$ are needed to estimate these experts within a error $\epsilon > 0$.

(ii) Routed experts. Likewise, exactly-specified and over-specified parameters β^*_{1j}, η^*_j, ν^*_j, for $j \in [k^*_2]$, have estimation rates of order $\widetilde{\mathcal{O}}_P(n^{-1/2})$ and $\widetilde{\mathcal{O}}_P(n^{-1/4})$, respectively. As the routed expert function $h_2(\cdot, \eta)$ is Lipschitz continuous, we deduce that the rates for estimating routed experts $h_2(x, \eta^*_j)$ also vary between $\widetilde{\mathcal{O}}_P(n^{-1/2})$ and $\widetilde{\mathcal{O}}_P(n^{-1/4})$ depending on the cardinality of the corresponding Voronoi cell $\mathcal{V}_{2,j}$. In summary, when both shared and routed expert functions are strongly identifiable, they enjoy the same estimation rates.

2.2 Linear Experts

In this section, we consider linear expert functions of the form $h_1(X, (\kappa_1, \kappa_0)) := \kappa_1^\top X + \kappa_0$ and $h_2(X, (\eta_1, \eta_0)) := \eta_1^\top X + \eta_0$. Then, the pair of ground-truth mixing measures (G^*_1, G^*_2) become

$G_1^* := \sum_{i=1}^{k_1^*} \omega_i^* \delta_{(\kappa_{1i}^*, \kappa_{0i}^*, \tau_i^*)}$ and $G_2^* := \sum_{i=1}^{k_2^*} \exp(\beta_{0i}^*) \delta_{(\beta_{1i}^*, \eta_{1i}^*, \eta_{0i}^*, \nu_i^*)}$. As discussed above, linear experts violate the strong identifiability condition due to the PDEs $\frac{\partial h_1}{\partial \kappa_0} \frac{\partial h_1}{\partial \kappa_0} = 1$ and $\frac{\partial h_2}{\partial \eta_1} = x \frac{\partial h_2}{\partial \eta_0}$. In turn, these PDEs lead to linear dependencies among the partial derivatives of the Gaussian p.d.f. π and of the function F defined below Proposition 1, given by $\frac{\partial^2 \pi}{\partial \kappa_0^2} = 2 \frac{\partial \pi}{\partial \tau}$ and $\frac{\partial F}{\partial \eta_1} = \frac{\partial^2 F}{\partial \beta_1 \partial \eta_0}$. These delicate relationships, which can be intuitively interpreted as interactions between the parameters κ_0 and τ, and among the parameters η_1, β_1 and η_0, affect the parameter and expert estimation rates. To overcome this issue, we consider instead a new Voronoi loss, given by

$$\begin{aligned}
\mathcal{D}_2((G_1, G_2), (G_1^*, G_2^*)) &:= \sum_{j=1}^{k_1^*} \Big| \sum_{i \in \mathcal{V}_{1,j}} \omega_i - \omega_j^* \Big| + \sum_{j=1}^{k_2^*} \Big| \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}) - \exp(\beta_{0j}^*) \Big| \\
&+ \sum_{\substack{j \in [k_1^*], \\ |\mathcal{V}_{1,j}| = 1}} \sum_{i \in \mathcal{V}_{1,j}} \omega_i (\|\Delta \kappa_{1ij}\| + |\Delta \kappa_{0ij}| + |\Delta \tau_{ij}|) + \sum_{\substack{j \in [k_1^*], \\ |\mathcal{V}_{1,j}| > 1}} \sum_{i \in \mathcal{V}_{1,j}} \omega_i (\|\Delta \kappa_{ij}\|^2 + |\Delta \kappa_{0ij}|^{r_{1,j}} + |\Delta \tau_{ij}|^{r_{1,j}/2}) \\
&+ \sum_{j \in [k_2^*] : |\mathcal{V}_{2,j}| = 1} \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}) (\|\Delta \beta_{1ij}\| + \|\Delta \eta_{1ij}\| + |\Delta \eta_{0ij}| + |\Delta \nu_{ij}|) \\
&+ \sum_{j \in [k_2^*] : |\mathcal{V}_{2,j}| > 1} \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}) (\|\Delta \beta_{1ij}\|^{r_{2,j}} + \|\Delta \eta_{1ij}\|^{r_{2,j}/2} + |\Delta \eta_{0ij}|^{r_{2,j}} + |\Delta \nu_{ij}|^{r_{2,j}/2}),
\end{aligned} \tag{6}$$

where we denote $\Delta \kappa_{1ij} := \kappa_{1i} - \kappa_{1j}^*$, $\Delta \kappa_{0ij} := \kappa_{0i} - \kappa_{0j}^*$, $\Delta \eta_{1ij} := \eta_{1i} - \eta_{1j}^*$ and $\Delta \eta_{0ij} := \eta_{0i} - \eta_{0j}^*$. In addition, we define $r_{1,j} := r_1(|\mathcal{V}_{1,j}|)$ and $r_{2,j} := r_2(|\mathcal{V}_{2,j}|)$, where the functions r_1 and r_2 stand for the solvability of polynomial equation systems specified in Appendix B. In particular, we have $r_1(2) = r_2(2) = 4$, $r_1(3) = r_2(3) = 6$, and $r_1(m), r_2(m) \geq 7$ for all $m \geq 4$.

Theorem 2. *Assume the expert functions h_1 and h_2 take linear forms. Then, the lower bound $\mathbb{E}_X[V(f_{G_1, G_2}(\cdot|X), f_{G_1^*, G_2^*}(\cdot|X))] \gtrsim \mathcal{D}_2((G_1, G_2), (G_1^*, G_2^*))$ holds for any $(G_1, G_2) \in \mathcal{G}_{k_1, k_2}(\Theta)$. As a consequence, we have*

$$\mathcal{D}_2(\widehat{G}_1^n, \widehat{G}_2^n), (G_1^*, G_2^*)) = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}}). \tag{7}$$

By comparing the Voronoi losses $\mathcal{D}_1$ and $\mathcal{D}_2$, we see that the estimation rates for exactly-specified shared and routed experts remain of parametric order $\widetilde{\mathcal{O}}_P(n^{-1/2})$. By contrast, there are changes in the estimation rates for the over-specified experts.

(i) Shared experts. The estimation rates for over-specified parameters κ_{1j}^*, κ_{0j}^*, τ_j^* are heterogeneous, of orders $\widetilde{\mathcal{O}}_P(n^{-1/4})$, $\widetilde{\mathcal{O}}_P(n^{-1/2r_{1,j}})$, $\widetilde{\mathcal{O}}_P(n^{-1/r_{1,j}})$, respectively. Since the input space is bounded, we have $|(\hat{\kappa}_{1i}^n)^\top x + \hat{\kappa}_{0i}^n - (\kappa_{1j}^*)^\top x - \kappa_{0j}^*| \lesssim \|\hat{\kappa}_{1i}^n - \kappa_{1j}^*\| + |\hat{\kappa}_{0i}^n - \kappa_{0j}^*|$. Then, it follows that the shared experts $(\kappa_{1j}^*)^\top x + \kappa_{0j}^*$ admit estimation rates of orders $\widetilde{\mathcal{O}}_P(n^{-1/2r_{1,j}})$. However, note that the rates for estimating their input-dependent terms $(\kappa_{1j}^*)^\top x$ are much faster, of order $\widetilde{\mathcal{O}}_P(n^{-1/4})$.

(ii) Routed experts. The estimation rates for over-specified parameters η_{1j}^*, ν_j^* are of orders $\widetilde{\mathcal{O}}_P(n^{-1/r_{2,j}})$, while those for β_{1j}^*, η_{0j}^* are slower, of orders $\widetilde{\mathcal{O}}_P(n^{-1/2r_{2,j}})$. By arguing similarly to the case of shared experts, the rates for estimating the routed experts $(\eta_{1j}^*)^\top x + \eta_{0j}^*$ and their input-dependent terms $(\eta_{1j}^*)^\top x$ depend on the parameter r_2 (related to the solvability of a certain system of polynomial equations) and are of orders $\widetilde{\mathcal{O}}_P(n^{-1/2r_{2,j}})$ and $\widetilde{\mathcal{O}}_P(n^{-1/r_{2,j}})$, respectively.

Table 2: Performance comparisons of different Sparse Mixture of Experts (SMoE) models on subsets of the SlimPajama dataset using a small-scale model with 158M parameters and large-scale model with 679M parameters. (SMoE-SG refers to SMoE Sigmoid Gating). PPL indicates the perplexity score.

	Small Models (158M)				**Large Models (679M)**			
	SMoE	**DeepSeek-V3**	**DeepSeek-V2**	**SMoE-SG**	**SMoE**	**DeepSeek-V3**	**DeepSeek-V2**	**SMoE-SG**
PPL ↓	13.63	**13.42**	13.49	13.61	9.51	9.49	9.52	**9.46**
LAMBADA	25.27%	25.49%	25.29%	25.43%	37.13%	36.88%	37.11%	37.56%
BLiMP	77.71%	77.20%	77.37%	77.38%	80.47%	81.28%	80.98%	81.08%
CBT	84.18%	84.40%	84.33%	84.23%	89.83%	89.65%	89.93%	89.57%
HellaSwag	29.43%	29.38%	29.38%	29.13%	37.49%	37.32%	37.14%	37.52%
PIQA	57.94%	59.14%	60.17%	58.92%	64.36%	65.72%	64.36%	64.91%
ARC-Challenge	21.20%	21.63%	20.52%	21.37%	23.09%	23.95%	24.21%	23.09%
RACE	30.11%	30.60%	31.02%	31.05%	33.03%	33.12%	33.17%	32.68%
SIQA	35.62%	35.57%	34.90%	34.90%	37.41%	38.59%	36.95%	37.67%
CommonSenseQA	24.65%	25.47%	24.98%	24.90%	26.54%	28.09%	27.35%	28.50%
Average	42.90%	**43.21%**	43.11%	43.04%	47.71%	**48.29%**	47.91%	48.06%

Notably, these rates become increasingly slow with the cardinality of the corresponding Voronoi cell $\mathcal{V}_{2,j}$. In particular, when $|\mathcal{V}_{2,j}| = 3$, they become $\widetilde{\mathcal{O}}_P(n^{-1/12})$ and $\widetilde{\mathcal{O}}_P(n^{-1/6})$, respectively.

(iii) Sample efficiency of the shared expert strategy. From the above observations, we see that shared experts have faster estimation rates than routed experts, i.e.$\widetilde{\mathcal{O}}_P(n^{-1/4})$ compared to $\widetilde{\mathcal{O}}_P(n^{-1/r_{2,j}})$. Furthermore, the estimation rates for shared experts in DeepSeekMoE are also faster than those for experts in MoE models without the shared expert strategy [56], which are also of the order $\widetilde{\mathcal{O}}_P(n^{-1/r_{2,j}})$. The punchline is that fewer data points are needed to estimate shared experts.

3 Experiments

In this section, we empirically validate the theoretical findings in the previous section. Using synthetic data, we demonstrate the convergence behavior of the maximum likelihood estimator $(\widehat{G}_1^n, \widehat{G}_2^n)$ towards the true mixing measure (G_1^*, G_2^*); we defer this experiment to Appendix G.1. In real-world scenarios, we evaluate our methodology on language modeling tasks using the SlimPajama corpus [66] (Section 3.1), and extend our evaluation to vision-language modeling benchmarks using the LLaVA architecture [43] integrated within the LIBMoE framework [57] (Section 3.2). Our empirical study compares four model configurations: Vanilla SMoE, DeepSeek-V3 (shared experts combined with normalized sigmoid gating), DeepSeek-V2 (shared experts with softmax routing), and SMoE Sigmoid Gating (normalized sigmoid gating without shared experts).

3.1 Language Modeling

Experimental Setup. We conduct the experiments on language modeling using subsets of the popular SLimPajama [66] dataset using Switch Transformer [20] baseline in two scales: small (158M parameters trained on 6.5B tokens) and large (679M parameters trained on 26.2B tokens). The models are configured with 66 total experts, utilizing top-8 expert routing in the baseline and a top-6 plus 2 shared experts routing scheme in the DeepSeek variants. We measure model performance in terms of perplexity and zero-shot accuracy across nine diverse downstream evaluation tasks [58, 74, 24, 81, 3, 11, 35, 63, 68]. Full experimental details are provided in Appendix I.1.

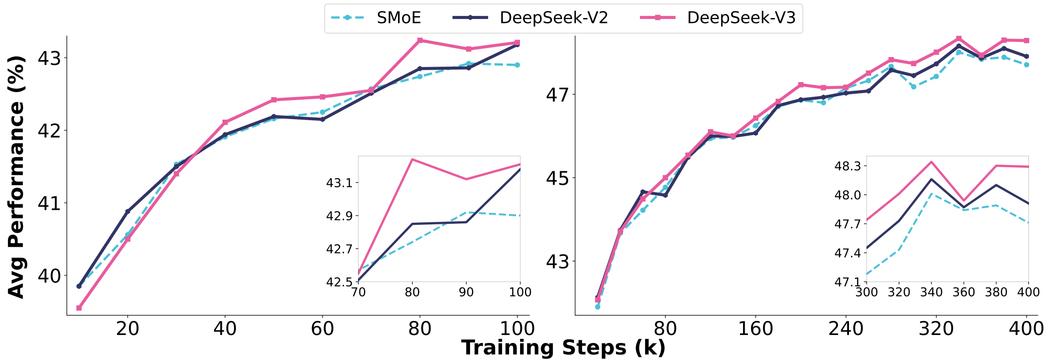


Figure 1: Average performance (%) over training steps in language modeling tasks. **Left:** Model with 158M parameters; **Right:** Model with 679M parameters.

Zero-shot performance on downstream tasks. Table 2 summarizes our primary experimental results for two model sizes trained on the SlimPajama dataset [66]. The results clearly demonstrate that both DeepSeek-V3 and DeepSeek-V2 consistently outperform the Vanilla SMoE baseline, achieving lower perplexity (PPL) scores and higher average accuracy across various downstream tasks for both model scales. Additionally, we integrated the normalized sigmoid router into the Vanilla SMoE architecture and observed that the SMoE Sigmoid Gating achieves superior performance compared to the Vanilla SMoE and, in some benchmarks, even surpasses the DeepSeek variants.

Convergence Rate. Figure 1 presents the average performance across various downstream tasks for DeepSeek-V3 and DeepSeek-V2 compared to the Vanilla SMoE. Across both model sizes, the DeepSeek variants demonstrate substantially faster convergence. Specifically, in both 158M and 679M parameter scales, DeepSeek-V3 and DeepSeek-V2 consistently reach the final performance of Vanilla SMoE using only 70-80% of the total training steps. Notably, DeepSeek-V3, which incorporates normalized sigmoid gating, demonstrates marginal improvements over DeepSeek-V2 in both convergence speed and final task performance. These results highlight the efficiency gains introduced by the shared expert and normalized sigmoid gating mechanisms and provide empirical support for our theoretical findings.

3.2 Vision-Language Modeling

Experimental Setup. We conduct experiments on the visual instruction tuning tasks [42] using the popular LLaVA architecture [44]. Building upon the LIBMoE framework [57], we adopt Phi3.5-mini [1] as the language model and SigLIP [82] as the vision encoder. Unlike LIBMoE, we sparse-upcycled [32] only the MLP Connector into 8 experts, employing a top-4 expert routing strategy, while the DeepSeek variants adopt a top-3 expert routing scheme with an additional shared expert, making our model approximately 4.4B parameters. To compare different SMoE algorithms, we use a subset of the LLaVA 1.5 dataset [42] (332K samples and 287M tokens) to train the models in the Visual Instruction Tuning (VIT) stage. Evaluation covers diverse benchmarks containing various vision-language capabilities, including perception, reasoning, OCR, instruction following, and more [31, 7, 40, 47, 65, 27, 83, 78, 45]. See Appendix I.2.

Performance. As summarized in Table 3, DeepSeek-V3 achieves the highest average score (51.75%),

Table 3: Vision-language model performance across benchmarks. (SMoE-SG refers to SMoE Sigmoid Gating)

	AI2D	MMStar	POPE	Science QA	TextVQA	GQA	MME-RW -Lite	MMMU Pro-S	OCR Bench	Average
SMoE	64.90%	41.66%	85.67%	81.61%	40.92%	60.19%	31.79%	25.61%	30.90%	51.47%
DeepSeek-V3	65.45%	41.40%	85.44%	81.94%	40.69%	60.01%	32.20%	26.01%	32.60%	**51.75%**
DeepSeek-V2	64.70%	41.55%	85.80%	82.20%	40.51%	60.15%	31.11%	25.72%	31.00%	51.41%
SMoE-SR	64.64%	41.51%	85.87%	82.17%	40.54%	60.07%	31.68%	25.95%	31.00%	51.49%

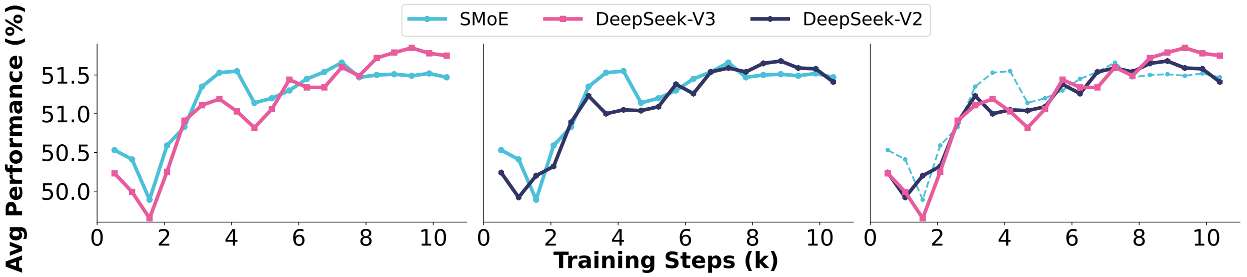


Figure 2: Average performance (%) over training steps on vision-language pretraining tasks. **Left:** Vanilla SMoE vs. DeepSeek-V3; **Center:** Vanilla SMoE vs. DeepSeek-V2; **Right:** DeepSeek-V2 vs. DeepSeek-V3.

outperforming the Vanilla SMoE (51.47%) and other model variants. Although DeepSeek-V2 shows slightly lower performance compared to other models, the difference remains marginal. Consistent with observations from language modeling experiments, additional evaluations conducted with Vanilla SMoE and the normalized sigmoid router show a similar pattern, confirming that the normalized sigmoid routing mechanism consistently enhances the performance of the standard SMoE architecture.

Convergence Rate. Figure 2 illustrates the performance progression over training steps, where both DeepSeek variants exhibit faster and more stable convergence compared to Vanilla SMoE. Notably, both DeepSeek-V2 and DeepSeek-V3 demonstrate accelerated convergence during the final stages of training. These results suggest that both shared expert integration and normalized routing significantly contribute to faster learning in vision-language pretraining.

3.3 Router Analysis

We now explore the router behavior by empirically examining the router saturation and change rate.

Router Saturation. Router Saturation, first introduced in OLMoE [51], quantifies the proportion of overlapping activated experts between the final checkpoint and an intermediary checkpoint at time t. It serves as a measure of the router's convergence over the course of training. A higher router saturation value indicates stronger alignment in expert selection, signifying that the router's decisions become increasingly consistent with its final configuration. The formal definition and formula are defined in the Appendix H.1. Figure 3 shows that, after 5% of training, up to ~60% of router decisions have already saturated. This early saturation aligns with prior findings in OLMoE [51] and OpenMoE [76], supporting the validity of our experimental setup. When comparing model configurations, we observe that models equipped with normalized sigmoid gating achieve noticeably faster saturation than those using softmax gating. In particular, the SMoE Sigmoid Gating exhibits consistently steeper saturation curves compared to Vanilla SMoE, reflecting more rapid convergence in expert selection. A similar pattern is observed in the comparison between DeepSeek-V3 and DeepSeek-V2 under the

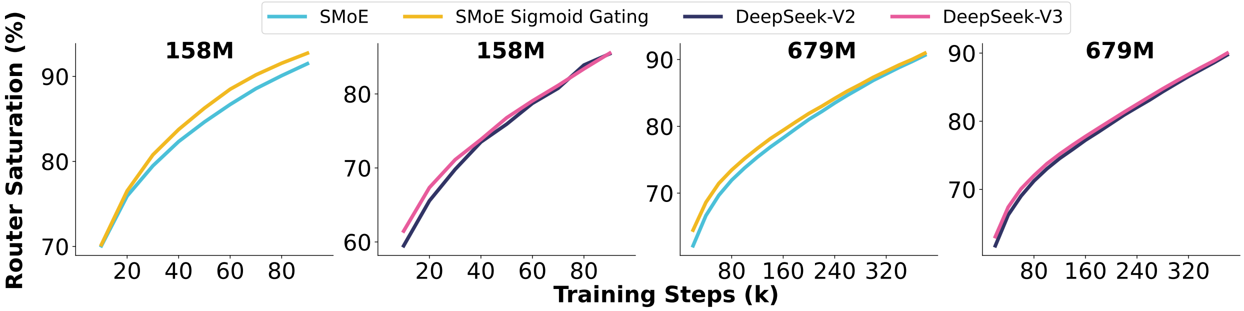


Figure 3: Evolution of router saturation (averaged across all layers) during training for language-modeling tasks with 158 M (left) and 679 M (right) parameter models. We compute saturation by comparing the routing to the top-8 experts with SMoE and SMoE Sigmoid Gating, and the top-6 experts with DeepSeek variants.

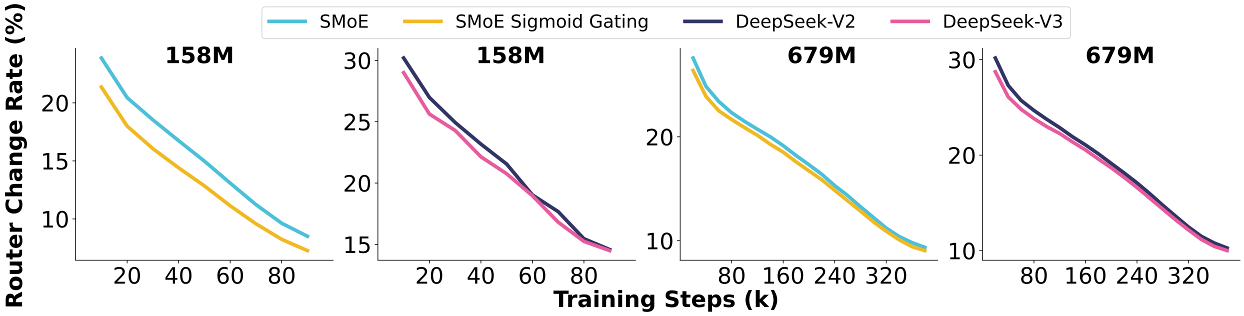


Figure 4: Router Change Rate (averaged across all layers) during training for language-modeling tasks with 158 M (left) and 679 M (right) parameter models. We compute router change rate by comparing the routing to the top-8 experts with SMoE and SMoE Sigmoid Gating, and the top-6 experts with DeepSeek variants.

shared expert configuration. These findings highlight the effectiveness of normalized sigmoid gating in accelerating router convergence, potentially reducing the training time required for convergence.

Router Change Rate. To evaluate the stability of the routing mechanism in Mixture-of-Experts (MoE) models during training, we introduce the Router Change Rate metric. This metric quantifies the proportion of expert activation decisions that change between consecutive checkpoints. A lower router change rate implies greater consistency in routing decisions over time, reflecting a more stable training process. The formal definition and computation details are provided in Appendix H.2. Figure 4 presents the router change rate comparison of different model configurations. We find that models employing normalized sigmoid gating have significantly lower change rates in both non-shared and shared expert settings. These findings underscore the efficiency of normalized sigmoid gating in stabilizing routing decisions throughout training. By reducing the routing fluctuation problem [13], this mechanism promotes a more consistent expert specialization, indicating that stable routing is critical in enhancing both optimization efficiency and final model performance.

4 Discussion

In this paper, we have presented an extensive study on the benefits of two fundamental ingredients of DeepSeekMoE architecture, namely the shared expert strategy and the normalized sigmoid gating mechanism. From the theoretical side, we perform a convergence analysis of expert estimation to investigate differences in sample efficiency. Our analysis reveals that the shared expert strategy leads to faster estimation rates for shared experts compared to routed experts and experts in the standard MoE. Furthermore, the estimation rates for routed experts become dramatically faster when replacing the softmax gating with the normalized sigmoid gating in DeepSeekMoE. Therefore, the incorporation of these two key factors into DeepSeekMoE significantly reduces the overall sample complexity for the estimation tasks.

From the empirical side, we validate our theoretical findings through extensive experiments and analysis on both synthetic and real-world datasets. Our results consistently demonstrate that both the shared experts strategy and the normalized sigmoid gating mechanism substantially affect the convergence rate and downstream performance in real-world scenarios. Moreover, these two ingredients also yield substantial gains in router convergence, routing stability, and expert utilization. Overall, our work provides both a principled understanding and robust empirical evidence for the effectiveness of these two components, offering valuable guidance for the design of future sparse mixture-of-experts.

Although our analysis confirms that using shared experts improves the sample complexity, it does not indicate how many shared experts should be employed to achieve the optimal configuration given a fixed computational budget. A potential approach to this problem is to derive a scaling law involving these quantities induced from extensive experiments as in [48]. However, since this direction goes beyond the scope of our work, we leave it for future development.

Appendices for "On DeepSeekMoE: Statistical Benefits of Shared Experts and Normalized Sigmoid Gating"

Contents

A On Normalized Sigmoid Gating

In this appendix, we conduct a convergence analysis of expert estimation in DeepSeek-V3's MoE to investigate the sample efficiency of the normalized sigmoid gating used in this architecture.

Problem setting. Assume that $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ are i.i.d. samples drawn from the Gaussian DeepSeek-V3's MOE whose conditional density function $g_{G_*}(y|x)$ is given by:

$$\begin{aligned} g_{G_1^*, G_2^*}(y|x) := \frac{1}{2} \sum_{i=1}^{k_1^*} \omega_i^* \pi(y|h_1(x, \kappa_i^*), \tau_i^*) \\ + \frac{1}{2} \sum_{i=1}^{k_2^*} \frac{\sigma((\beta_{1i}^*)^\top x + \beta_{0i}^*)}{\sum_{j=1}^{k_2^*} \sigma((\beta_{1j}^*)^\top x + \beta_{0j}^*)} \cdot \pi(y|h_2(x, \eta_i^*), \nu_i^*), \end{aligned} \tag{8}$$

where $\sigma : \mathbb{R} \to (0, \infty)$ stands for the sigmoid function, that is, $\sigma(z) := \frac{1}{1+\exp(-z)}$, for all $z \in \mathbb{R}$. By abuse of notations, we define the pair of ground-truth mixing measures (G_1^*, G_2^*) under this setting as $G_1^* := \sum_{i=1}^{k_1^*} \omega_i^* \delta_{(\kappa_i^*, \tau_i^*)}$ and $G_2^* := \sum_{i=1}^{k_2^*} \sigma(\beta_{0i}^*) \delta_{(\beta_{1i}^*, \eta_i^*, \nu_i^*)}$. Here, we still impose all the assumptions used for Section 2 on this analysis.

Maximum likelihood estimation (MLE). Under the above setting, the MLE defined in equation (2) is rewritten as

$$(\widetilde{G}_1^n, \widetilde{G}_2^n) \in \underset{(G_1, G_2) \in \mathcal{G}_{k_1, k_2}(\Theta)}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} \log(g_{G_1, G_2}(Y_i|X_i)), \tag{9}$$

where $\mathcal{G}_{k_1,k_2}(\Theta) := \mathcal{G}_{k_1}(\Theta_1) \times \mathcal{G}_{k_2}(\Theta)$ denotes the set of mixing measure pairs (G_1, G_2) with at most k_1 and k_2 atoms, respectively, that is,

$$\begin{aligned} \mathcal{G}_{k_1}(\Theta_1) &:= \Big\{ G_1 = \sum_{i=1}^{k_1'} \omega_i \delta_{(\kappa_i, \tau_i)} : 1 \leq k_1' \leq k_1 \Big\}, \\ \mathcal{G}_{k_2}(\Theta_2) &:= \Big\{ G_2 = \sum_{i=1}^{k_2'} \sigma(\beta_{0i}) \delta_{(\beta_{1i}, \eta_i^*, \nu_i^*)} : 1 \leq k_2' \leq k_2 \Big\}. \end{aligned}$$

Given the MLE $(\widetilde{G}_1^n, \widetilde{G}_2^n)$ in equation (9), we proceed to establish the convergence rate of density estimation $g_{\widetilde{G}_1^n, \widetilde{G}_2^n}$. However, there are some changes in the gating convergence behavior compared to that in DeepSeekMoE due to the structure of the sigmoid function.

The convergence of normalized sigmoid gating. Recall that we fit the ground-truth DeepSeek-V3's MoE model (8) with a mixture of $k_1 > k_1^*$ shared experts and $k_2 > k_2^*$ routed experts. Then, there must be some gorund-truth routed experts approximated by more than one fitted routed experts. As a result, the sum of weights of these fitted routed experts is expected to converge to the weight of the ground-truth routed experts, for example,

$$\sum_{i=1}^{2} \frac{\sigma((\hat{\beta}_{1i}^n)^\top x + \hat{\beta}_{0i}^n)}{\sum_{j=1}^{k_2^n} \sigma((\hat{\beta}_{1j}^n)^\top x + \hat{\beta}_{0j}^n)} \to \frac{\sigma((\beta_{11}^*)^\top x + \beta_{01}^*)}{\sum_{j=1}^{k_2^*} \sigma((\beta_{1j}^*)^\top x + \beta_{0j}^*)},$$

for almost every x. Since the denominator $\sum_{j=1}^{k_2^n}\sigma((\hat{\beta}_{1j}^n)^\top x+\hat{\beta}_{0j}^n)$ should converge to its counterpart $\sum_{j=1}^{k_2^*}\sigma((\beta_{1j}^*)^\top x+\beta_{0j}^*)$. Then, it must hold that

$$\sum_{i=1}^{2}\sigma((\hat{\beta}_{1i}^n)^\top x+\hat{\beta}_{0i}^n)\to\sigma((\beta_{11}^*)^\top x+\beta_{01}^*),$$

as $n\to\infty$, for almost every x. This result occurs only if $\beta_{11}^*=0_d$. Therefore, we will divide our analysis into two complement regimes for the over-specified parameters β_{1i}^*:

Sparse regime. All over-specified parameters β_{1i}^* equal zero vector;

Dense regime. Not all over-specified parameters β_{1i}^* equal zero vector.

It is worth noting that the sparse regime of parameters rarely occurs in practice. However, for completeness, we will perform the convergence analysis of expert estimation under both the sparse and dense regimes in Appendix A.1 and Appendix A.2, respectively.

A.1 Sparse Regime

To begin with, let us derive the density estimation rate for the sparse regime in Proposition 2.

Proposition 2. *Under the sparse regime, the density estimation $g_{\widetilde{G}_1^n,\widetilde{G}_2^n}(Y|X)$ converges to the true density $g_{G_1^*,G_2^*}(Y|X)$ at the following rate:*

$$\mathbb{E}_X[V(g_{\widetilde{G}_1^n,\widetilde{G}_2^n}(\cdot|X),g_{G_1^*,G_2^*}(\cdot|X))]=\mathcal{O}_P([\log(n)/n]^{\frac{1}{2}}).$$

Since the sigmoid function is Lipschitz continuous, the proof of this proposition can be done similarly to that of Proposition 1, which is provided in Appendix E.1. The result of Proposition 2 indicates that the density estimation $g_{\widetilde{G}_1^n,\widetilde{G}_2^n}$ converges to the ground-truth density $g_{G_1^*,G_2^*}$ under the Total Variation distance at the parametric rate of order $\widetilde{\mathcal{O}}_P(n^{-1/2})$.

Voronoi loss. Next, we construct Voronoi loss tailored to the sparse regime as

$$\begin{aligned}\mathcal{D}_3((G_1,G_2),(G_1^*,G_2^*)):=&\sum_{j=1}^{k_1^*}\Big|\sum_{i\in\mathcal{V}_{1,j}}\omega_i-\omega_j^*\Big|+\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}\Big|\sum_{i\in\mathcal{V}_{2,j}}\sigma(\beta_{0i})-\sigma(\beta_{0j}^*)\Big|\\&+\sum_{\substack{j\in[k_1^*],\\|\mathcal{V}_{1,j}|=1}}\sum_{i\in\mathcal{V}_{1,j}}\omega_i(\|\Delta\kappa_{ij}\|+|\Delta\tau_{ij}|)+\sum_{\substack{j\in[k_2^*],\\|\mathcal{V}_{2,j}|=1}}\sum_{i\in\mathcal{V}_{2,j}}(\|\Delta\beta_{1ij}\|+|\Delta\beta_{0ij}|+\|\Delta\eta_{ij}\|+|\Delta\nu_{ij}|)\\&+\sum_{\substack{j\in[k_1^*],\\|\mathcal{V}_{1,j}|>1}}\sum_{i\in\mathcal{V}_{1,j}}\omega_i(\|\Delta\kappa_{ij}\|^2+|\Delta\tau_{ij}|^2)+\sum_{\substack{j\in[k_2^*],\\|\mathcal{V}_{2,j}|>1}}\sum_{i\in\mathcal{V}_{2,j}}(\|\Delta\beta_{1ij}\|^2+\|\Delta\eta_{ij}\|^2+|\Delta\nu_{ij}|^2),\end{aligned}\tag{10}$$

where we denote $\Delta\beta_{0ij}:=\beta_{0i}-\beta_{0j}^*$. Given the above loss function, we are now able to capture parameter and expert estimation rates under the sparse regime in the following theorem.

Theorem 3. *Suppose that the expert functions h_1 and h_2 are strongly identifiable. Then, the lower bound $\mathbb{E}_X[V(g_{G_1,G_2}(\cdot|X), g_{G_1^*,G_2^*}(\cdot|X))] \gtrsim \mathcal{D}_3((G_1, G_2), (G_1^*, G_2^*))$ holds for any $(G_1, G_2) \in \mathcal{G}_{k_1,k_2}(\Theta)$. As a consequence, we have*

$$\mathcal{D}_3(\widetilde{G}_1^n, \widetilde{G}_2^n), (G_1^*, G_2^*)) = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}}).$$

The proof of Theorem 3 is provided in Appendix D.3. From the formulations of Voronoi losses $\mathcal{D}_1$ and $\mathcal{D}_3$ in equations (4) and (10), respectively, we observe that shared experts and routed experts which satisfy the strong identifiability condition admit the same estimation rates as those in Theorem 1. In particular, the rates for estimating both types of experts are of orders $\widetilde{\mathcal{O}}_P(n^{-1/2})$ and $\widetilde{\mathcal{O}}_P(n^{-1/4})$ when they are exactly-specified and over-specified, respectively. In other words, the normalized sigmoid gating does not have clear advantages over the standard softmax gating under the sparse regime. However, it should be noted that the sparse regime is less likely to occur in practice than the dense regime. Thus, we continue the comparison of sample efficiency between the two gatings under the dense regime in the next section.

A.2 Dense Regime

Next, under the dense regime, note that the ground-truth model is misspecified, that is, the density estimation $g_{\widetilde{G}_1^n,\widetilde{G}_2^n}$ converges to the missepcified density function $g_{G_1^*,\check{G}_2}$ rather than the ground-truth density $g_{G_1^*,G_2^*}$, where $\check{G}_2 \in \overline{\mathcal{G}}_{k_2}(\Theta_2) := \arg\min_{G_2 \in \mathcal{G}_{k_2}(\Theta_2)\setminus\mathcal{G}_{k_2^*}(\Theta_2)} \mathrm{KL}(g_{G_1^*,G_2} \| g_{G_1^*,G_2^*})$. Following the result of Proposition 2, we are also able to establish the parametric density estimation rate under the dense regime in the following corollary.

Corollary 1. *Under the dense regime, the density estimation $g_{\widetilde{G}_1^n,\widetilde{G}_2^n}$ converges to the density $g_{G_1^*,\check{G}_2}$ at the rate:* $\inf_{\check{G}_2 \in \overline{\mathcal{G}}_{k_2}(\Theta_2)} \mathbb{E}_X[V(g_{\widetilde{G}_1^n,\widetilde{G}_2^n}(\cdot|X), g_{G_1^*,\check{G}_2}(\cdot|X))] = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}})$.

Subsequently, we focus on characterizing parameter and expert estimation rates under the dense regime by establishing the Total Variation lower bound

$$\inf_{(G_1^*,\check{G}_2) \in \overline{\mathcal{G}}_{k_1,k_2}(\Theta)} \mathbb{E}_X[V(g_{G_1,G_2}(\cdot|X), g_{G_1^*,\check{G}_2}(\cdot|X))] \gtrsim \mathcal{D}_4((G_1, G_2), (G_1^*, \check{G}_2)),$$

where $\mathcal{D}_4$ is a Voronoi loss that will be defined later in equation (11). Recall that a key step in deriving this lower bound is to decompose the density difference $g_{\widetilde{G}_1^n,\widetilde{G}_2^n}(Y|X) - g_{G_1^*,\check{G}_2}(Y|X)$ into linearly independent terms using Taylor expansions to the functions $x \mapsto \pi(Y|h_1(x,\kappa),\tau)$ and $x \mapsto \sigma(\beta_1^\top x + \beta_0)\pi(Y|h_2(x,\eta),\nu)$ w.r.t their parameters (κ,τ) and $(\beta_1,\beta_0,\eta,\nu)$, respectively. Due to the gating change, it is necessary to introduce a new condition on the routed expert function h_2 to ensure linear independence among terms in the Taylor expansions.

Definition 2 (Weak Identifiability)**.** *We say that the expert function $x \mapsto h_2(x,\eta)$ is weakly identifiable if it is differentiable w.r.t its parameter η, and if for any $k_2 \geq 1$ and $\eta_1, \eta_2, \ldots, \eta_{k_2}$, the following set is linearly independent w.r.t x:*

$$\left\{ \frac{\partial h_2}{\partial \eta^{(u_2)}}(x, \eta_i) : i \in [k_2],\ u_2 \in [d_2] \right\}.$$

Examples. It can be validated that even linear experts of the form $h_2(x,(\eta_1,\eta_0)) := \eta_1^\top x + \eta_0$ satisfy the weak identifiability condition. Note that the strong identifiability condition in Definition 1 implies the weak identifiability condition. Therefore, two-layer FFNs $h_2(x,(\eta_2,\eta_1,\eta_0)) := \eta_2 \mathrm{ReLU}(\eta_1^\top x + \eta_0)$ are also weakly identifiable. On the other hand, input-free experts $h_2(x,\eta) = c(\eta)$ does not meet the weak identifiability condition.

Voronoi loss. Now, we build a Voronoi loss to capture parameter estimation rates under the dense regime, which is given by

$$\mathcal{D}_4((G_1,G_2),(G_1^*,\check{G}_2)) := \sum_{j=1}^{k_1^*}\Big|\sum_{i\in\mathcal{V}_{1,j}}\omega_i - \omega_j^*\Big| + \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|=1}\sum_{i\in\mathcal{V}_{1,j}}\omega_i(\|\Delta\kappa_{ij}\| + |\Delta\tau_{ij}|)$$

$$+ \sum_{\substack{j\in[k_1^*],\\ |\mathcal{V}_{1,j}|>1}}\sum_{i\in\mathcal{V}_{1,j}}\omega_i(\|\Delta\kappa_{ij}\|^2 + |\Delta\tau_{ij}|^2) + \sum_{j=1}^{k_2^*}\sum_{i\in\mathcal{V}_{2,j}}(\|\beta_{1i}-\check{\beta}_{1j}\| + |\beta_{0i}-\check{\beta}_{0j}|$$

$$+ \|\eta_i - \check{\eta}_j\| + |\nu_i - \check{\nu}_j|). \tag{11}$$

Given the above loss, we are now ready to present results for the convergence rates of parameter estimation and expert estimation in Theorem 4, whose proof can be found in Appendix D.4.

Theorem 4. *Suppose that the shared expert function h_1 is strongly identifiable, while the routed expert function h_2 is weakly identifiable. Then, the lower bound*

$$\inf_{(G_1^*,\check{G}_2)\in\overline{\mathcal{G}}_{k_1,k_2}(\Theta)} \mathbb{E}_X[V(g_{G_1,G_2}(\cdot|X), g_{G_1^*,\check{G}_2}(\cdot|X))] \gtrsim \mathcal{D}_4((G_1,G_2),(G_1^*,\check{G}_2))$$

holds for any $(G_1,G_2)\in\mathcal{G}_{k_1,k_2}(\Theta)$. As a consequence, we have

$$\inf_{(G_1^*,\check{G}_2)\in\overline{\mathcal{G}}_{k_1,k_2}(\Theta)} \mathcal{D}_4(\widetilde{G}_1^n,\widetilde{G}_2^n),(G_1^*,\check{G}_2)) = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}}).$$

It can be seen from the formulation of the Voronoi loss $\mathcal{D}_4$ that the estimation rates for shared experts remain unchanged compared to those in Theorem 3, which are of the orders $\widetilde{\mathcal{O}}_P(n^{-1/2})$ for exactly-specified ones and $\widetilde{\mathcal{O}}_P(n^{-1/4})$ for over-specified ones. However, there are changes in the estimation rates for routed experts.

(i) Routed experts: In particular, the convergence rates of parameter estimation $\widetilde{\eta}_i^n$ are of parametric order $\widetilde{\mathcal{O}}_P(n^{-1/2})$. Since the routed expert function $h_2(x,\eta)$ is Lipschitz continuous w.r.t its parameter η, then the rates for estimating both exactly-specified and over-specified routed experts are of order $\widetilde{\mathcal{O}}_P(n^{-1/2})$. These rates are substantially faster than those when using the standard softmax gating in Theorem 1 and Theorem 2, which are of orders $\widetilde{\mathcal{O}}_P(n^{-1/4})$ and $\widetilde{\mathcal{O}}_P(n^{-1/r_2(|\mathcal{V}_{2,j}|)})$, respectively.

(ii) Sample efficiency of the normalized sigmoid gating: As a result, when using the normalized sigmoid gating, then we need only $\mathcal{O}(\epsilon^{-2})$ to approximate routed experts with a given error ϵ, even if they are of linear form. On the other hand, when using the softmax gating, it requires $\mathcal{O}(\epsilon^{-4})$ data points to estimate strongly identifiable experts. Furthermore, if the routed experts are of linear form, then we need $\mathcal{O}(\epsilon^{-r_2(|\mathcal{V}_{2,j}|)})$ data points to estimate, which is equivalent to $\mathcal{O}(\epsilon^{-12})$ when these routed experts have three fitted experts, that is, $|\mathcal{V}_{2,j}| = 3$. Hence, we claim that the normalized sigmoid gating helps improve the sample efficiency of DeepSeekMoE.

B Systems of Polynomial Equations

In this appendix, we will provide a formal definition of the functions r_1 and r_2 involved in the Voronoi loss $\mathcal{D}_2$ defined in equation (6).

Definition of the function r_1. To capture estimation rates for shared expert parameters in Section 2.2, it is necessary to consider the solvability of a system of polynomial equations previously studied in [25]. More specifically, for each $m \geq 2$, let $r_1(m)$ be the smallest natural number r such that the system:

$$\sum_{i=1}^{m} \sum_{\substack{n_1, n_2 \in \mathbb{N}: \\ n_1 + 2n_2 = \ell}} \frac{s_{3i}^2 \; s_{1i}^{n_1} \; s_{2i}^{n_2}}{n_1! \; n_2!} = 0, \quad \ell = 1, 2, \ldots, r, \tag{12}$$

does not admit any non-trivial solutions for the unknown variables $\{s_{1i}, s_{2i}, s_{3i}\}_{i=1}^{m}$. Here, we call a solution non-trivial if all the values of s_{3i} are non-zero, whereas at least one among s_{1i} is different from zero. In the following proposition, we provide the values of the function r_1 at some specific points $m \in \mathbb{N}$.

Proposition 3 (Proposition 2.1, [25])**.** *For $m = 2$, we get $r_1(m) = 4$, while for $m = 3$, we have $r_1(m) = 6$. When $m \geq 4$, we have $r_1(m) \geq 7$.*

The proof of Proposition 3 can be found in [25].

Definition of the function r_2. To characterize estimation rates for routed expert parameters in Section 2.2, we need to take into account the solvability of another system of polynomial equations studied in [56], which is given by

$$\sum_{i=1}^{m} \sum_{\alpha \in \mathcal{I}_{\ell_1, \ell_2}} \frac{t_{5i}^2 \; t_{1i}^{\alpha_1} \; t_{2i}^{\alpha_2} \; t_{3i}^{\alpha_3} \; t_{4i}^{\alpha_4}}{\alpha_1! \; \alpha_2! \; \alpha_3! \; \alpha_4!} = 0, \tag{13}$$

for all $\ell_1, \ell_2 \geq 0$ satisfying $1 \leq \ell_1 + \ell_2 \leq r$, where

$$\mathcal{I}_{\ell_1, \ell_2} := \{\alpha = (\alpha_i)_{i=1}^{4} \in \mathbb{N}^d \times \mathbb{N}^d \times \mathbb{N} \times \mathbb{N} : \alpha_1 + \alpha_2 = \ell_1, \alpha_3 + 2\alpha_4 = \ell_2 - |\alpha_2|\}.$$

Then, we define $r_2(m)$ as the smallest natural number r such that the system in equation (13) has no non-trivial solutions for the unknown variables $\{t_{5i}, t_{1i}, t_{2i}, t_{3i}, t_{4i}\}_{i=1}^{m}$. Here, a solution is called non-trivial if all the values of t_{5i} are different from, while at least one among t_{4i} is non-zero. The following proposition provides a relation between the two functions r_1 and r_2 as well as specify the values of $r_2(m)$ at some points $m \in \mathbb{N}$.

Proposition 4 (Lemma 1, [56])**.** *The function r_2 is upper bounded by the function r_1, that is, $r_2(m) \leq r_1(m)$, for all $m \in \mathbb{N}$. In addition, we have $r_2(2) = 4$, $r_2(3) = 6$ and $r(m) \geq 7$ when $m \geq 4$.*

The proof of Lemma 4 can be found in [56].

C Related Works

There have been two primary lines of works on understanding MoE models in the literature.

From a statistical perspective, Zeevi et al. [80] investigated the representation power of a mixture of generalized linear experts when using this model to approximate target functions belonging to a Sobolev class. Next, Mendes et al. [50] performed a convergence analysis of MLE under the MoE with experts being polynomial regression models, offering an important insight for finding the optimal configuration of the number of experts and their sizes. After that, considering data generated from a Gaussian MoE with covariate-free gating, Ho et al. [26] established an *algebraic independence* condition on the location and scale functions of the Gaussian density to characterize which choices of this pair will lead to faster convergence rates of parameter estimation. Then, this analysis was extended to more practical yet challenging settings of dense and sparse softmax gating Gaussian MoE in [56] and [54], respectively. These works demonstrated that parameter and expert estimation rates hinged on the solvability of some systems of polynomial equations and became significantly slow as the number of experts increased. Lastly, Nguyen et al. [55] considered a MoE-based regression framework where the regression function took the form of MoE with standard softmax gating, dense-to-sparse gating, and hierarchical softmax gating, respectively. Their convergence analysis of least squares estimation provided critical implications on the design of expert structures. In particular, it indicated that feed-forward expert networks equipped with the sigmoid function or the Gaussian linear error unit (GELU) activation function admitted estimation rates of polynomial orders, while experts of polynomial forms had much slower estimation rates, of exponential orders.

From a deep learning perspective, Chen et al. [8] took into account a classification problem with cluster structures using MoE models. In particular, they justified the ability of the gating network to learn the cluster-center features, enabling the model to separate a big complex problem into simpler ones, each of which will be handled by the corresponding specialized experts. Furthermore, theories for applications of MoE in continual learning [39, 37], domain adaptation [53, 9], and language modeling [59, 17] have also been extensively explored in the literature. Interestingly, self-attention mechanism in the Transformers architecture [73] has recently been shown to be represented by a mixture of linear experts with quadratic softmax gating [2, 77], leading to numerous advances in parameter-efficient fine-tuning methods [71, 36].

However, to the best of our knowledge, no prior work has been done to identify the theoretical properties of the DeepSeekMoE architecture.

D Proof of Main Results

D.1 Proof of Theorem 1

Proof overview. Recall that our goal is to demonstrate that the following lower bound holds for any $G \in \mathcal{G}_{k_1,k_2}(\Theta)$:

$$\mathbb{E}_X[V(f_{G_1,G_2}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X))] \gtrsim \mathcal{D}_1((G_1, G_2), (G_1^*, G_2^*)). \tag{14}$$

Our proof will be divided into two main parts. Firstly, we aim to establish the local part of the bound (14), that is,

$$\lim_{\varepsilon\to 0} \inf_{(G_1,G_2)\in\mathcal{G}_{k_1,k_2}(\Theta):\mathcal{D}_1((G_1,G_2),(G_1^*,G_2^*))\leq\varepsilon} \frac{\mathbb{E}_X[V(f_{G_1,G_2}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X))]}{\mathcal{D}_1((G_1,G_2),(G_1^*,G_2^*))} > 0. \tag{15}$$

The above result implies that there exists a positive constant ε' such that

$$\inf_{(G_1,G_2)\in\mathcal{G}_{k_1,k_2}(\Theta):\mathcal{D}_1((G_1,G_2),(G_1^*,G_2^*))\leq\varepsilon'} \frac{\mathbb{E}_X[V(f_{G_1,G_2}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X))]}{\mathcal{D}_1((G_1,G_2),(G_1^*,G_2^*))} > 0.$$

Then, we complete the proof by deriving the following global part of the bound (14):

$$\inf_{(G_1,G_2)\in\mathcal{G}_{k_1,k_2}(\Theta):\mathcal{D}_1((G_1,G_2),(G_1^*,G_2^*))>\varepsilon'} \frac{\mathbb{E}_X[V(f_{G_1,G_2}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X))]}{\mathcal{D}_1((G_1,G_2),(G_1^*,G_2^*))} > 0. \tag{16}$$

Proof for the local part (15)**:** Assume by contrary that the claim in equation (15) does not hold. Then, we can find a sequence of mixing measure pairs (G_1^n, G_2^n) taking the form $G_1^n := \sum_{i=1}^{k_1^n} \omega_i^n \delta_{(\kappa_i^n,\tau_i^n)}$, $G_2^n := \sum_{i=1}^{k_2^n} \exp(\beta_{0i}^n)\delta_{(\beta_{1i}^n,\eta_i^n,\nu_i^n)}$ for $n\in\mathbb{N}$ such that $\mathcal{D}_{1n} := \mathcal{D}_1((G_1^n,G_2^n),(G_1^*,G_2^*)) \to 0$ and

$$\mathbb{E}_X[V(f_{G_1^n,G_2^n}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X))]/\mathcal{D}_{1n} \to 0, \tag{17}$$

as $n\to\infty$. As our proof argument is asymptotic, we may assume that the number of shared and routed experts k_1^n, k_2^n do not vary with the sample size n. In addition, we also assume that Voronoi cells are independent of n, that is, $\mathcal{V}_{1,j_1} = \mathcal{V}_{1,j_1}(G_1^n)$ and $\mathcal{V}_{2,j_2} = \mathcal{V}_{2,j_2}(G_2^n)$, for all $j_1\in[k_1^*]$ and $j_2\in[k_2^*]$. Then, we can represent the Voronoi loss $\mathcal{D}_{1n}$ as

$$\begin{aligned}
\mathcal{D}_{1n} &= \sum_{j=1}^{k_1^*}\Big|\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n - \omega_j^*\Big| + \sum_{j=1}^{k_2^*}\Big|\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta_{0i}^n) - \exp(\beta_{0j}^*)\Big| \\
&+ \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|=1}\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n(\|\Delta\kappa_{ij}^n\| + |\Delta\tau_{ij}^n|) + \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta_{0i}^n)(\|\Delta\beta_{1ij}^n\| + \|\Delta\eta_{ij}^n\| + |\Delta\nu_{ij}^n|) \\
&+ \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|>1}\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n(\|\Delta\kappa_{ij}^n\|^2 + |\Delta\tau_{ij}^n|^2) + \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta_{0i}^n)(\|\Delta\beta_{1ij}^n\|^2 + \|\Delta\eta_{ij}^n\|^2 + |\Delta\nu_{ij}^n|^2),
\end{aligned} \tag{18}$$

where we denote $\Delta\kappa_{ij}^n := \kappa_i^n - \kappa_j^*$, $\Delta\tau_{ij}^n := \tau_i^n - \tau_j^*$, $\Delta\beta_{1ij}^n := \beta_{1i}^n - \beta_{1j}^*$, $\Delta\eta_{ij}^n := \eta_i^n - \eta_j^*$, and $\Delta\nu_{ij}^n := \nu_i^n - \nu_j^*$. Recall that $\mathcal{D}_{1n}\to 0$ as $n\to\infty$, then it follows that $\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n \to \omega_j^*$, $(\kappa_i^n,\tau_i^n)\to(\kappa_j^*,\tau_j^*)$ as $n\to\infty$ for all $i\in\mathcal{V}_{1,j}$ and $j\in[k_1^*]$. Furthermore, we also have $\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta_{0i}^n) - \exp(\beta_{0j}^*)$, $(\beta_{1i}^n,\eta_i^n,\nu_i^n)\to(\beta_{1j}^*,\eta_j^*,\nu_j^*)$ as $n\to\infty$ for all $i\in\mathcal{V}_{2,j}$ and $j\in[k_2^*]$.

Subsequently, we partition the rest of this proof into three main stages:

Stage 1 - Density Decomposition: In this stage, we focus on decomposing the density difference $f_{G_1^n,G_2^n}(Y|X) - f_{G_1^*,G_2^*}(Y|X)$. For ease of presentation, let us denote

$$q_{G_1^n}(Y|X) := \sum_{i=1}^{k_1^n}\omega_i^n\pi(Y|h_1(X,\kappa_i^n),\tau_i^n),$$

$$q_{G_1^*}(Y|X) := \sum_{i=1}^{k_1^*} \omega_i^* \pi(Y|h_1(X,\kappa_i^*), \tau_i^*),$$

$$p_{G_2^n}(Y|X) := \sum_{i=1}^{k_2^n} \frac{\exp((\beta_{1i}^n)^\top X + \beta_{0i}^n)}{\sum_{j=1}^{k_2^n} \exp((\beta_{1j}^n)^\top X + \beta_{0j}^n)} \cdot \pi(Y|h_2(X,\eta_i^n), \nu_i^n),$$

$$p_{G_2^*}(Y|X) := \sum_{i=1}^{k_2^*} \frac{\exp((\beta_{1i}^*)^\top X + \beta_{0i}^*)}{\sum_{j=1}^{k_2^*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)} \cdot \pi(Y|h_2(X,\eta_i^*), \nu_i^*).$$

Then, we have

$$f_{G_1^n,G_2^n}(Y|X) - f_{G_1^*,G_2^*}(Y|X) = \frac{1}{2}\left[(q_{G_1^n}(Y|X) - q_{G_1^*}(Y|X)) + (p_{G_2^n}(Y|X) - p_{G_2^*}(Y|X))\right].$$

Stage 1.1: In this step, we decompose the term $q_{G_1^n}(Y|X) - q_{G_1^*}(Y|X)$ as

$$\begin{aligned} q_{G_1^n}(Y|X) - q_{G_1^*}(Y|X) &= \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|=1} \sum_{i\in\mathcal{V}_{1,j}} \omega_i^n [\pi(Y|h_1(X,\kappa_i^n),\tau_i^n) - \pi(Y|h_1(X,\kappa_j^*),\tau_j^*)] \\ &+ \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|>1} \sum_{i\in\mathcal{V}_{1,j}} \omega_i^n [\pi(Y|h_1(X,\kappa_i^n),\tau_i^n) - \pi(Y|h_1(X,\kappa_j^*),\tau_j^*)] \\ &+ \sum_{j=1}^{k_1^*} \Big(\sum_{i\in\mathcal{V}_{1,j}} \omega_i^n - \omega_j^*\Big) \pi(Y|h_1(X,\kappa_j^*),\tau_j^*) \\ &:= A_{n,1}(Y|X) + A_{n,2}(Y|X) + A_{n,0}(Y|X). \end{aligned}$$

By applying the first-order and second-order Taylor expansions to the function $\pi(Y|h_1(X,\kappa_i^n),\tau_i^n))$ around the point (κ_j^*,τ_j^*), respectively, we have

$$A_{n,1}(Y|X) = \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|=1} \sum_{i\in\mathcal{V}_{1,j}} \omega_i^n \sum_{|\alpha|=1} \frac{1}{\alpha!} (\Delta\kappa_{ij}^n)^{\alpha_1} (\Delta\tau_{ij}^n)^{\alpha_2} \cdot \frac{\partial \pi}{\partial\kappa^{\alpha_1}\partial\tau^{\alpha_2}}(Y|h_1(X,\kappa_j^*),\tau_j^*) + R_{n,1}(Y|X),$$

$$A_{n,2}(Y|X) = \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|>1} \sum_{i\in\mathcal{V}_{1,j}} \omega_i^n \sum_{|\alpha|=1}^{2} \frac{1}{\alpha!} (\Delta\kappa_{ij}^n)^{\alpha_1} (\Delta\tau_{ij}^n)^{\alpha_2} \cdot \frac{\partial^{|\alpha|} \pi}{\partial\kappa^{\alpha_1}\partial\tau^{\alpha_2}}(Y|h_1(X,\kappa_j^*),\tau_j^*) + R_{n,2}(Y|X),$$

where $R_{n,1}(Y|X)$ and $R_{n,2}(Y|X)$ are the Taylor remainders such that $R_{n,1}(Y|X)/\mathcal{D}_{1n} \to 0$ as $n \to \infty$. By the chain rule, the first-order derivatives of the function π with respect to its parameters κ and τ are given by

$$\begin{aligned} \frac{\partial \pi}{\partial \kappa^{(u_1)}}(Y|h_1(X,\kappa_j^*),\tau_j^*) &= \frac{\partial h_1}{\partial \kappa^{(u_1)}}(X,\kappa_j^*)\frac{\partial \pi}{\partial h_1}(Y|h_1(X,\kappa_j^*),\tau_j^*), \\ \frac{\partial \pi}{\partial \tau}(Y|h_1(X,\kappa_j^*),\tau_j^*) &= \frac{1}{2}\frac{\partial^2 \pi}{\partial h_1^2}(Y|h_1(X,\kappa_j^*),\tau_j^*), \end{aligned}$$

for all $u_1 \in [d_1]$. Analogously, the second-order derivatives of the function π w.r.t its parameters are calculated as

$$\frac{\partial^2 \pi}{\partial \kappa^{(u_1)}\partial \kappa^{(v_1)}}(Y|h_1(X,\kappa_j^*),\tau_j^*) = \frac{\partial^2 h_1}{\partial \kappa^{(u_1)}\partial \kappa^{(v_1)}}(X,\kappa_j^*)\frac{\partial \pi}{\partial h_1}(Y|h_1(X,\kappa_j^*),\tau_j^*)$$

$$
\begin{aligned}
&+ \frac{\partial h_1}{\partial \kappa^{(u_1)}}(X,\kappa_j^*)\frac{\partial h_1}{\partial \kappa^{(v_1)}}(X,\kappa_j^*)\frac{\partial^2 \pi}{\partial h_1^2}(Y|h_1(X,\kappa_j^*),\tau_j^*),\\
\frac{\partial^2 \pi}{\partial \tau^2}(Y|h_1(X,\kappa_j^*),\tau_j^*) &= \frac{1}{4}\frac{\partial^4 \pi}{\partial h_1^4}(Y|h_1(X,\kappa_j^*),\tau_j^*),\\
\frac{\partial^2 \pi}{\partial \kappa^{(u_1)}\partial \tau}(Y|h_1(X,\kappa_j^*),\tau_j^*) &= \frac{1}{2}\frac{\partial h_1}{\partial \kappa^{(u_1)}}(X,\kappa_j^*)\frac{\partial^3 \pi}{\partial h_1^3}(Y|h_1(X,\kappa_j^*),\tau_j^*),
\end{aligned}
$$

for all $u_1, v_1 \in [d_1]$. Combine the above results, we can rewrite $A_{n,1}(Y|X)$ as

$$
A_{n,1}(Y|X) = \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|=1}\Big[A_{n,1,1}^{(j)}(X)\frac{\partial \pi}{\partial h_1}(Y|h_1(X,\kappa_j^*),\tau_j^*) + A_{n,1,2}^{(j)}(X)\frac{\partial^2 \pi}{\partial h_1^2}(Y|h_1(X,\kappa_j^*),\tau_j^*)\Big] + R_{n,1}(Y|X),
$$

where we denote

$$
\begin{aligned}
A_{n,1,1}^{(j)}(X) &:= \sum_{i\in\mathcal{V}_{1,j}}\omega_i^n\sum_{u_1=1}^{d_1}(\Delta\kappa_{ij}^n)^{(u_1)}\frac{\partial h_1}{\partial \kappa^{(u_1)}}(X,\kappa_j^*),\\
A_{n,1,2}^{(j)}(X) &:= \sum_{i\in\mathcal{V}_{1,j}}\omega_i^n\frac{1}{2}(\Delta\tau_{ij}^n),
\end{aligned}
$$

for all $j \in [k_1^*]$ such that $|\mathcal{V}_{1,j}| = 1$. Similarly, the quantity $A_{n,2}(Y|X)$ can be represented as

$$
\begin{aligned}
A_{n,2}(Y|X) = \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|>1}\Big[\, & A_{n,2,1}^{(j)}(X)\frac{\partial \pi}{\partial h_1}(Y|h_1(X,\kappa_j^*),\tau_j^*) + A_{n,2,2}^{(j)}(X)\frac{\partial^2 \pi}{\partial h_1^2}(Y|h_1(X,\kappa_j^*),\tau_j^*)\\
&+ A_{n,2,3}^{(j)}(X)\frac{\partial^3 \pi}{\partial h_1^3}(Y|h_1(X,\kappa_j^*),\tau_j^*) + A_{n,2,4}^{(j)}(X)\frac{\partial^4 \pi}{\partial h_1^4}(Y|h_1(X,\kappa_j^*),\tau_j^*)\Big] + R_{n,2}(Y|X),
\end{aligned}
$$

where we denote

$$
\begin{aligned}
A_{n,2,1}^{(j)}(X) &:= \sum_{i\in\mathcal{V}_{1,j}}\omega_i^n\Big(\sum_{u_1=1}^{d_1}(\Delta\kappa_{ij}^n)^{(u_1)}\frac{\partial h_1}{\partial \kappa^{(u_1)}}(X,\kappa_j^*) + \sum_{u_1,v_1=1}^{d_1}\frac{(\Delta\kappa_{ij}^n)^{(u_1)}(\Delta\kappa_{ij}^n)^{(v_1)}}{1+1_{\{u_1=v_1\}}}\frac{\partial^2 h_1}{\partial \kappa^{(u_1)}\partial \kappa^{(v_1)}}(X,\kappa_j^*)\Big),\\
A_{n,2,2}^{(j)}(X) &:= \sum_{i\in\mathcal{V}_{1,j}}\omega_i^n\Big(\frac{1}{2}(\Delta\tau_{ij}^n) + \sum_{u_1,v_1=1}^{d_1}\frac{(\Delta\kappa_{ij}^n)^{(u_1)}(\Delta\kappa_{ij}^n)^{(v_1)}}{1+1_{\{u_1=v_1\}}}\frac{\partial h_1}{\partial \kappa^{(u_1)}}(X,\kappa_j^*)\frac{\partial h_1}{\partial \kappa^{(v_1)}}(X,\kappa_j^*)\Big),\\
A_{n,2,3}^{(j)}(X) &:= \sum_{i\in\mathcal{V}_{1,j}}\omega_i^n\sum_{u_1=1}^{d_1}\frac{1}{2}(\Delta\kappa_{ij}^n)^{(u_1)}(\Delta\tau_{ij}^n)\frac{\partial h_1}{\partial \kappa^{(u_1)}}(X,\kappa_j^*),\\
A_{n,2,4}^{(j)}(X) &:= \sum_{i\in\mathcal{V}_{1,j}}\omega_i^n\frac{1}{8}(\Delta\tau_{ij}^n)^2,
\end{aligned}
$$

for all $j \in [k_1^*]$ such that $|\mathcal{V}_{1,j}| > 1$.

Stage 1.2: In this step, we decompose the term $Q_n(Y|X) := \Big[\sum_{j=1}^{k_2^*}\exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)\Big]\cdot[p_{G_2^n}(Y|X) - p_{G_2^*}(Y|X)]$. By denoting $F(Y|X;\beta_1,\eta,\nu) := \exp(\beta_1^\top X)\pi(Y|h_2(X,\eta),\nu)$ and $H(Y|X;\beta_1) :=$

$\exp(\beta_1^\top X)p_{G_2}(Y|X)$, we can represent $Q_n(Y|X)$ as

$$\begin{aligned}
Q_n(Y|X) &= \sum_{j=1}^{k_2^*}\sum_{i\in\mathcal{V}_{2,j}} \exp(\beta_{0i}^n)[F(Y|X;\beta_{1i}^n,\eta_i^n,\nu_i^n) - F(Y|X;\beta_{1j}^*,\eta_j^*,\nu_j^*)] \\
&- \sum_{j=1}^{k_2^*}\sum_{i\in\mathcal{V}_{2,j}} \exp(\beta_{0i}^n)[H(Y|X;\beta_{1i}^n) - H(Y|X;\beta_{1j}^*)] \\
&+ \sum_{j=1}^{k_2^*}\Big(\sum_{i\in\mathcal{V}_{2,j}} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^*)\Big)[F(Y|X;\beta_{1j}^*,\eta_j^*,\nu_j^*) - H(Y|X;\beta_{1j}^*)] \\
&:= B_n(Y|X) - C_n(Y|X) + E_n(Y|X).
\end{aligned}$$

Stage 1.2.1: In this step, we decompose the term $B_n(Y|X)$:

$$\begin{aligned}
B_n(Y|X) &= \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\sum_{i\in\mathcal{V}_{2,j}} \exp(\beta_{0i}^n)[F(Y|X;\beta_{1i}^n,\eta_i^n,\nu_i^n) - F(Y|X;\beta_{1j}^*,\eta_j^*,\nu_j^*)] \\
&+ \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}\sum_{i\in\mathcal{V}_{2,j}} \exp(\beta_{0i}^n)[F(Y|X;\beta_{1i}^n,\eta_i^n,\nu_i^n) - F(Y|X;\beta_{1j}^*,\eta_j^*,\nu_j^*)] \\
&:= B_{n,1}(Y|X) + B_{n,2}(Y|X).
\end{aligned}$$

By applying the first-order and second-order Taylor expansions to the function $F(Y|X;\beta_{1i}^n,\eta_i^n,\nu_i^n)$ around the point $(\beta_{1j}^*,\eta_j^*,\nu_j^*)$, we have

$$\begin{aligned}
B_{n,1}(Y|X) &= \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\sum_{i\in\mathcal{V}_{2,j}} \exp(\beta_{0i}^n)\sum_{|\alpha|=1}\frac{1}{\alpha!}(\Delta\beta_{1ij}^n)^{\alpha_1}(\Delta\eta_{ij}^n)^{\alpha_2}(\Delta\nu_{ij}^n)^{\alpha_3} \\
&\qquad\times\frac{\partial F}{\partial\beta_1^{\alpha_1}\partial\eta^{\alpha_2}\partial\nu^{\alpha_3}}(Y|X;\beta_{1j}^*,\eta_j^*,\nu_j^*) + R_{n,3}(Y|X), \\
B_{n,2}(Y|X) &= \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}\sum_{i\in\mathcal{V}_{2,j}} \exp(\beta_{0i}^n)\sum_{|\alpha|=1}^{2}\frac{1}{\alpha!}(\Delta\beta_{1ij}^n)^{\alpha_1}(\Delta\eta_{ij}^n)^{\alpha_2}(\Delta\nu_{ij}^n)^{\alpha_3} \\
&\qquad\times\frac{\partial^{|\alpha|} F}{\partial\beta_1^{\alpha_1}\partial\eta^{\alpha_2}\partial\nu^{\alpha_3}}(Y|X;\beta_{1j}^*,\eta_j^*,\nu_j^*) + R_{n,4}(Y|X),
\end{aligned}$$

where $R_{n,3}(Y|X)$ and $R_{n,4}(Y|X)$ are the Taylor remainders such that $R_{n,3}(Y|X)/\mathcal{D}_{1n} \to 0$ and $R_{n,4}(Y|X)/\mathcal{D}_{1n} \to 0$ as $n\to\infty$. By means of the chain rule, the first-order derivatives of the function F w.r.t its parameters β_1,η,ν are given by

$$\begin{aligned}
\frac{\partial F}{\partial\beta_1^{(u)}}(Y|X;\beta_{1j}^*,\eta_j^*,\nu_j^*) &= X^{(u)}\exp((\beta_{1j}^*)^\top X)\pi(Y|h_2(X,\eta_j^*),\nu_j^*), \\
\frac{\partial F}{\partial\eta^{(u_2)}}(Y|X;\beta_{1j}^*,\eta_j^*,\nu_j^*) &= \frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\eta_j^*)\exp((\beta_{1j}^*)^\top X)\frac{\partial\pi}{\partial h_2}(Y|h_2(X,\eta_j^*),\nu_j^*), \\
\frac{\partial F}{\partial\nu}(Y|X;\beta_{1j}^*,\eta_j^*,\nu_j^*) &= \frac{1}{2}\exp((\beta_{1j}^*)^\top X)\frac{\partial^2\pi}{\partial h_2^2}(Y|h_2(X,\eta_j^*),\nu_j^*),
\end{aligned}$$

for all $u_2 \in [d_2]$. Similarly, we can derive the second-order derivatives of the function F w.r.t its parameters as follows:

$$\begin{aligned}
\frac{\partial^2 F}{\partial \beta_1^{(u)} \partial \beta_1^{(v)}}(Y|X;\beta_{1j}^*,\eta_j^*,\nu_j^*) &= X^{(u)}X^{(v)}\exp((\beta_{1j}^*)^\top X)\pi(Y|h_2(X,\eta_j^*),\nu_j^*), \\
\frac{\partial^2 F}{\partial \eta^{(u_2)} \partial \eta^{(v_2)}}(Y|X;\beta_{1j}^*,\eta_j^*,\nu_j^*) &= \frac{\partial^2 h_2}{\partial \eta^{(u_2)}\partial \eta^{(v_2)}}(X,\eta_j^*)\exp((\beta_{1j}^*)^\top X)\frac{\partial \pi}{\partial h_2}(Y|h_2(X,\eta_j^*),\nu_j^*) \\
&\quad + \frac{\partial h_2}{\partial \eta^{(u_2)}}(X,\eta_j^*)\frac{\partial h_2}{\partial \eta^{(v_2)}}(X,\eta_j^*)\exp((\beta_{1j}^*)^\top X)\frac{\partial^2 \pi}{\partial h_2^2}(Y|h_2(X,\eta_j^*),\nu_j^*), \\
\frac{\partial^2 F}{\partial \nu^2}(Y|X;\beta_{1j}^*,\eta_j^*,\nu_j^*) &= \frac{1}{4}\exp((\beta_{1j}^*)^\top X)\frac{\partial^4 \pi}{\partial h_2^4}(Y|h_2(X,\eta_j^*),\nu_j^*),
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial^2 F}{\partial \beta_1^{(u)} \partial \eta^{(v_2)}}(Y|X;\beta_{1j}^*,\eta_j^*,\nu_j^*) &= X^{(u)}\frac{\partial h_2}{\partial \eta^{(v_2)}}(X,\eta_j^*)\exp((\beta_{1j}^*)^\top X)\frac{\partial \pi}{\partial h_2}(Y|h_2(X,\eta_j^*),\nu_j^*), \\
\frac{\partial^2 F}{\partial \beta_1^{(u)} \partial \nu}(Y|X;\beta_{1j}^*,\eta_j^*,\nu_j^*) &= \frac{1}{2}X^{(u)}\exp((\beta_{1j}^*)^\top X)\frac{\partial^2 \pi}{\partial h_2^2}(Y|h_2(X,\eta_j^*),\nu_j^*), \\
\frac{\partial^2 F}{\partial \eta^{(u_2)} \partial \nu}(Y|X;\beta_{1j}^*,\eta_j^*,\nu_j^*) &= \frac{1}{2}\frac{\partial h_2}{\partial \eta^{(u_2)}}(X,\eta_j^*)\exp((\beta_{1j}^*)^\top X)\frac{\partial^3 \pi}{\partial h_2^3}(Y|h_2(X,\eta_j^*),\nu_j^*),
\end{aligned}$$

for all $u_2, v_2 \in [d_2]$. Putting the above results together, we can rewrite $B_{n,1}(Y|X)$ as

$$\begin{aligned}
B_{n,1}(Y|X) = \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1} \Big[& B_{n,1,0}^{(j)}(X)\pi(Y|h_2(X,\eta_j^*),\nu_j^*) + B_{n,1,1}^{(j)}(X)\frac{\partial \pi}{\partial h_2}(Y|h_2(X,\eta_j^*),\nu_j^*) \\
& + B_{n,1,2}^{(j)}(X)\frac{\partial^2 \pi}{\partial h_2^2}(Y|h_2(X,\eta_j^*),\nu_j^*)\Big] + R_{n,3}(Y|X),
\end{aligned}$$

where we denote

$$\begin{aligned}
B_{n,1,0}^{(j)}(X) &:= \sum_{i\in\mathcal{V}_{2,j}}\exp(\beta_{0i}^n)\sum_{u=1}^{d}(\Delta\beta_{1ij}^n)^{(u)}X^{(u)}\exp((\beta_{1j}^*)^\top X), \\
B_{n,1,1}^{(j)}(X) &:= \sum_{i\in\mathcal{V}_{2,j}}\exp(\beta_{0i}^n)\sum_{u_2=1}^{d_2}(\Delta\eta_{ij}^n)^{(u_2)}\frac{\partial h_2}{\partial \eta^{(u_2)}}(X,\eta_j^*)\exp((\beta_{1j}^*)^\top X), \\
B_{n,1,2}^{(j)}(X) &:= \sum_{i\in\mathcal{V}_{2,j}}\exp(\beta_{0i}^n)\frac{1}{2}(\Delta\nu_{ij}^n)\exp((\beta_{1j}^*)^\top X),
\end{aligned}$$

for all $j \in [k_2^*]$ such that $|\mathcal{V}_{2,j}| = 1$. Analogously, we can represent the term $B_{n,2}(Y|X)$ as

$$B_{n,2}(Y|X) = \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\sum_{\rho=0}^{4}B_{n,2,\rho}^{(j)}(X)\frac{\partial^\rho \pi}{\partial h_2^\rho}(Y|h_2(X,\eta_j^*),\nu_j^*) + R_{n,4}(Y|X),$$

where we define

$$B^{(j)}_{n,2,0}(X) := \sum_{i\in\mathcal{V}_{2,j}} \exp(\beta^n_{0i})\Bigg[\sum_{u=1}^{d}(\Delta\beta^n_{1ij})^{(u)}X^{(u)} + \sum_{u,v=1}^{d}\frac{(\Delta\beta^n_{1ij})^{(u)}(\Delta\beta^n_{1ij})^{(v)}}{1+1_{\{u=v\}}}X^{(u)}X^{(v)}\Bigg]\exp((\beta^*_{1j})^\top X),$$

$$B^{(j)}_{n,2,1}(X) := \sum_{i\in\mathcal{V}_{2,j}} \exp(\beta^n_{0i})\Bigg[\sum_{u_2=1}^{d_2}(\Delta\eta^n_{ij})^{(u_2)}\frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\eta^*_j) + \sum_{u_2,v_2=1}^{d_2}\frac{(\Delta\eta^n_{ij})^{(u_2)}(\Delta\eta^n_{ij})^{(v_2)}}{1+1_{\{u_2=v_2\}}}\frac{\partial^2 h_2}{\partial\eta^{(u_2)}\partial\eta^{(v_2)}}(X,\eta^*_j)$$
$$+\sum_{u=1}^{d}\sum_{v_2=1}^{d_2}(\Delta\beta^n_{1ij})^{(u)}(\Delta\eta^n_{ij})^{(v_2)}X^{(u)}\frac{\partial h_2}{\partial\eta^{(v_2)}}(X,\eta^*_j)\Bigg]\exp((\beta^*_{1j})^\top X),$$

$$B^{(j)}_{n,2,2}(X) := \sum_{i\in\mathcal{V}_{2,j}} \exp(\beta^n_{0i})\Bigg[\frac{1}{2}(\Delta\nu^n_{ij}) + \sum_{u_2,v_2=1}^{d_2}\frac{(\Delta\eta^n_{ij})^{(u_2)}(\Delta\eta^n_{ij})^{(v_2)}}{1+1_{\{u_2=v_2\}}}\frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\eta^*_j)\frac{\partial h_2}{\partial\eta^{(v_2)}}(X,\eta^*_j)$$
$$+\sum_{u=1}^{d}\frac{1}{2}(\Delta\beta^n_{1ij})^{(u)}(\Delta\nu^n_{ij})X^{(u)}\Bigg]\exp((\beta^*_{1j})^\top X),$$

$$B^{(j)}_{n,2,3}(X) := \sum_{i\in\mathcal{V}_{2,j}} \exp(\beta^n_{0i})\sum_{u_2=1}^{d_2}\frac{1}{2}(\Delta\eta^n_{ij})^{(u_2)}(\Delta\nu^n_{ij})\frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\eta^*_j)\exp((\beta^*_{1j})^\top X),$$

$$B^{(j)}_{n,2,4}(X) := \sum_{i\in\mathcal{V}_{2,j}} \exp(\beta^n_{0i})\frac{1}{8}(\Delta\nu^n_{ij})^2\exp((\beta^*_{1j})^\top X),$$

for all $j\in[k^*_2]$ such that $|\mathcal{V}_{2,j}|>1$.

Stage 1.2.2: In this step, we decompose the term $C_n(Y|X)$:

$$\begin{aligned}C_n(Y|X) &= \sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|=1}\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})[H(Y|X;\beta^n_{1i})-H(Y|X;\beta^*_{1j})]\\ &+\sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|>1}\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})[H(Y|X;\beta^n_{1i})-H(Y|X;\beta^*_{1j})]\\ &:= C_{n,1}(Y|X)+C_{n,2}(Y|X).\end{aligned}$$

By means of the first-order and second-order Taylor expansions to the function $H(Y|X;\beta^n_{1i})$ around the point β^*_{1j}, we get

$$C_{n,1}(Y|X) = \sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|=1}\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})\sum_{u=1}^{d}(\Delta\beta^n_{1ij})^{(u)}X^{(u)}H(Y|X;\beta^*_{1j})+R_{n,5}(Y|X),$$

$$C_{n,2}(Y|X) = \sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|>1}\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})\Bigg[\sum_{u=1}^{d}(\Delta\beta^n_{1ij})^{(u)}X^{(u)}H(Y|X;\beta^*_{1j})$$
$$+\sum_{u,v=1}^{d}\frac{(\Delta\beta^n_{1ij})^{(u)}(\Delta\beta^n_{1ij})^{(v)}}{1+1_{\{u=v\}}}X^{(u)}X^{(v)}H(Y|X;\beta^*_{1j})\Bigg]+R_{n,6}(Y|X),$$

where $R_{n,5}(Y|X)$ and $R_{n,6}(Y|X)$ are the Taylor remainders such that $R_{n,5}(Y|X)/\mathcal{D}_{1n} \to 0$ and $R_{n,6}(Y|X)/\mathcal{D}_{1n} \to 0$ as $n \to \infty$.

Putting the above decompositions together, we can view $A_{n,0}(Y|X)/\mathcal{D}_{1n}$, $[A_{n,1}(Y|X) - R_{n,1}(Y|X)]/\mathcal{D}_{1n}$, $[A_{n,2}(Y|X) - R_{n,2}(Y|X)]/\mathcal{D}_{1n}$, $[B_{n,1}(Y|X) - R_{n,3}(Y|X)]/\mathcal{D}_{1n}$, $[B_{n,2}(Y|X) - R_{n,4}(Y|X)]/\mathcal{D}_{1n}$, $[C_{n,1}(Y|X) - R_{n,5}(Y|X)]/\mathcal{D}_{1n}$, $[C_{n,2}(Y|X) - R_{n,6}(Y|X)]/\mathcal{D}_{1n}$, and $E_n(Y|X)/\mathcal{D}_{1n}$ as a combination of elements of the following sets

$$\mathcal{S}_{0,j} := \{\pi(Y|h_1(X,\kappa^*_j),\tau^*_j)\},$$

$$\mathcal{S}_{1,j} := \left\{\frac{\partial h_1}{\partial \kappa^{(u_1)}}(X,\kappa^*_j)\frac{\partial \pi}{\partial h_1}(Y|h_1(X,\kappa^*_j),\tau^*_j),\ \frac{\partial^2 h_1}{\partial \kappa^{(u_1)}\partial \kappa^{(v_1)}}(X,\kappa^*_j)\frac{\partial \pi}{\partial h_1}(Y|h_1(X,\kappa^*_j),\tau^*_j) : u_1, v_1 \in [d_1]\right\},$$

$$\mathcal{S}_{2,j} := \left\{\frac{\partial^2 \pi}{\partial h_1^2}(Y|h_1(X,\kappa^*_j),\tau^*_j),\ \frac{\partial h_1}{\partial \kappa^{(u_1)}}(X,\kappa^*_j)\frac{\partial h_1}{\partial \kappa^{(v_1)}}(X,\kappa^*_j)\frac{\partial^2 \pi}{\partial h_1^2}(Y|h_1(X,\kappa^*_j),\tau^*_j) : u_1, v_1 \in [d_1]\right\},$$

$$\mathcal{S}_{3,j} := \left\{\frac{\partial h_1}{\partial \kappa^{(u_1)}}(X,\kappa^*_j)\frac{\partial^3 \pi}{\partial h_1^3}(Y|h_1(X,\kappa^*_j),\tau^*_j) : u_1, v_1 \in [d_1]\right\},$$

$$\mathcal{S}_{4,j} := \left\{\frac{\partial^4 \pi}{\partial h_1^4}(Y|h_1(X,\kappa^*_j),\tau^*_j) : u_1, v_1 \in [d_1]\right\},$$

for all $j \in [k^*_1]$, and

$$\mathcal{T}_{0,j} := \{F(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j),\ X^{(u_2)}F(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j),\ X^{(u_2)}X^{(v_2)}F(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j) : u_2, v_2 \in [d_2]\},$$

$$\mathcal{T}_{1,j} := \left\{\frac{\partial h_2}{\partial \eta^{(u_2)}}(X,\eta^*_j)F_1(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j),\ \frac{\partial^2 h_2}{\partial \eta^{(u_2)}\partial \eta^{(v_2)}}(X,\eta^*_j)F_1(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j),\right.$$
$$\left. X^{(u_2)}\frac{\partial h_2}{\partial \eta^{(u_2)}}(X,\eta^*_j)F_1(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j) : u_2, v_2 \in [d_2]\right\},$$

$$\mathcal{T}_{2,j} := \left\{F_2(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j),\ \frac{\partial h_2}{\partial \eta^{(u_2)}}(X,\eta^*_j)\frac{\partial h_2}{\partial \eta^{(v_2)}}(X,\eta^*_j)F_2(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j),\right.$$
$$\left. X^{(u_2)}F_2(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j) : u_2, v_2 \in [d_2]\right\},$$

$$\mathcal{T}_{3,j} := \left\{\frac{\partial h_2}{\partial \eta^{(u_2)}}(X,\eta^*_j)F_3(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j) : u_2 \in [d_2]\right\},$$

$$\mathcal{T}_{4,j} := \{F_4(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j)\},$$

$$\mathcal{T}_{5,j} := \{H(Y|X;\beta^*_{1j}),\ X^{(u)}H(Y|X;\beta^*_{1j}),\ X^{(u)}X^{(v)}H(Y|X;\beta^*_{1j}) : u, v \in [d]\},$$

where we denote

$$F_\rho(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j) := \exp((\beta^*_{1j})^\top X)\frac{\partial^\rho \pi}{\partial h_1^\rho}(Y|h_2(X,\eta^*_j),\nu^*_j),$$

for all $\rho \in [4]$ and $j \in [k^*_2]$.

Stage 2 - Non-vanishing coefficients: In this stage, we show by contradiction that not all the coefficients in the representations of $A_{n,0}(Y|X)/\mathcal{D}_{1n}$, $[A_{n,1}(Y|X)-R_{n,1}(Y|X)]/\mathcal{D}_{1n}$, $[A_{n,2}(Y|X)-R_{n,2}(Y|X)]/\mathcal{D}_{1n}$, $[B_{n,1}(Y|X)-R_{n,3}(Y|X)]/\mathcal{D}_{1n}$, $[B_{n,2}(Y|X)-R_{n,4}(Y|X)]/\mathcal{D}_{1n}$, $[C_{n,1}(Y|X)-R_{n,5}(Y|X)]/\mathcal{D}_{1n}$, $[C_{n,2}(Y|X)-R_{n,6}(Y|X)]/\mathcal{D}_{1n}$, and $E_n(Y|X)/\mathcal{D}_{1n}$ converge to zero as $n\to\infty$. In particular, we assume that all those coefficients go to zero. Then, by looking into the coefficients of the terms:

- $\pi(Y|h_1(X,\kappa^*_j),\tau^*_j)$ for $j\in[k^*_1]$, we have $\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j=1}^{k^*_1}\Big|\sum_{i\in\mathcal{V}_{1,j}}\omega^n_i-\omega^*_j\Big|\to 0$;
- $\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa^*_j)\frac{\partial\pi}{\partial h_1}(Y|h_1(X,\kappa^*_j),\tau^*_j)$ for $j\in[k^*_1]:|\mathcal{V}_{1,j}|=1$ and $u_1\in[d_1]$, we have
$$\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j\in[k^*_1]:|\mathcal{V}_{1,j}|=1}\sum_{i\in\mathcal{V}_{1,j}}\omega^n_i\|\Delta\kappa^n_{ij}\|\to 0;$$
- $\frac{\partial^2\pi}{\partial h_1^2}(Y|h_1(X,\kappa^*_j),\tau^*_j)$ for $j\in[k^*_1]:|\mathcal{V}_{1,j}|=1$, we have
$$\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j\in[k^*_1]:|\mathcal{V}_{1,j}|=1}\sum_{i\in\mathcal{V}_{1,j}}\omega^n_i|\Delta\tau^n_{ij}|\to 0;$$
- $\big[\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa^*_j)\big]^2\frac{\partial^2\pi}{\partial h_1^2}(Y|h_1(X,\kappa^*_j),\tau^*_j)$ for $j\in[k^*_1]:|\mathcal{V}_{1,j}|>1$ and $u_1\in[d_1]$, we have
$$\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j\in[k^*_1]:|\mathcal{V}_{1,j}|>1}\sum_{i\in\mathcal{V}_{1,j}}\omega^n_i\|\Delta\kappa^n_{ij}\|^2\to 0;$$
- $\frac{\partial^4\pi}{\partial h_1^4}(Y|h_1(X,\kappa^*_j),\tau^*_j)$ for $j\in[k^*_1]:|\mathcal{V}_{1,j}|>1$ and $u_1\in[d_1]$, we have
$$\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j\in[k^*_1]:|\mathcal{V}_{1,j}|>1}\sum_{i\in\mathcal{V}_{1,j}}\omega^n_i|\Delta\tau^n_{ij}|^2\to 0;$$
- $F(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j)$ for $j\in[k^*_2]$, we have $\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j=1}^{k^*_2}\Big|\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})-\exp(\beta^*_{1j})\Big|\to 0$;
- $X^{(u)}F(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j)$ for $j\in[k^*_2]:|\mathcal{V}_{2,j}|=1$ and $u\in[d]$, we have
$$\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|=1}\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})\|\Delta\beta^n_{1ij}\|\to 0;$$
- $\frac{\partial h_2}{\partial\eta^{(u_2)}}F_1(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j)$ for $j\in[k^*_2]:|\mathcal{V}_{2,j}|=1$ and $u_2\in[d_2]$, we have
$$\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|=1}\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})\|\Delta\eta^n_{ij}\|\to 0;$$
- $F_2(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j)$ for $j\in[k^*_2]:|\mathcal{V}_{2,j}|=1$, we have
$$\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|=1}\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})|\Delta\nu^n_{ij}|\to 0;$$

- $X^{(u)}X^{(v)}F(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j)$ for $j\in[k^*_2]:|\mathcal{V}_{2,j}|>1$ and $u,v\in[d]$, we have

$$\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|>1}\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})\|\Delta\beta^n_{1ij}\|^2\to 0;$$

- $\left[\frac{\partial^2 h_2}{\partial\eta^{(u_2)}}(X,\eta^*_j)\right]^2F_2(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j)$ for $j\in[k^*_2]:|\mathcal{V}_{2,j}|>1$ and $u_2\in[d_2]$, we have

$$\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|>1}\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})\|\Delta\eta^n_{ij}\|\to 0;$$

- $F_4(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j)$ for $j\in[k^*_2]:|\mathcal{V}_{2,j}|>1$, we have

$$\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|>1}\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})|\Delta\nu^n_{ij}|^2\to 0;$$

Taking the sum of the above limits, we deduce $1=\frac{1}{\mathcal{D}_{1n}}\cdot\mathcal{D}_{1n}\to 0$ as $n\to\infty$, which is a contradiction. Thus, at least one among the coefficients in the representations of $A_{n,0}(Y|X)/\mathcal{D}_{1n}$, $[A_{n,1}(Y|X)-R_{n,1}(Y|X)]/\mathcal{D}_{1n}$, $[A_{n,2}(Y|X)-R_{n,2}(Y|X)]/\mathcal{D}_{1n}$, $[B_{n,1}(Y|X)-R_{n,3}(Y|X)]/\mathcal{D}_{1n}$, $[B_{n,2}(Y|X)-R_{n,4}(Y|X)]/\mathcal{D}_{1n}$, $[C_{n,1}(Y|X)-R_{n,5}(Y|X)]/\mathcal{D}_{1n}$, $[C_{n,2}(Y|X)-R_{n,6}(Y|X)]/\mathcal{D}_{1n}$, and $E_n(Y|X)/\mathcal{D}_{1n}$ does not converge to zero.

Stage 3 - Fatou's lemma contradiction: In this stage, we use the Fatou's lemma to show a contradiction to the result of Stage 2. For that purpose, let us denote m_n as the maximum of the absolute values of the coefficients in the representations of $A_{n,0}(Y|X)/\mathcal{D}_{1n}$, $[A_{n,1}(Y|X)-R_{n,1}(Y|X)]/\mathcal{D}_{1n}$, $[A_{n,2}(Y|X)-R_{n,2}(Y|X)]/\mathcal{D}_{1n}$, $[B_{n,1}(Y|X)-R_{n,3}(Y|X)]/\mathcal{D}_{1n}$, $[B_{n,2}(Y|X)-R_{n,4}(Y|X)]/\mathcal{D}_{1n}$, $[C_{n,1}(Y|X)-R_{n,5}(Y|X)]/\mathcal{D}_{1n}$, $[C_{n,2}(Y|X)-R_{n,6}(Y|X)]/\mathcal{D}_{1n}$, and $E_n(Y|X)/\mathcal{D}_{1n}$. It follows from the result of Stage 2 that $1/m_n\not\to\infty$ as $n\to\infty$. In addition, we also denote

$$\frac{1}{m_n\mathcal{D}_{1n}}\cdot\sum_{i\in\mathcal{V}_{1,j}}\omega^n_i(\Delta\kappa^n_{ij})^{(u_1)}\to s^{(u_1)}_{1,j},\quad \frac{1}{m_n\mathcal{D}_{1n}}\cdot\sum_{i\in\mathcal{V}_{1,j}}\omega^n_i(\Delta\tau^n_{ij})\to s_{2,j},$$
$$\frac{1}{m_n\mathcal{D}_{1n}}\cdot\sum_{i\in\mathcal{V}_{1,j}}\omega^n_i(\Delta\kappa^n_{ij})^{(u_1)}(\Delta\kappa^n_{ij})^{(v_1)}\to s^{(u_1v_1)}_{3,j},\quad \frac{1}{m_n\mathcal{D}_{1n}}\cdot\sum_{i\in\mathcal{V}_{1,j}}\omega^n_i(\Delta\tau^n_{ij})^2\to s_{4,j},$$
$$\frac{1}{m_n\mathcal{D}_{1n}}\cdot\sum_{i\in\mathcal{V}_{1,j}}\omega^n_i(\Delta\kappa^n_{ij})^{(u_1)}(\Delta\tau^n_{ij})\to s^{(u_1)}_{5,j},\quad \frac{1}{m_n\mathcal{D}_{1n}}\cdot\Big(\sum_{i\in\mathcal{V}_{1,j}}\omega^n_i-\omega^*_j\Big)\to s_{0,j},$$

for all $j\in[k^*_1]$ and

$$\frac{1}{m_n\mathcal{D}_{1n}}\cdot\Big(\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})-\exp(\beta^*_{0j})\Big)\to t_{0,j},\quad \frac{1}{m_n\mathcal{D}_{1n}}\cdot\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})(\Delta\beta^n_{1ij})^{(u)}\to t^{(u)}_{1,j},$$
$$\frac{1}{m_n\mathcal{D}_{1n}}\cdot\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})(\Delta\eta^n_{ij})^{(u_2)}\to t^{(u_2)}_{2,j},\quad \frac{1}{m_n\mathcal{D}_{1n}}\cdot\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})(\Delta\nu^n_{ij})\to t_{3,j},$$

$$
\begin{gathered}
\frac{1}{m_n\mathcal{D}_{1n}}\cdot\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})(\Delta\beta^n_{1ij})^{(u)}(\Delta\beta^n_{1ij})^{(v)}\to t^{(uv)}_{4,j},\quad \frac{1}{m_n\mathcal{D}_{1n}}\cdot\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})(\Delta\eta^n_{ij})^{(u_2)}(\Delta\eta^n_{ij})^{(v_2)}\to t^{(u_2v_2)}_{5,j},\\
\frac{1}{m_n\mathcal{D}_{1n}}\cdot\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})(\Delta\nu^n_{ij})^{2}\to t_{6,j},\quad \frac{1}{m_n\mathcal{D}_{1n}}\cdot\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})(\Delta\beta^n_{1ij})^{(u)}(\Delta\eta^n_{ij})^{(v_2)}\to t^{(uv_2)}_{7,j},\\
\frac{1}{m_n\mathcal{D}_{1n}}\cdot\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})(\Delta\beta^n_{1ij})^{(u)}(\Delta\nu^n_{ij})\to t^{(u)}_{8,j},\quad \frac{1}{m_n\mathcal{D}_{1n}}\cdot\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})(\Delta\eta^n_{ij})^{(u_2)}(\Delta\nu^n_{ij})\to t^{(u_2)}_{9,j},
\end{gathered}
$$

for all $j\in[k^*_2]$ as $n\to\infty$. Due to the result of Stage 2, at least one among the above limits is different from zero. Recall from equation (17) that we have

$$
\mathbb{E}_X[V(f_{G^n_1,G^n_2}(\cdot|X),f_{G^*_1,G^*_2}(\cdot|X))]/\mathcal{D}_{1n}\to 0,
$$

Furthermore, according to the Fatou's lemma, we get

$$
\lim_{n\to\infty}\frac{\mathbb{E}_X[V(f_{G^n_1,G^n_2}(\cdot|X),f_{G^*_1,G^*_2}(\cdot|X))]}{m_n\mathcal{D}_{1n}}\geq\int\liminf_{n\to\infty}\frac{|f_{G^n_1,G^n_2}(Y|X)-f_{G^*_1,G^*_2}(Y|X)|}{2m_n\mathcal{D}_{1n}}\mathrm{d}(X,Y).
$$

Then, we deduce $[f_{G^n_1,G^n_2}(Y|X)-f_{G^*_1,G^*_2}(Y|X)]/[m_n\mathcal{D}_{1n}]\to 0$ as $n\to\infty$ for almost surely (X,Y). Since the input space is bounded and the parameter space is compact, the quantity $\sum_{j=1}^{k^*_2}\exp((\beta^*_{1j})^\top X+\beta^*_{0j})$ is bounded. Thus, we also have

$$
\Big[\sum_{j=1}^{k^*_2}\exp((\beta^*_{1j})^\top X+\beta^*_{0j})\Big][f_{G^n_1,G^n_2}(Y|X)-f_{G^*_1,G^*_2}(Y|X)]/[m_n\mathcal{D}_{1n}]\to 0,
$$

implying that

$$
\frac{1}{2}\Big[\sum_{j=1}^{k^*_2}\exp((\beta^*_{1j})^\top X+\beta^*_{0j})\Big]\cdot\frac{q_{G^n_1}(Y|X)-q_{G^*_1}(Y|X)}{m_n\mathcal{D}_{1n}}+\frac{1}{2}\frac{Q_n(Y|X)}{m_n\mathcal{D}_{1n}}\to 0.
$$

as $n\to\infty$ for almost surely (X,Y). From the decomposition of the terms $q_{G^n_1}(Y|X)-q_{G^*_1}(Y|X)$ and $Q_n(Y|X)$ in Stage 1, we have

$$
\begin{aligned}
&\frac{1}{2}\Big[\sum_{j=1}^{k^*_2}\exp((\beta^*_{1j})^\top X+\beta^*_{0j})\Big]\cdot\frac{A_{n,2}(Y|X)+A_{n,1}(Y|X)+A_{n,0}(Y|X)}{m_n\mathcal{D}_{1n}}\\
&+\frac{1}{2}\frac{B_{n,1}(Y|X)+B_{n,2}(Y|X)-C_{n,1}(Y|X)-C_{n,2}(Y|X)+E_n(Y|X)}{m_n\mathcal{D}_{1n}}\to 0.
\end{aligned}\tag{19}
$$

Denote $F_{\rho,j}(Y|X):=F_\rho(Y|X;\beta^*_{1j},\eta^*_j,\nu^*_j)$ and $H_j(Y|X):=H(Y|X,\beta^*_{1j})$, we have

$$
\begin{aligned}
\lim_{n\to\infty}\frac{A_{n,0}(Y|X)}{m_n\mathcal{D}_{1n}}&=\sum_{j=1}^{k^*_1}s_{0,j}\pi(Y|h_1(X,\kappa^*_j),\tau^*_j),\\
\lim_{n\to\infty}\frac{A_{n,1}(Y|X)}{m_n\mathcal{D}_{1n}}&=\sum_{j\in[k^*_1]:|\mathcal{V}_{1,j}|=1}\Big[\sum_{u_1=1}^{d_1}s^{(u_1)}_{1,j}\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa^*_j)\frac{\partial\pi}{\partial h_1}(Y|h_1(X,\kappa^*_j),\tau^*_j)
\end{aligned}
$$

$$+\frac{1}{2}s_{2,j}\frac{\partial^2\pi}{\partial h_1^2}(Y|h_1(X,\kappa_j^*),\tau_j^*)\Big],$$

$$\lim_{n\to\infty}\frac{A_{n,2}(Y|X)}{m_n\mathcal{D}_{1n}}=\sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|>1}\Big[\Big(\sum_{u_1=1}^{d_1}s_{1,j}^{(u_1)}\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*)+\sum_{u_1,v_1=1}^{d_1}\frac{s_{3,j}^{(u_1v_1)}}{1+1_{\{u_1=v_1\}}}\frac{\partial^2 h_1}{\partial\kappa^{(u_1)}\partial\kappa^{(v_1)}}(X,\kappa_j^*)\Big)$$

$$\times\frac{\partial\pi}{\partial h_1}(Y|h_1(X,\kappa_j^*),\tau_j^*)+\Big(\frac{1}{2}s_{2,j}+\sum_{u_1,v_1=1}^{d_1}\frac{s_{3,j}^{(u_1v_1)}}{1+1_{\{u_1=v_1\}}}\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*)\frac{\partial h_1}{\partial\kappa^{(v_1)}}(X,\kappa_j^*)\Big)\frac{\partial^2\pi}{\partial h_1^2}(Y|h_1(X,\kappa_j^*),\tau_j^*)$$

$$+\Big(\frac{1}{2}\sum_{u_1=1}^{d_1}s_{5,j}^{(u_1)}\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*)\Big)\frac{\partial^3\pi}{\partial h_1^3}(Y|h_1(X,\kappa_j^*),\tau_j^*)+\frac{1}{8}s_{4,j}\frac{\partial^4\pi}{\partial h_1^4}(Y|h_1(X,\kappa_j^*),\tau_j^*)\Big],$$

and

$$\lim_{n\to\infty}\frac{B_{n,1}(Y|X)}{m_n\mathcal{D}_{1n}}=\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\Big[\sum_{u=1}^{d}t_{1,j}^{(u)}X^{(u)}F_{0,j}(Y|X)+\sum_{u_2=1}^{d_2}t_{2,j}^{(u_2)}\frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\eta_j^*)F_{1,j}(Y|X)+\frac{1}{2}t_{3,j}F_{2,j}(Y|X)\Big],$$

$$\lim_{n\to\infty}\frac{B_{n,2}(Y|X)}{m_n\mathcal{D}_{1n}}=\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}\Big[\Big(\sum_{u=1}^{d}t_{1,j}^{(u)}X^{(u)}+\sum_{u,v=1}^{d}t_{4,j}^{(uv)}X^{(u)}X^{(v)}\Big)F_{0,j}(Y|X)$$

$$+\Big(\sum_{u_2=1}^{d_2}t_{2,j}^{(u_2)}\frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\eta_j^*)+\sum_{u_2,v_2=1}^{d_2}t_{5,j}^{(u_2v_2)}\frac{\partial^2 h_2}{\partial\eta^{(u_2)}\partial\eta^{(v_2)}}(X,\eta_j^*)+\sum_{u=1}^{d}\sum_{v_2=1}^{d_2}t_{7,j}^{(uv_2)}X^{(u)}\frac{\partial h_2}{\partial\eta^{(v_2)}}(X,\eta_j^*)\Big)F_{1,j}(Y|X)$$

$$+\Big(\frac{1}{2}t_{3,j}+\sum_{u_2,v_2=1}^{d_2}t_{5,j}^{(u_2v_2)}\frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\eta_j^*)\frac{\partial h_2}{\partial\eta^{(v_2)}}(X,\eta_j^*)+\sum_{u=1}^{d}\frac{1}{2}t_{8,j}^{(u)}X^{(u)}\Big)F_{2,j}(Y|X)$$

$$+\Big(\sum_{u_2=1}^{d_2}\frac{1}{2}t_{9,j}^{(u_2)}\frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\eta_j^*)\Big)F_{3,j}(Y|X)+\frac{1}{8}t_{6,j}F_{4,j}(Y|X)\Big],$$

and

$$\lim_{n\to\infty}\frac{C_{n,1}(Y|X)}{m_n\mathcal{D}_{1n}}=\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\sum_{u=1}^{d}t_{1,j}^{(u)}X^{(u)}H_j(Y|X),$$

$$\lim_{n\to\infty}\frac{C_{n,2}(Y|X)}{m_n\mathcal{D}_{1n}}=\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}\Big(\sum_{u=1}^{d}t_{1,j}^{(u)}X^{(u)}+\sum_{u,v=1}^{d}t_{4,j}^{(uv)}X^{(u)}X^{(v)}\Big)H_j(Y|X),$$

$$\lim_{n\to\infty}\frac{E_n(Y|X)}{m_n\mathcal{D}_{1n}}=\sum_{j=1}^{k_2^*}t_{0,j}[F_{0,j}(Y|X)-H_j(Y|X)].$$

It is worth noting that for almost every X, the set

$$\Big\{\Big[\sum_{j=1}^{k_2^*}\exp((\beta_{1j}^*)^\top X+\beta_{0j}^*)\Big]\frac{\partial^\rho\pi}{\partial h_1^\rho}(Y|h_1(X,\kappa_j^*),\tau_j^*):0\leq\rho\leq 4,\ j\in[k_1^*]\Big\}$$

$$\cup\Big\{F_\rho(Y|X;\beta_{1j}^*,\eta_j^*,\nu_j^*),\ H(Y|X;\beta_{1j}^*):0\leq\rho\leq 4,\ j\in[k_2^*]\Big\}$$

is linearly independent w.r.t Y. Therefore, it follows that the coefficients of those terms in the limit in equation (19) become zero.

For $j \in [k_1^*]$, by looking at the coefficient of the term $\Big[\sum_{j=1}^{k_2^*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)\Big]\pi(Y|h_1(X,\kappa_j^*),\tau_j^*)$, we have $s_{0,j} = 0$.

For $j \in [k_1^*]$ such that $|\mathcal{V}_{1,j}| = 1$, by considering the coefficients of

- $\Big[\sum_{j=1}^{k_2^*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)\Big]\frac{\partial \pi}{\partial h_1}(Y|h_1(X,\kappa_j^*),\tau_j^*)$, we have $\sum_{u_1=1}^{d_1} s_{1,j}^{(u_1)} \frac{\partial h_1}{\partial \kappa^{(u_1)}}(X,\kappa_j^*) = 0$ for almost every X. Since the expert function h_1 is strongly identifiable, we get $s_{1,j}^{(u_1)} = 0$ for all $u_1 \in [d_1]$;
- $\Big[\sum_{j=1}^{k_2^*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)\Big]\frac{\partial^2 \pi}{\partial h_1^2}(Y|h_1(X,\kappa_j^*),\tau_j^*)$, we have $s_{2,j} = 0$.

For $j \in [k_1^*]$ such that $|\mathcal{V}_{1,j}| > 1$, by considering the coefficients of

- $\Big[\sum_{j=1}^{k_2^*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)\Big]\frac{\partial \pi}{\partial h_1}(Y|h_1(X,\kappa_j^*),\tau_j^*)$, we have

  $$\sum_{u_1=1}^{d_1} s_{1,j}\frac{\partial h_1}{\partial \kappa^{(u_1)}}(X,\kappa_j^*) + \sum_{u_1,v_1=1}^{d_1} \frac{s_{3,j}^{(u_1 v_1)}}{1 + 1_{\{u_1=v_1\}}}\frac{\partial^2 h_1}{\partial \kappa^{(u_1)}\partial \kappa^{(v_1)}}(X,\kappa_j^*) = 0,$$

  for almost every X. Since the expert function h_1 satisfies the strong identifiability condition, we get $s_{1,j}^{(u_1)} = s_{3,j}^{(u_1 v_1)} = 0$ for all $u_1, v_1 \in [d_1]$;
- $\Big[\sum_{j=1}^{k_2^*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)\Big]\frac{\partial^2 \pi}{\partial h_1^2}(Y|h_1(X,\kappa_j^*),\tau_j^*)$, we have

  $$\frac{1}{2}s_{2,j} + \sum_{u_1,v_1=1}^{d_1} \frac{s_{3,j}^{(u_1 v_1)}}{1 + 1_{\{u_1=v_1\}}}\frac{\partial h_1}{\partial \kappa^{(u_1)}}(X,\kappa_j^*)\frac{\partial h_1}{\partial \kappa^{(v_1)}}(X,\kappa_j^*) = 0,$$

  for almost every X. Since $s_{3,j}^{(u_1 v_1)} = 0$ for all $u_1, v_1 \in [d_1]$, we deduce $s_{2,j} = 0$;
- $\Big[\sum_{j=1}^{k_2^*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)\Big]\frac{\partial^3 \pi}{\partial h_1^3}(Y|h_1(X,\kappa_j^*),\tau_j^*)$, we have $\frac{1}{2}\sum_{u_1=1}^{d_1} s_{5,j}^{(u_1)}\frac{\partial h_1}{\partial \kappa^{(u_1)}}(X,\kappa_j^*) = 0$, for almost every X. As the expert function h_1 meets the strong identifiability condition, we get $s_{5,j}^{(u_1)} = 0$ for all $u_1 \in [d_1]$;
- $\Big[\sum_{j=1}^{k_2^*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)\Big]\frac{\partial^4 \pi}{\partial h_1^4}(Y|h_1(X,\kappa_j^*),\tau_j^*)$, we have $s_{4,j} = 0$.

For $j \in [k_2^*]$ such that $|\mathcal{V}_{2,j}| = 1$, by considering the coefficients of

- $F_{0,j}(Y|X)$, we have $t_{0,j} + \sum_{u=1}^{d} t_{1,j}^{(u)} X^{(u)} = 0$, for almost every X. Then, we deduce $t_{0,j} = t_{1,j}^{(u)} = 0$ for all $u \in [d]$;
- $F_{1,j}(Y|X)$, we have $\sum_{u_2=1}^{d_2} t_{2,j}^{(u_2)}\frac{\partial h_2}{\partial \eta^{(u_2)}}(X,\eta_j^*)$, for almost every X. As the expert function h_2 is strongly identifiable, we get $t_{2,j}^{(u_2)} = 0$ for all $u_2 \in [d_2]$;

- $F_{2,j}(Y|X)$, we have $t_{3,j} = 0$.

For $j \in [k_2^*]$ such that $|\mathcal{V}_{2,j}| > 1$, by considering the coefficients of

- $F_{0,j}(Y|X)$, we have $t_{0,j} + \sum_{u=1}^{d} t_{1,j}^{(u)} X^{(u)} + \sum_{u,v=1}^{d} t_{4,j}^{(uv)} X^{(u)} X^{(v)} = 0$, for almost surely X. Then, we get $t_{0,j} = t_{1,j}^{(u)} = t_{4,j}^{(uv)}$ for all $u, v \in [d]$.
- $F_{1,j}(Y|X)$, we have

$$\sum_{u_2=1}^{d_2} t_{2,j}^{(u_2)} \frac{\partial h_2}{\partial \eta^{(u_2)}}(X, \eta_j^*) + \sum_{u_2,v_2=1}^{d_2} t_{5,j}^{(u_2 v_2)} \frac{\partial^2 h_2}{\partial \eta^{(u_2)} \partial \eta^{(v_2)}}(X, \eta_j^*) + \sum_{u=1}^{d} \sum_{v_2=1}^{d_2} t_{7,j}^{(uv_2)} X^{(u)} \frac{\partial h_2}{\partial \eta^{(v_2)}}(X, \eta_j^*) = 0,$$

  for almost every X. As the expert function h_2 meets the strong identifiability condition, we get $t_{2,j}^{(u_2)} = t_{5,j}^{(u_2 v_2)} = t_{7,j}^{(uv_2)} = 0$ for all $u_2, v_2 \in [d_2]$ and $u \in [d]$;
- $F_{2,j}(Y|X)$, we have

$$\frac{1}{2} t_{3,j} + \sum_{u_2,v_2=1}^{d_2} t_{5,j}^{(u_2 v_2)} \frac{\partial h_2}{\partial \eta^{(u_2)}}(X, \eta_j^*) \frac{\partial h_2}{\partial \eta^{(v_2)}}(X, \eta_j^*) + \sum_{u=1}^{d} \frac{1}{2} t_{8,j}^{(u)} X^{(u)} = 0,$$

  for almost every X. Since $t_{5,j}^{(u_2 v_2)} = 0$ for all $u_2, v_2 \in [d_2]$, we deduce $\frac{1}{2} t_{3,j} + \sum_{u=1}^{d} \frac{1}{2} t_{8,j}^{(u)} X^{(u)} = 0$, for almost every X. Then, we get $t_{3,j} = t_{8,j}^{(u)} = 0$ for all $u_2, v_2 \in [d_2]$ and $u \in [d]$;
- $F_{3,j}(Y|X)$, we have $\sum_{u_2=1}^{d_2} \frac{1}{2} t_{9,j}^{(u_2)} \frac{\partial h_2}{\partial \eta^{(u_2)}}(X, \eta_j^*) = 0$, for almost every X. As the expert function h_2 is strongly identifiable, we get $t_{9,j}^{(u_2)}$ for all $u_2 \in [d_2]$;
- $F_{4,j}(Y|X)$, we have $t_{6,j} = 0$.

Putting the above results together, we have (i) $s_{0,j} = s_{1,j}^{(u_1)} = s_{2,j} = s_{3,j}^{(u_1 v_1)} = s_{4,j} = s_{5,j}^{(u_1)} = 0$ for all $j \in [k_1^*]$ and $u_1, v_1 \in [d_1]$; (ii) $t_{0,j} = t_{1,j}^{(u)} = t_{2,j}^{(u_2)} = t_{3,j} = t_{4,j}^{(uv)} = t_{5,j}^{(u_2 v_2)} = t_{6,j} = t_{7,j}^{uv_2} = t_{8,j}^{(u)} = t_{9,j}^{(u_2)} = 0$ for all $j \in [k_2^*]$, $u, v \in [d]$ and $u_2, v_2 \in [d_2]$. This contradicts to the fact that at least one among them is non-zero. Consequently, we achieve the local part in equation (15), that is,

$$\lim_{\varepsilon \to 0} \inf_{(G_1,G_2) \in \mathcal{G}_{k_1,k_2}(\Theta) : \mathcal{D}_1((G_1,G_2),(G_1^*,G_2^*)) \leq \varepsilon} \frac{\mathbb{E}_X[V(f_{G_1,G_2}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X))]}{\mathcal{D}_1((G_1,G_2),(G_1^*,G_2^*))} > 0.$$

The local part indicates that there exists a positive constant ε' such that

$$\inf_{(G_1,G_2) \in \mathcal{G}_{k_1,k_2}(\Theta) : \mathcal{D}_1((G_1,G_2),(G_1^*,G_2^*)) \leq \varepsilon} \frac{\mathbb{E}_X[V(f_{G_1,G_2}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X))]}{\mathcal{D}_1((G_1,G_2),(G_1^*,G_2^*))} > 0.$$

Proof for the global part (16)**:** Thus, it is sufficient to establish the global part

$$\inf_{(G_1,G_2) \in \mathcal{G}_{k_1,k_2}(\Theta) : \mathcal{D}_1((G_1,G_2),(G_1^*,G_2^*)) > \varepsilon'} \frac{\mathbb{E}_X[V(f_{G_1,G_2}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X))]}{\mathcal{D}_1((G_1,G_2),(G_1^*,G_2^*))} > 0.$$

Suppose that the global part does not hold, then there exists a sequence of mixing measure pairs $(\tilde{G}_1^n, \tilde{G}_2^n)$ satisfying $\mathcal{D}_1((\tilde{G}_1^n, \tilde{G}_2^n), (G_1^*, G_2^*)) > \varepsilon'$ and $\lim_{n\to\infty} \frac{\mathbb{E}_X[V(f_{\tilde{G}_1^n,\tilde{G}_2^n}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X))]}{\mathcal{D}_1((\tilde{G}_1^n,\tilde{G}_2^n),(G_1^*,G_2^*))} = 0$. In other words, we have

$$\lim_{n\to\infty} \mathbb{E}_X[V(f_{\tilde{G}_1^n,\tilde{G}_2^n}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X))] = 0.$$

Recall that the parameter space Θ is compact, then we can replace the sequence $(\tilde{G}_1^n, \tilde{G}_2^n)$ by one of its subsequences which converges to some pair of mixing measures $(\tilde{G}_1, \tilde{G}_2)$. Due to the fact that $\mathcal{D}_1((\tilde{G}_1^n, \tilde{G}_2^n), (G_1^*, G_2^*)) > \varepsilon'$, we get $\mathcal{D}_1((\tilde{G}_1, \tilde{G}_2), (G_1^*, G_2^*)) > \varepsilon'$. Next, by applying the Fatou's lemma, we have

$$\begin{aligned} 0 = \lim_{n\to\infty} \mathbb{E}_X[V(f_{\tilde{G}_1^n,\tilde{G}_2^n}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X))] &\geq \frac{1}{2}\int \liminf_{n\to\infty} \Big| f_{\tilde{G}_1^n,\tilde{G}_2^n}(Y|X), f_{G_1^*,G_2^*}(Y|X)\Big| \mathrm{d}(X,Y) \\ &= \frac{1}{2}\int \Big| f_{\tilde{G}_1,\tilde{G}_2}(Y|X) - f_{G_1^*,G_2^*}(Y|X)\Big| \mathrm{d}(X,Y). \end{aligned}$$

The above result implies that $f_{\tilde{G}_1,\tilde{G}_2}(Y|X) = f_{G_1^*,G_2^*}(Y|X)$ for almost surely (X,Y). According to Proposition 5, we deduce $(\tilde{G}_1, \tilde{G}_2) \equiv (G_1^*, G_2^*)$, indicating that $\mathcal{D}_1((\tilde{G}_1, \tilde{G}_2), (G_1^*, G_2^*)) = 0$. This contradicts the fact that $\mathcal{D}_1((\tilde{G}_1, \tilde{G}_2), (G_1^*, G_2^*)) > \varepsilon' > 0$. Hence, we obtain the global part (16) and complete the proof.

D.2 Proof of Theorem 2

By employing arguments used in Appendix D.1, it is sufficient to establish the local part

$$\lim_{\varepsilon\to 0} \inf_{(G_1,G_2)\in\mathcal{G}_{k_1,k_2}(\Theta):\mathcal{D}_2((G_1,G_2),(G_1^*,G_2^*))\leq\varepsilon} \frac{\mathbb{E}_X[V(f_{G_1,G_2}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X))]}{\mathcal{D}_2((G_1,G_2),(G_1^*,G_2^*))} > 0, \tag{20}$$

and the global part

$$\inf_{(G_1,G_2)\in\mathcal{G}_{k_1,k_2}(\Theta):\mathcal{D}_2((G_1,G_2),(G_1^*,G_2^*))>\varepsilon'} \frac{\mathbb{E}_X[V(f_{G_1,G_2}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X))]}{\mathcal{D}_2((G_1,G_2),(G_1^*,G_2^*))} > 0. \tag{21}$$

in this appendix. As the global part (21) can be derived similarly to Appendix D.1, we omit its proof here. Thus, we will focus on showing only the local part (20). Assume by contrary that the local part is not true. Then, there exists a sequence of mixing measure pairs (G_1^n, G_2^n) taking the form $G_1^n := \sum_{i=1}^{k_1^n} \omega_i^n \delta_{(\kappa_{1i}^n,\kappa_{0i}^n,\tau_i^n)}$, $G_2^n := \sum_{i=1}^{k_2^n} \exp(\beta_{0i}^n)\delta_{(\beta_{1i}^n,\eta_{1i}^n,\eta_{0i}^n,\nu_i^n)}$ for $n \in \mathbb{N}$ such that $\mathcal{D}_{2n} := \mathcal{D}_2((G_1^n, G_2^n), (G_1^*, G_2^*)) \to 0$ and

$$\mathbb{E}_X[V(f_{G_1^n,G_2^n}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X))]/\mathcal{D}_{2n} \to 0, \tag{22}$$

as $n \to \infty$. Here, we may assume WLOG that the number of shared experts and routed experts k_1^n, k_2^n and Voronoi cells $\mathcal{V}_{1,j} = \mathcal{V}_{1,j}(G_1^n)$, $\mathcal{V}_{2,j} = \mathcal{V}_{2,j}(G_2^n)$ do not change with the sample size n. Then, the Voronoi loss $\mathcal{D}_{2n}$ can be rewritten as

$$\mathcal{D}_{2n} = \sum_{j=1}^{k_1^*} \Big| \sum_{i\in\mathcal{V}_{1,j}} \omega_i^n - \omega_j^* \Big| + \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1} \Big| \sum_{i\in\mathcal{V}_{2,j}} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^*) \Big|$$

$$
\begin{aligned}
&+ \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|=1} \sum_{i\in\mathcal{V}_{1,j}} \omega_i^n(\|\Delta\kappa_{1ij}^n\| + |\Delta\kappa_{0ij}^n| + |\Delta\tau_{ij}^n|) \\
&+ \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|>1} \sum_{i\in\mathcal{V}_{1,j}} \omega_i^n(\|\Delta\kappa_{1ij}^n\|^2 + |\Delta\kappa_{0ij}^n|^{r_{1,j}} + |\Delta\tau_{ij}^n|^{r_{1,j}/2}) \\
&+ \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1} \sum_{i\in\mathcal{V}_{2,j}} \exp(\beta_{0i}^n)(\|\Delta\beta_{1ij}^n\| + \|\Delta\eta_{1ij}^n\| + |\Delta\eta_{0ij}^n| + |\Delta\nu_{ij}^n|) \\
&+ \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1} \sum_{i\in\mathcal{V}_{2,j}} \exp(\beta_{0i}^n)(\|\Delta\beta_{1ij}^n\|^{r_{2,j}} + \|\Delta\eta_{1ij}^n\|^{r_{2,j}/2} + |\Delta\eta_{0ij}^n|^{r_{2,j}} + |\Delta\nu_{ij}^n|^{r_{2,j}/2}), \qquad (23)
\end{aligned}
$$

where we denote $\Delta\kappa_{1ij}^n := \kappa_{1i}^n - \kappa_{1j}^*$, $\Delta\kappa_{0ij}^n := \kappa_{0i}^n - \kappa_{0j}^*$, $\Delta\eta_{1ij}^n := \eta_{1i}^n - \eta_{1j}^*$, and $\Delta\eta_{0ij}^n := \eta_{0i}^n - \eta_{0j}^*$. Since $\mathcal{D}_{2n} \to 0$ as $n \to \infty$, then the above formulation indicates that as $n \to \infty$, we have $\sum_{i\in\mathcal{V}_{1,j}} \omega_i^n \to \omega_j^*$, $(\kappa_{1i}^n, \kappa_{0i}^n, \tau_i^n) \to (\kappa_{1j}^*, \kappa_{0j}^*, \tau_j^*)$ as $n \to \infty$ for all $i \in \mathcal{V}_{1,j}$ and $j \in [k_1^*]$. Furthermore, we also have $\sum_{i\in\mathcal{V}_{2,j}} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^*)$, $(\beta_{1i}^n, \eta_{1i}^n, \eta_{0i}^n, \nu_i^n) \to (\beta_{1j}^*, \eta_{1j}^*, \eta_{0j}^*, \nu_j^*)$ as $n \to \infty$ for all $i \in \mathcal{V}_{2,j}$ and $j \in [k_2^*]$.

Next, we divide the rest of this proof into three main steps:

Stage 1 - Density Decomposition: In this stage, we aim to decompose the density discrepancy $f_{G_1^n,G_2^n}(Y|X) - f_{G_1^*,G_2^*}(Y|X)$. For ease of presentation, we denote

$$
\begin{aligned}
q_{G_1^n}(Y|X) &:= \sum_{i=1}^{k_1^n} \omega_i^n \pi(Y|(\kappa_{1i}^n)^\top X + \kappa_{0i}^n, \tau_i^n), \\
q_{G_1^*}(Y|X) &:= \sum_{i=1}^{k_1^*} \omega_i^* \pi(Y|(\kappa_{1j}^*)^\top X + \kappa_{0j}^*, \tau_i^*), \\
p_{G_2^n}(Y|X) &:= \sum_{i=1}^{k_2^n} \frac{\exp((\beta_{1i}^n)^\top X + \beta_{0i}^n)}{\sum_{j=1}^{k_2^n} \exp((\beta_{1j}^n)^\top X + \beta_{0j}^n)} \cdot \pi(Y|(\eta_{1i}^n)^\top X + \eta_{0i}^n, \nu_i^n), \\
p_{G_2^*}(Y|X) &:= \sum_{i=1}^{k_2^*} \frac{\exp((\beta_{1i}^*)^\top X + \beta_{0i}^*)}{\sum_{j=1}^{k_2^*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)} \cdot \pi(Y|(\eta_{1j}^*)^\top X + \eta_{0j}^*, \nu_i^*).
\end{aligned}
$$

Given the above notations, we get

$$
f_{G_1^n,G_2^n}(Y|X) - f_{G_1^*,G_2^*}(Y|X) = \frac{1}{2}\left[(q_{G_1^n}(Y|X) - q_{G_1^*}(Y|X)) + (p_{G_2^n}(Y|X) - p_{G_2^*}(Y|X))\right].
$$

Stage 1.1: Firstly, we decompose the term $q_{G_1^n}(Y|X) - q_{G_1^*}(Y|X)$ as

$$
\begin{aligned}
q_{G_1^n}(Y|X) - q_{G_1^*}(Y|X) &= \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|=1} \sum_{i\in\mathcal{V}_{1,j}} \omega_i^n[\pi(Y|(\kappa_{1i}^n)^\top X + \kappa_{0i}^n, \tau_i^n) - \pi(Y|(\kappa_{1j}^*)^\top X + \kappa_{0j}^*, \tau_j^*)] \\
&+ \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|>1} \sum_{i\in\mathcal{V}_{1,j}} \omega_i^n[\pi(Y|(\kappa_{1i}^n)^\top X + \kappa_{0i}^n, \tau_i^n) - \pi(Y|(\kappa_{1j}^*)^\top X + \kappa_{0j}^*, \tau_j^*)] \\
&+ \sum_{j=1}^{k_1^*} \Big(\sum_{i\in\mathcal{V}_{1,j}} \omega_i^n - \omega_j^*\Big)\pi(Y|(\kappa_{1j}^*)^\top X + \kappa_{0j}^*, \tau_j^*)
\end{aligned}
$$

$$:= A_{n,1}(Y|X) + A_{n,2}(Y|X) + A_{n,0}(Y|X).$$

By applying the first-order Taylor expansion to the function $\pi(Y|(\kappa^n_{1i})^\top X + \kappa^n_{0i}, \tau^n_i)$ around the point $(\kappa^*_{1j}, \kappa^*_{0j}, \tau^*_j)$, the term $A_{n,1}(Y|X)$ is rewritten as

$$\begin{aligned}
A_{n,1}(Y|X) &= \sum_{j\in[k^*_1]:|\mathcal{V}_{1,j}|=1} \sum_{i\in\mathcal{V}_{1,j}} \sum_{|\alpha|=1} \frac{\omega^n_i}{\alpha!}(\Delta\kappa^n_{1ij})^{\alpha_1}(\Delta\kappa^n_{0ij})^{\alpha_2}(\Delta\tau^n_{ij})^{\alpha_3} \\
&\qquad \times \frac{\partial^{|\alpha_1|+\alpha_2+\alpha_3}\pi}{\partial\kappa_1^{\alpha_1}\partial\kappa_2^{\alpha_2}\partial\tau^{\alpha_3}}(Y|(\kappa^*_{1j})^\top X + \kappa^*_{0j}, \tau^*_j) + R_{n,1}(Y|X) \\
&= \sum_{j\in[k^*_1]:|\mathcal{V}_{1,j}|=1} \sum_{i\in\mathcal{V}_{1,j}} \sum_{|\alpha|=1} \frac{\omega^n_i}{2^{\alpha_3}\alpha!}(\Delta\kappa^n_{1ij})^{\alpha_1}(\Delta\kappa^n_{0ij})^{\alpha_2}(\Delta\tau^n_{ij})^{\alpha_3} \\
&\qquad \times X^{\alpha_1}\frac{\partial^{|\alpha_1|+\alpha_2+2\alpha_3}\pi}{\partial h_1^{|\alpha_1|+\alpha_2+2\alpha_3}}(Y|(\kappa^*_{1j})^\top X + \kappa^*_{0j}, \tau^*_j) + R_{n,1}(Y|X) \\
&= \sum_{j\in[k^*_1]:|\mathcal{V}_{1,j}|=1} \sum_{|\alpha_1|=0}^{1} \sum_{\ell=1_{\{|\alpha_1|=0\}}}^{2(1-|\alpha_1|)} A^{(j)}_{n,\alpha_1,\ell}\cdot X^{\alpha_1}\frac{\partial^{|\alpha_1|+\ell}\pi}{\partial h_1^{|\alpha_1|+\ell}}(Y|(\kappa^*_{1j})^\top X + \kappa^*_{0j}, \tau^*_j) + R_{n,1}(Y|X),
\end{aligned}$$

where $R_{n,1}(Y|X)$ is a Taylor remainder such that $R_{n,1}(Y|X)/\mathcal{D}_{2n} \to 0$ as $n\to\infty$, and

$$A^{(j)}_{n,\alpha_1,\ell} := \sum_{i\in\mathcal{V}_{1,j}} \sum_{\alpha_2+2\alpha_3=\ell} \frac{\omega^n_i}{2^{\alpha_3}\alpha!}(\Delta\kappa^n_{1ij})^{\alpha_1}(\Delta\kappa^n_{0ij})^{\alpha_2}(\Delta\tau^n_{ij})^{\alpha_3},$$

for all $j\in[k^*_1]$, $\alpha_1\in\mathbb{N}^d$ and $\ell\in\mathbb{N}$ such that $(\alpha_1,\ell)\neq(0_d,0)$. Meanwhile, by applying the Taylor expansion of the order $r_{1,j} := r_1(|\mathcal{V}_{1,j}|)$ to the function $\pi(Y|(\kappa^n_{1i})^\top X + \kappa^n_{0i}, \tau^n_i)$ around the point $(\kappa^*_{1j}, \kappa^*_{0j}, \tau^*_j)$, we rewrite the term $A_{n,2}(Y|X)$ as

$$A_{n,2}(Y|X) = \sum_{j\in[k^*_1]:|\mathcal{V}_{1,j}|>1} \sum_{|\alpha_1|=0}^{r_{1,j}} \sum_{\ell=1_{\{|\alpha_1|=0\}}}^{2(r_{1,j}-|\alpha_1|)} A^{(j)}_{n,\alpha_1,\ell}\cdot X^{\alpha_1}\frac{\partial^{|\alpha_1|+\ell}\pi}{\partial h_1^{|\alpha_1|+\ell}}(Y|(\kappa^*_{1j})^\top X + \kappa^*_{0j}, \tau^*_j) + R_{n,2}(Y|X),$$

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

Stage 1.2: Next, we attempt to decompose the term $Q_n(Y|X) := \Big[\sum_{j=1}^{k^*_2}\exp((\beta^*_{1j})^\top X + \beta^*_{0j})\Big]\cdot [p_{G^n_2}(Y|X) - p_{G^*_2}(Y|X)]$. By denoting $F(Y|X;\beta_1,\eta_1,\eta_0,\nu) := \exp(\beta_1^\top X)\pi(Y|(\eta_1)^\top X + \eta_0, \nu)$ and $H(Y|X;\beta_1) := \exp(\beta_1^\top X)p_{G_2}(Y|X)$, we can represent $Q_n(Y|X)$ as

$$\begin{aligned}
Q_n(Y|X) &= \sum_{j=1}^{k^*_2}\sum_{i\in\mathcal{V}_{2,j}} \exp(\beta^n_{0i})[F(Y|X;\beta^n_{1i},\eta^n_{1i},\eta^n_{0i},\nu^n_i) - F(Y|X;\beta^*_{1j},\eta^*_{1j},\eta^*_{0j},\nu^*_j)] \\
&\quad - \sum_{j=1}^{k^*_2}\sum_{i\in\mathcal{V}_{2,j}} \exp(\beta^n_{0i})[H(Y|X;\beta^n_{1i}) - H(Y|X;\beta^*_{1j})] \\
&\quad + \sum_{j=1}^{k^*_2}\Big(\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i}) - \exp(\beta^*_{0j})\Big)[F(Y|X;\beta^*_{1j},\eta^*_{1j},\eta^*_{0j},\nu^*_j) - H(Y|X;\beta^*_{1j})]
\end{aligned}$$

$$:= B_n(Y|X) - C_n(Y|X) + E_n(Y|X).$$

Stage 1.2.1: In this step, we decompose the term $B_n(Y|X)$:

$$\begin{aligned}
B_n(Y|X) &= \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta_{0i}^n)[F(Y|X;\beta_{1i}^n,\eta_{1i}^n,\eta_{0i}^n,\nu_i^n) - F(Y|X;\beta_{1j}^*,\eta_{1j}^*,\eta_{0j}^*,\nu_j^*)]\\
&\quad + \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta_{0i}^n)[F(Y|X;\beta_{1i}^n,\eta_{1i}^n,\eta_{0i}^n,\nu_i^n) - F(Y|X;\beta_{1j}^*,\eta_{1j}^*,\eta_{0j}^*,\nu_j^*)]\\
&:= B_{n,1}(Y|X) + B_{n,2}(Y|X).
\end{aligned}$$

By applying the first-order Taylor expansion to the function $F(Y|X;\beta_{1i}^n,\eta_{1i}^n,\eta_{0i}^n,\nu_i^n)$ around the point $(\beta_{1j}^*,\eta_{1j}^*,\eta_{0j}^*,\nu_j^*)$, we have

$$\begin{aligned}
B_{n,1}(Y|X) &= \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta_{0i}^n)\sum_{|\alpha|=1}\frac{1}{\alpha!}(\Delta\beta_{1ij}^n)^{\alpha_1}(\Delta\eta_{1ij}^n)^{\alpha_2}(\Delta\eta_{0ij}^n)^{\alpha_3}(\Delta\nu_{ij}^n)^{\alpha_4}\\
&\qquad\qquad\times\frac{\partial^{|\alpha_1|+|\alpha_2|+\alpha_3+\alpha_4}F}{\partial\beta_1^{\alpha_1}\partial\eta_1^{\alpha_2}\partial\eta_0^{\alpha_3}\partial\nu^{\alpha_4}}(Y|X;\beta_{1j}^*,\eta_{1j}^*,\eta_{0j}^*,\nu_j^*) + R_{n,3}(Y|X)\\
&= \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\sum_{i\in\mathcal{V}_{2,j}}\sum_{|\alpha|=1}\frac{\exp(\beta_{0i}^n)}{2^{\alpha_4}\alpha!}(\Delta\beta_{1ij}^n)^{\alpha_1}(\Delta\eta_{1ij}^n)^{\alpha_2}(\Delta\eta_{0ij}^n)^{\alpha_3}(\Delta\nu_{ij}^n)^{\alpha_4}\\
&\qquad\qquad\times X^{\alpha_1+\alpha_2}\exp((\beta_{1j}^*)^\top X)\frac{\partial^{|\alpha_2|+\alpha_3+2\alpha_4}\pi}{\partial h_2^{|\alpha_2|+\alpha_3+2\alpha_4}}(Y|(\eta_{1j}^*)^\top X+\eta_{0j}^*,\nu_j^*) + R_{n,3}(Y|X)\\
&= \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\sum_{|\ell_1|+\ell_2=1}^{2}B_{n,\ell_1,\ell_2}^{(j)}\cdot X^{\ell_1}\exp((\beta_{1j}^*)^\top X)\frac{\partial^{\ell_2}\pi}{\partial h_2^{\ell_2}}(Y|(\eta_{1j}^*)^\top X+\eta_{0j}^*,\nu_j^*) + R_{n,3}(Y|X),
\end{aligned}$$

where $R_{n,3}(Y|X)$ is the Taylor remainder such that $R_{n,3}(Y|X)/\mathcal{D}_{2n}\to 0$, and

$$B_{n,\ell_1,\ell_2}^{(j)} := \sum_{i\in\mathcal{V}_{2,j}}\sum_{\alpha\in\mathcal{I}_{\ell_1,\ell_2}}\frac{\exp(\beta_{0i}^n)}{2^{\alpha_4}\alpha!}(\Delta\beta_{1ij}^n)^{\alpha_1}(\Delta\eta_{1ij}^n)^{\alpha_2}(\Delta\eta_{0ij}^n)^{\alpha_3}(\Delta\nu_{ij}^n)^{\alpha_4},$$

for all $j\in[k_2^*]$, $\ell_1\in\mathbb{N}^d$, and $\ell_2\in\mathbb{N}$ such that $(\ell_1,\ell_2)\neq(0_d,0)$, where we define

$$\mathcal{I}_{\ell_1,\ell_2} := \{\alpha=(\alpha_i)_{i=1}^4\in\mathbb{N}^d\times\mathbb{N}^d\times\mathbb{N}\times\mathbb{N} : \alpha_1+\alpha_2=\ell_1, \alpha_3+2\alpha_4=\ell_2-|\alpha_2|\}.$$

By applying the Taylor expansion of the order $r_{2,j} := r_2(|\mathcal{V}_{2,j}|)$ to the function $F(Y|X;\beta_{1i}^n,\eta_{1i}^n,\eta_{0i}^n,\nu_i^n)$ around the point $(\beta_{1j}^*,\eta_{1j}^*,\eta_{0j}^*,\nu_j^*)$, we have

$$B_{n,2}(Y|X) = \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\sum_{|\ell_1|+\ell_2=1}^{2r_{2,j}}B_{n,\ell_1,\ell_2}^{(j)}\cdot X^{\ell_1}\exp((\beta_{1j}^*)^\top X)\frac{\partial^{\ell_2}\pi}{\partial h_2^{\ell_2}}(Y|(\eta_{1j}^*)^\top X+\eta_{0j}^*,\nu_j^*) + R_{n,4}(Y|X),$$

where $R_{n,4}(Y|X)$ is the Taylor remainder such that $R_{n,4}(Y|X)/\mathcal{D}_{2n}\to 0$.

Stage 1.2.2: In this step, we decompose the term $C_n(Y|X)$:

$$C_n(Y|X) = \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta_{0i}^n)[H(Y|X;\beta_{1i}^n) - H(Y|X;\beta_{1j}^*)]$$

$$
\begin{aligned}
&+ \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1} \sum_{i\in\mathcal{V}_{2,j}} \exp(\beta_{0i}^n)[H(Y|X;\beta_{1i}^n) - H(Y|X;\beta_{1j}^*)] \\
&:= C_{n,1}(Y|X) + C_{n,2}(Y|X).
\end{aligned}
$$

By means of the first-order and second-order Taylor expansions to the function $H(Y|X;\beta_{1i}^n)$ around the point β_{1j}^*, the term $C_{n,1}(Y|X)$ can be represented as

$$
\begin{aligned}
C_{n,1}(Y|X) &= \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1} \sum_{i\in\mathcal{V}_{2,j}} \exp(\beta_{0i}^n) \sum_{|\gamma|=1} \frac{1}{\gamma!}(\Delta\beta_{1ij}^n)^{\gamma} \frac{\partial^{|\gamma|}H}{\partial\beta_1^{\gamma}}(Y|X;\beta_{1j}^*) + R_{n,5}(Y|X) \\
&= \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1} \sum_{i\in\mathcal{V}_{2,j}} \sum_{|\gamma|=1} \frac{\exp(\beta_{0i}^n)}{\gamma!}(\Delta\beta_{1ij}^n)^{\gamma} \cdot X^{\gamma} \exp((\beta_{1j}^*)^{\top}X) p_{G_2^n}(Y|X) + R_{n,5}(Y|X) \\
&= \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1} \sum_{|\gamma|=1} C_{n,\gamma}^{(j)} \cdot X^{\gamma} \exp((\beta_{1j}^*)^{\top}X) p_{G_2^n}(Y|X) + R_{n,5}(Y|X),
\end{aligned}
$$

where $R_{n,5}(Y|X)$ is the Taylor remainder such that $R_{n,5}(Y|X)/\mathcal{D}_{2n} \to 0$, and

$$
C_{n,\gamma}^{(j)} := \sum_{i\in\mathcal{V}_{2,j}} \frac{\exp(\beta_{0i}^n)}{\gamma!}(\Delta\beta_{1ij}^n)^{\gamma},
$$

for all $j \in [k_2^*]$ and $\gamma \in \mathbb{N}^d \setminus \{0_d\}$. Analogously, by applying the second-order Taylor expansion to the function $H(Y|X;\beta_{1i}^n)$ around the point β_{1j}^*, we represent the term $C_{n,2}(Y|X)$ as

$$
C_{n,2}(Y|X) = \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1} \sum_{|\gamma|=1}^{2} C_{n,\gamma}^{(j)} \cdot X^{\gamma} \exp((\beta_{1j}^*)^{\top}X) p_{G_2^n}(Y|X) + R_{n,6}(Y|X),
$$

where $R_{n,6}(Y|X)$ is the Taylor remainder such that $R_{n,6}(Y|X)/\mathcal{D}_{2n} \to 0$.

Combining the above decompositions of $A_n(Y|X)$, $B_n(Y|X)$, and $C_n(Y|X)$ together, we obtain

$$
\begin{aligned}
&\Big[\sum_{j=1}^{k_2^*} \exp((\beta_{1j}^*)^{\top}X + \beta_{0j}^*)\Big] \cdot [f_{G_1^n,G_2^n}(Y|X) - f_{G_1^*,G_2^*}(Y|X)] \\
&= \Big[\sum_{j=1}^{k_2^*} \exp((\beta_{1j}^*)^{\top}X + \beta_{0j}^*)\Big] \frac{1}{2} \sum_{j\in[k_1^*]} \sum_{|\alpha_1|=0}^{r_{1,j}} \sum_{\ell=0}^{2(r_{1,j}-|\alpha_1|)} A_{n,\alpha_1,\ell}^{(j)} \cdot X^{\alpha_1} \frac{\partial^{|\alpha_1|+\ell}\pi}{\partial h_1^{|\alpha_1|+\ell}}(Y|(\kappa_{1j}^*)^{\top}X + \kappa_{0j}^*, \tau_j^*) \\
&+ \frac{1}{2} \sum_{j\in[k_2^*]} \sum_{|\ell_1|+\ell_2=0}^{2r_{2,j}} B_{n,\ell_1,\ell_2}^{(j)} \cdot X^{\ell_1} \exp((\beta_{1j}^*)^{\top}X) \frac{\partial^{\ell_2}\pi}{\partial h_2^{\ell_2}}(Y|(\eta_{1j}^*)^{\top}X + \eta_{0j}^*, \nu_j^*) \\
&- \frac{1}{2} \sum_{j\in[k_2^*]} \sum_{|\gamma|=0}^{1+1_{\{|\mathcal{V}_{2,j}|>1\}}} C_{n,\gamma}^{(j)} \cdot X^{\gamma} \exp((\beta_{1j}^*)^{\top}X) p_{G_2^n}(Y|X) \\
&+ \frac{1}{2}\Big[\sum_{j=1}^{k_2^*} \exp((\beta_{1j}^*)^{\top}X + \beta_{0j}^*)\Big][R_{n,1}(Y|X) + R_{n,2}(Y|X)]
\end{aligned}
$$

$$+\frac{1}{2}[R_{n,3}(Y|X)+R_{n,4}(Y|X)-R_{n,5}(Y|X)-R_{n,6}(Y|X)], \tag{24}$$

with a convention that $r_{1,j}=1$ for $j\in[k_1^*]:|\mathcal{V}_{1,j}|=1$ and $r_{2,j}=1$ for $j\in[k_2^*]:|\mathcal{V}_{2,j}|$, where we define

$$A^{(j)}_{n,0_d,0}:=\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n-\omega_j^*,\qquad j\in[k_1^*]$$
$$B^{(j)}_{n,0_d,0}:=\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta_{0i}^n)-\exp(\beta_{0j}^*),\qquad j\in[k_2^*]$$
$$C^{(j)}_{n,0_d}:=\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta_{0i}^n)-\exp(\beta_{0j}^*),\qquad j\in[k_2^*].$$

Stage 2 - Non-vanishing coefficients: In this stage, we demonstrate that at least one among the terms $A^{(j)}_{n,\alpha_1,\ell}/\mathcal{D}_{2n}$, $B^{(j)}_{n,\ell_1,\ell_2}/\mathcal{D}_{2n}$, and $C^{(j)}_{n,\gamma}/\mathcal{D}_{2n}$ does not converge to zero as $n\to\infty$. In particular, we assume that all these terms go to zero. Then, by looking at the terms $A^{(j)}_{n,\alpha_1,\ell}$,

- For $j\in[k_1^*]$ and $|\alpha_1|=\ell=0$, we have $\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j=1}^{k_1^*}\Big|\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n-\omega_j^*\Big|\to 0$;
- For $j\in[k_1^*]:|\mathcal{V}_{1,j}|=1$, $\alpha_1\in\mathbb{N}^d:|\alpha_1|=1$ and $\ell=0$, we have
$$\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|=1}\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n\|\Delta\kappa_{1ij}^n\|\to 0;$$
- For $j\in[k_1^*]:|\mathcal{V}_{1,j}|=1$, $\alpha_1=0_d$ and $\ell=1$, we have
$$\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|=1}\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n|\Delta\kappa_{0ij}^n|\to 0;$$
- For $j\in[k_1^*]:|\mathcal{V}_{1,j}|=1$, $\alpha_1\in\mathbb{N}^d:|\alpha_1|=1$ and $\ell=2$ we have
$$\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|=1}\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n|\Delta\tau_{ij}^n|\to 0;$$
- For $j\in[k_1^*]:|\mathcal{V}_{1,j}|>1$, $\alpha_1=2e_u$, where $e_u\in\mathbb{N}^d$ is a one-hot vector with the u-th entry being one while other entries being zero, for $u\in[d]$, we have
$$\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|>1}\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n\|\Delta\kappa_{1ij}^n\|^2\to 0;$$

Next, by considering the terms $B^{(j)}_{n,\ell_1,\ell_2}$

- For $j\in[k_2^*]$ and $|\ell_1|=\ell_2=0$, we have $\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{j=1}^{k_2^*}\Big|\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta_{0i}^n)-\exp(\beta_{1j}^*)\Big|\to 0$;

- For $j \in [k_2^*] : |\mathcal{V}_{2,j}| = 1$, $\ell_1 = e_u$ for $u \in [d]$, and $\ell_2 = 0$, we have

$$\frac{1}{\mathcal{D}_{2n}} \cdot \sum_{j \in [k_2^*]:|\mathcal{V}_{2,j}|=1} \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}^n) \|\Delta \beta_{1ij}^n\| \to 0;$$

- For $j \in [k_2^*] : |\mathcal{V}_{2,j}| = 1$, $\ell_1 = e_u$ for $u \in [d]$, and $\ell_2 = 1$, we have

$$\frac{1}{\mathcal{D}_{2n}} \cdot \sum_{j \in [k_2^*]:|\mathcal{V}_{2,j}|=1} \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}^n) \|\Delta \eta_{1ij}^n\| \to 0;$$

- For $j \in [k_2^*] : |\mathcal{V}_{2,j}| = 1$, $\ell_1 = 0_d$ and $\ell_2 = 1$, we have

$$\frac{1}{\mathcal{D}_{2n}} \cdot \sum_{j \in [k_2^*]:|\mathcal{V}_{2,j}|=1} \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}^n) |\Delta \eta_{0ij}^n| \to 0;$$

- For $j \in [k_2^*] : |\mathcal{V}_{2,j}| = 1$, $\ell_1 = 0_d$, and $\ell_2 = 2$ we have

$$\frac{1}{\mathcal{D}_{2n}} \cdot \sum_{j \in [k_2^*]:|\mathcal{V}_{2,j}|=1} \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}^n) |\Delta \nu_{ij}^n| \to 0;$$

Taking the sum of the above limits, we deduce

$$\begin{aligned} \frac{1}{\mathcal{D}_{2n}} \cdot \Bigg[\sum_{j=1}^{k_1^*} \Big| \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n - \omega_j^* \Big| + \sum_{j \in [k_2^*]:|\mathcal{V}_{2,j}|>1} \Big| \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^*) \Big| \\ + \sum_{j \in [k_1^*]:|\mathcal{V}_{1,j}|=1} \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n (\|\Delta \kappa_{1ij}^n\| + |\Delta \kappa_{0ij}^n| + |\Delta \tau_{ij}^n|) + \sum_{j \in [k_1^*]:|\mathcal{V}_{1,j}|>1} \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n \|\Delta \kappa_{1ij}^n\|^2 \\ + \sum_{j \in [k_2^*]:|\mathcal{V}_{2,j}|=1} \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}^n) (\|\Delta \beta_{1ij}^n\| + \|\Delta \eta_{1ij}^n\| + |\Delta \eta_{0ij}^n| + |\Delta \nu_{ij}^n|) \Bigg] \to 0, \end{aligned}$$

as $n \to \infty$. From the formulation of the Voronoi loss $\mathcal{D}_{2n}$ in equation (23), it follows that

$$\begin{aligned} \frac{1}{\mathcal{D}_{2n}} \Bigg[\sum_{j \in [k_1^*]:|\mathcal{V}_{1,j}|>1} \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n (|\Delta \kappa_{0ij}^n|^{r_{1,j}} + |\Delta \tau_{ij}^n|^{r_{1,j}/2}) \\ + \sum_{j \in [k_2^*]:|\mathcal{V}_{2,j}|>1} \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}^n) (\|\Delta \beta_{1ij}^n\|^{r_{2,j}} + \|\Delta \eta_{1ij}^n\|^{r_{2,j}/2} + |\Delta \eta_{0ij}^n|^{r_{2,j}} + |\Delta \nu_{ij}^n|^{r_{2,j}/2}) \Bigg] \not\to 0, \end{aligned} \quad (25)$$

as $n \to \infty$. Then, we consider two following cases:

Case I: $\frac{1}{\mathcal{D}_{2n}} \sum_{j \in [k_1^*]:|\mathcal{V}_{1,j}|>1} \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n (|\Delta \kappa_{0ij}^n|^{r_{1,j}} + |\Delta \tau_{ij}^n|^{r_{1,j}/2}) \not\to 0$ as $n \to \infty$.

In this case, there exists some index $j' \in [k_1^*] : |\mathcal{V}_{1,j'}| > 1$ such that

$$\frac{1}{\mathcal{D}_{2n}} \cdot \sum_{i \in \mathcal{V}_{1,j'}} \omega_i^n (|\Delta \kappa_{0ij'}^n|^{r_{1,j'}} + |\Delta \tau_{ij'}^n|^{r_{1,j'}/2}) \not\to 0, \quad (26)$$

as $n \to \infty$. WLOG, we may assume that $j' = 1$. Recall that the term $A^{(j)}_{n,\alpha_1,\ell}/\mathcal{D}_{2n} \to 0$ as $n \to \infty$ for all $0 \leq |\alpha_1| \leq r_{1,j}$ and $0 \leq \ell \leq 2(r_{1,j} - |\alpha_1|)$. Then, by dividing the ratio $A^{(1)}_{n,0_d,\ell}$ by the left hand side of equation (26), we get

$$\frac{\sum_{i\in\mathcal{V}_{1,1}}\sum_{\alpha_2+2\alpha_3=\ell}\frac{\omega_i^n}{2^{\alpha_3}\alpha_2!\alpha_3!}(\Delta\kappa_{1i1})^{\alpha_2}(\Delta\tau_{i1})^{\alpha_3}}{\sum_{i\in\mathcal{V}_{1,1}}\omega_i^n(|\Delta\kappa^n_{0i1}|^{r_{1,1}}+|\Delta\tau^n_{i1}|^{r_{1,1}/2})} \to 0, \tag{27}$$

as $n \to \infty$ for all $0 \leq \ell \leq 2r_{1,1}$.

Let us denote $M_{n,1} := \max\{|\Delta\kappa^n_{0i1}|, |\Delta\tau^n_{i1}| : i \in \mathcal{V}_{1,1}\}$ and $W_{n,1} := \max\{\omega^n_i : i \in \mathcal{V}_{1,1}\}$. Since the sequence $(\omega^n_i/W_{n,1})_n$ is bounded below, we can replace it by its subsequence that admits the limit $s^2_{1i} := \lim_{n\to\infty}\omega^n_i/W_{n,1} > 0$. It should be noted that at least one among the terms s^2_{1i}, for $i \in \mathcal{V}_{1,1}$, is equal to 1. Next, we denote $(\Delta\kappa^n_{0i1})/M_{n,1} \to s_{2i}$ and $(\Delta\tau^n_{i1})/[2M^2_{n,1}] \to s_{3i}$ for all $i \in \mathcal{V}_{1,1}$. Similarly, at least one of each of the s_{2i} and s_{3i} is equal to 1 or -1. Then, by dividing both the numerators and the denominators of the left hand side of equation (27) by $W_{n,1}M^\ell_{n,1}$, we obtain the following system of polynomial equations:

$$\sum_{i\in\mathcal{V}_{1,1}}\sum_{\alpha_2+2\alpha_3=\ell}\frac{s^2_{1i}\ s^{\alpha_2}_{2i}\ s^{\alpha_3}_{3i}}{\alpha_2!\alpha_3!} = 0, \qquad 1 \leq \ell \leq r_{1,1}.$$

According to the definition of the term $r_{1,1}$, the above system does not admit any non-trivial solutions, which is a contradiction. Thus, Case I cannot occur.

Case II: $\frac{1}{\mathcal{D}_{2n}}\sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|>1}\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta^n_{0i})(\|\Delta\beta^n_{1ij}\|^{r_{2,j}}+\|\Delta\eta^n_{1ij}\|^{r_{2,j}/2}+|\Delta\eta^n_{0ij}|^{r_{2,j}}+|\Delta\nu^n_{ij}|^{r_{2,j}/2}) \not\to 0$ as $n \to \infty$.

In this case, we can find some index $j' \in [k^*_2] : |\mathcal{V}_{2,j'}| > 1$ such that

$$\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{i\in\mathcal{V}_{2,j'}}\exp(\beta^n_{0i})(\|\Delta\beta^n_{1ij'}\|^{r_{2,j'}}+\|\Delta\eta^n_{1ij'}\|^{r_{2,j'}/2}+|\Delta\eta^n_{0ij'}|^{r_{2,j'}}+|\Delta\nu^n_{ij'}|^{r_{2,j'}/2}) \not\to 0, \tag{28}$$

as $n \to \infty$. WLOG, we may assume that $j' = 1$. Recall that the term $B^{(j)}_{n,\ell_1,\ell_2}/\mathcal{D}_{2n} \to 0$ as $n \to \infty$ for all $j \in [k^*_2]$ and $(\ell_1, \ell_2) \in \mathbb{N}^d \times \mathbb{N} : 0 \leq |\ell_1| + \ell_2 \leq 2r_{2,j}$. Then, by dividing the ratio $B^{(1)}_{n,\ell_1,\ell_2}$ by the left hand side of equation (28), we get

$$\frac{\sum_{i\in\mathcal{V}_{2,1}}\sum_{\alpha\in\mathcal{I}_{\ell_1,\ell_2}}\frac{\exp(\beta^n_{0i})}{2^{\alpha_4}\alpha!}(\Delta\beta^n_{1i1})^{\alpha_1}(\Delta\eta^n_{1i1})^{\alpha_2}(\Delta\eta^n_{0i1})^{\alpha_3}(\Delta\nu^n_{i1})^{\alpha_4}}{\sum_{i\in\mathcal{V}_{2,1}}\exp(\beta^n_{0i})(\|\Delta\beta^n_{1i1}\|^{r_{2,1}}+\|\Delta\eta^n_{1i1}\|^{r_{2,1}/2}+|\Delta\eta^n_{0i1}|^{r_{2,1}}+|\Delta\nu^n_{i1}|^{r_{2,1}/2})} \to 0, \tag{29}$$

as $n \to \infty$ for all $(\ell_1, \ell_2) \in \mathbb{N}^d \times \mathbb{N} : 0 \leq |\ell_1| + \ell_2 \leq 2r_{2,1}$.

Let us denote $M_{n,2} := \max\{\|\Delta\beta^n_{1i1}\|, \|\Delta\eta^n_{1i1}\|, |\Delta\eta^n_{0i1}|, |\Delta\nu^n_{i1}| : i \in \mathcal{V}_{2,1}\}$ and $W_{n,2} := \max\{\exp(\beta^n_{0i}) : i \in \mathcal{V}_{2,1}\}$. Since the sequence $(\exp(\beta^n_{0i})/W_{n,2})_n$ is bounded below, we can replace it by its subsequence that admits the limit $t^2_{5i} := \lim_{n\to\infty}\exp(\beta^n_{0i})/W_{n,2} > 0$. It should be noted that at least one among the terms t^2_{5i}, for $i \in \mathcal{V}_{2,1}$, is equal to 1. Next, we denote

$$(\Delta\beta^n_{1i1})/M_{n,2} \to t_{1i}, \qquad (\Delta\eta^n_{1i1})/M^2_{n,2} \to t_{2i},$$

$$(\Delta\eta_{0i1}^{n})/M_{n,2} \to t_{3i}, \qquad (\Delta\nu_{i1}^{n})/[2M_{n,2}^{2}] \to t_{4i},$$

for all $i \in \mathcal{V}_{2,1}$. Similarly, at least one of each of the t_{1i}, t_{2i}, t_{3i}, and t_{4i}, is equal to 1 or -1. Then, by dividing both the numerators and the denominators of the left hand side of equation (29) by $W_{n,2}M_{n,2}^{|\ell_1|+\ell_2}$, we obtain the following system of polynomial equations:

$$\sum_{i\in\mathcal{V}_{2,1}}\sum_{\alpha\in\mathcal{I}_{\ell_1,\ell_2}}\frac{1}{\alpha!}\cdot t_{5i}^{2}\ t_{1i}^{\alpha_1}\ t_{2i}^{\alpha_2}\ t_{3i}^{\alpha_3}\ t_{4i}^{\alpha_4}=0, \qquad 1\leq|\ell_1|+\ell_2\leq r_{2,1}.$$

According to the definition of the term $r_{2,1}$, the above system does not admit any non-trivial solutions, which is a contradiction. Thus, Case II cannot occur.

The fact that both Case I and Case II cannot occur contradicts the result of equation (25). Thus, not all the terms $A_{n,\alpha_1,\ell}^{(j)}/\mathcal{D}_{2n}$, $B_{n,\ell_1,\ell_2}^{(j)}/\mathcal{D}_{2n}$, and $C_{n,\gamma}^{(j)}/\mathcal{D}_{2n}$ converge to zero as $n\to\infty$.

Stage 3 - Fatou's lemma contradiction: We denote by m_n the maximum of the absolute values of the ratios $A_{n,\alpha_1,\ell}^{(j)}/\mathcal{D}_{2n}$, $B_{n,\ell_1,\ell_2}^{(j)}/\mathcal{D}_{2n}$, and $C_{n,\gamma}^{(j)}/\mathcal{D}_{2n}$. It follows from the result of Stage that $1/m_n \not\to \infty$ as $n\to\infty$. Then, by means of the Fatou's lemma, we have

$$\lim_{n\to\infty}\frac{\mathbb{E}_X[V(f_{G_1^n,G_2^n}(\cdot|X),f_{G_1^*,G_2^*}(\cdot|X))]}{m_n\mathcal{D}_{2n}}\geq\int\liminf_{n\to\infty}\frac{|f_{G_1^n,G_2^n}(Y|X)-f_{G_1^*,G_2^*}(Y|X)|}{2m_n\mathcal{D}_{2n}}\mathrm{d}(X,Y).$$

Then, we deduce $[f_{G_1^n,G_2^n}(Y|X)-f_{G_1^*,G_2^*}(Y|X)]/[m_n\mathcal{D}_{1n}]\to 0$ as $n\to\infty$ for almost surely (X,Y). Since the input space is bounded and the parameter space is compact, the quantity $\sum_{j=1}^{k_2^*}\exp((\beta_{1j}^*)^\top X+\beta_{0j}^*)$ is bounded. Thus, we also have

$$\frac{1}{m_n\mathcal{D}_{2n}}\Big[\sum_{j=1}^{k_2^*}\exp((\beta_{1j}^*)^\top X+\beta_{0j}^*)\Big][f_{G_1^n,G_2^n}(Y|X)-f_{G_1^*,G_2^*}(Y|X)]\to 0,$$

as $n\to\infty$ for almost surely (X,Y). Let us denote

$$\begin{aligned}\frac{1}{m_n\mathcal{D}_{2n}}A_{n,\alpha_1,\ell}^{(j)}&\to a_{\alpha_1,\ell}^{(j)},\\ \frac{1}{m_n\mathcal{D}_{2n}}B_{n,\ell_1,\ell_2}^{(j)}&\to b_{\ell_1,\ell_2}^{(j)},\\ \frac{1}{m_n\mathcal{D}_{2n}}C_{n,\gamma}^{(j)}&\to c_{\gamma}^{(j)},\end{aligned}$$

as $n\to\infty$ with a note that at least one among them is non-zero. From equation (24), we deduce

$$\begin{aligned}&\Big[\sum_{j=1}^{k_2^*}\exp((\beta_{1j}^*)^\top X+\beta_{0j}^*)\Big]\frac{1}{2}\sum_{j\in[k_1^*]}\sum_{|\alpha_1|=0}^{r_{1,j}}\sum_{\ell=0}^{2(r_{1,j}-|\alpha_1|)}a_{\alpha_1,\ell}^{(j)}\cdot X^{\alpha_1}\frac{\partial^{|\alpha_1|+\ell}\pi}{\partial h_1^{|\alpha_1|+\ell}}(Y|(\kappa_{1j}^*)^\top X+\kappa_{0j}^*,\tau_j^*)\\ &+\frac{1}{2}\sum_{j\in[k_2^*]}\sum_{|\ell_1|+\ell_2=0}^{2r_{2,j}}b_{\ell_1,\ell_2}^{(j)}\cdot X^{\ell_1}\exp((\beta_{1j}^*)^\top X)\frac{\partial^{\ell_2}\pi}{\partial h_2^{\ell_2}}(Y|(\eta_{1j}^*)^\top X+\eta_{0j}^*,\nu_j^*)\end{aligned}$$

$$- \frac{1}{2} \sum_{j \in [k_2^*]} \sum_{|\gamma|=0}^{1+1_{\{|\mathcal{V}_{2,j}|>1\}}} c_{\gamma}^{(j)} \cdot X^{\gamma} \exp((\beta_{1j}^*)^{\top} X) p_{G_2^*}(Y|X) \Big] \to 0,$$

as $n \to \infty$ for almost surely (X, Y). Since the set

$$\left\{ X^{\alpha_1} \frac{\partial^{|\alpha_1|+\ell} \pi}{\partial h_1^{|\alpha_1|+\ell}} (Y|(\kappa_{1j}^*)^{\top} X + \kappa_{0j}^*, \tau_j^*) : j \in [k_1^*], 0 \leq |\alpha_1| \leq r_{1,j}, 0 \leq \ell \leq 2(r_{1,j} - |\alpha_1|) \right\}$$

$$\cup \left\{ X^{\ell_1} \exp((\beta_{1j}^*)^{\top} X) \frac{\partial^{\ell_2} \pi}{\partial h_2^{\ell_2}} (Y|(\eta_{1j}^*)^{\top} X + \eta_{0j}^*, \nu_j^*), \ X^{\gamma} \exp((\beta_{1j}^*)^{\top} X) p_{G_2^*}(Y|X) : \right.$$

$$\left. j \in [k_2^*], 0 \leq |\ell_1| + \ell_2 \leq 2 r_{2,j}, 0 \leq |\gamma| \leq 2 \right\}$$

is linearly independent w.r.t ..., we obtain $a_{\alpha_1,\ell}^{(j)}$ for all $j \in [k_1^*]$, $\alpha_1 \in \mathbb{N}^d$, $\ell \in \mathbb{N}$, and $b_{\ell_1,\ell_2}^{(j)} = c_{\gamma}^{(j)} = 0$ for all $j \in [k_2^*]$, $(\ell_1, \ell_2) \in \mathbb{N}^d \times \mathbb{N}$, $\gamma \in \mathbb{N}^d$. This result contradicts the fact that not all the terms $a_{\alpha_1,\ell}^{(j)}$, $b_{\ell_1,\ell_2}^{(j)}$, and $c_{\gamma}^{(j)}$ equal zero. Hence, we achieve the local part in equation (20) and complete the proof.

D.3 Proof of Theorem 3

By leveraging the proof framework in Appendix D.1, we also focus on demonstrating the local part

$$\lim_{\varepsilon \to 0} \inf_{(G_1,G_2) \in \mathcal{G}_{k_1,k_2}(\Theta) : \mathcal{D}_3((G_1,G_2),(G_1^*,G_2^*)) \leq \varepsilon} \frac{\mathbb{E}_X[V(g_{G_1,G_2}(\cdot|X), g_{G_1^*,G_2^*}(\cdot|X))]}{\mathcal{D}_3((G_1,G_2),(G_1^*,G_2^*))} > 0, \tag{30}$$

and the global part

$$\inf_{(G_1,G_2) \in \mathcal{G}_{k_1,k_2}(\Theta) : \mathcal{D}_3((G_1,G_2),(G_1^*,G_2^*)) > \varepsilon'} \frac{\mathbb{E}_X[V(g_{G_1,G_2}(\cdot|X), g_{G_1^*,G_2^*}(\cdot|X))]}{\mathcal{D}_3((G_1,G_2),(G_1^*,G_2^*))} > 0. \tag{31}$$

in this appendix. Note that since the global part (31) can be argued in a similar fashion to Appendix D.1, its derivation is omitted here. Therefore, it is sufficient to establish the local part (30). Suppose that the local part does not hold. Then, there exists a sequence of mixing measure pairs (G_1^n, G_2^n) taking the form $G_1^n := \sum_{i=1}^{k_1^n} \omega_i^n \delta_{(\kappa_i^n, \tau_i^n)}$, $G_2^n := \sum_{i=1}^{k_2^n} \sigma(\beta_{0i}^n) \delta_{(\beta_{1i}^n, \eta_i^n, \nu_i^n)}$ for $n \in \mathbb{N}$ such that $\mathcal{D}_{3n} := \mathcal{D}_3((G_1^n, G_2^n), (G_1^*, G_2^*)) \to 0$ and

$$\mathbb{E}_X[V(g_{G_1^n,G_2^n}(\cdot|X), g_{G_1^*,G_2^*}(\cdot|X))]/\mathcal{D}_{3n} \to 0, \tag{32}$$

as $n \to \infty$. Here, we may assume WLOG that the number of shared experts and routed experts k_1^n, k_2^n and Voronoi cells $\mathcal{V}_{1,j} = \mathcal{V}_{1,j}(G_1^n)$, $\mathcal{V}_{2,j} = \mathcal{V}_{2,j}(G_2^n)$ do not change with the sample size n. Then, the Voronoi loss $\mathcal{D}_{3n}$ can be rewritten as

$$\mathcal{D}_{3n} = \sum_{j=1}^{k_1^*} \Big| \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n - \omega_j^* \Big| + \sum_{j \in [k_2^*] : |\mathcal{V}_{2,j}| > 1} \Big| \sum_{i \in \mathcal{V}_{2,j}} \sigma(\beta_{0i}^n) - \sigma(\beta_{0j}^*) \Big|$$

$$+ \sum_{j \in [k_1^*] : |\mathcal{V}_{1,j}| = 1} \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n (\|\Delta \kappa_{ij}^n\| + |\Delta \tau_{ij}^n|) + \sum_{j \in [k_2^*] : |\mathcal{V}_{2,j}| = 1} \sum_{i \in \mathcal{V}_{2,j}} (\|\Delta \beta_{1ij}^n\| + |\Delta \beta_{0ij}^n| + \|\Delta \eta_{ij}^n\| + |\Delta \nu_{ij}^n|)$$

$$+\sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|>1}\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n(\|\Delta\kappa_{ij}^n\|^2+|\Delta\tau_{ij}^n|^2)+\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}\sum_{i\in\mathcal{V}_{2,j}}(\|\Delta\beta_{1ij}^n\|^2+\|\Delta\eta_{ij}^n\|^2+|\Delta\nu_{ij}^n|^2), \tag{33}$$

where we denote $\Delta\beta_{0ij}^n:=\beta_{0i}^n-\beta_{0j}^*$. Since $\mathcal{D}_{3n}\to 0$ as $n\to\infty$, then the above formulation indicates that as $n\to\infty$, we have

- For $j\in[k_1^*]$ and $i\in\mathcal{V}_{1,j}$: $\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n\to\omega_j^*$, $(\kappa_i^n,\tau_i^n)\to(\kappa_j^*,\tau_j^*)$;
- For $j\in[k_2^*]:|\mathcal{V}_{2,j}|=1$ and $i\in\mathcal{V}_{2,j}$: $(\beta_{1i}^n,\beta_{0i}^n,\eta_i^n,\nu_i^n)\to(\beta_{1j}^*,\beta_{0j}^*,\eta_j^*,\nu_j^*)$;
- For $j\in[k_2^*]:|\mathcal{V}_{2,j}|>1$ and $i\in\mathcal{V}_{2,j}$: $\sum_{i\in\mathcal{V}_{2,j}}\sigma(\beta_{0i}^n)-\sigma(\beta_{0j}^*)$, $(\beta_{1i}^n,\eta_i^n,\nu_i^n)\to(\beta_{1j}^*,\eta_j^*,\nu_j^*)$.

Now, we divide the proof into three main stages:

Stage 1 - Density Decomposition: In this stage, we aim to decompose the density discrepancy $g_{G_1^n,G_2^n}(Y|X)-g_{G_1^*,G_2^*}(Y|X)$. For ease of presentation, we denote

$$\begin{aligned}
q_{G_1^n}(Y|X)&:=\sum_{i=1}^{k_1^n}\omega_i^n\pi(Y|h_1(X,\kappa_i^n),\tau_i^n),\\
q_{G_1^*}(Y|X)&:=\sum_{i=1}^{k_1^*}\omega_i^*\pi(Y|h_1(X,\kappa_i^*),\tau_i^*),\\
p_{G_2^n}(Y|X)&:=\sum_{i=1}^{k_2^n}\frac{\sigma((\beta_{1i}^n)^\top X+\beta_{0i}^n)}{\sum_{j=1}^{k_2^n}\sigma((\beta_{1j}^n)^\top X+\beta_{0j}^n)}\cdot\pi(Y|h_2(X,\eta_i^n),\nu_i^n),\\
p_{G_2^*}(Y|X)&:=\sum_{i=1}^{k_2^*}\frac{\sigma((\beta_{1i}^*)^\top X+\beta_{0i}^*)}{\sum_{j=1}^{k_2^*}\sigma((\beta_{1j}^*)^\top X+\beta_{0j}^*)}\cdot\pi(Y|h_2(X,\eta_i^*),\nu_i^*).
\end{aligned}$$

Given the above notations, we get

$$g_{G_1^n,G_2^n}(Y|X)-g_{G_1^*,G_2^*}(Y|X)=\frac{1}{2}\left[(q_{G_1^n}(Y|X)-q_{G_1^*}(Y|X))+(p_{G_2^n}(Y|X)-p_{G_2^*}(Y|X))\right].$$

Stage 1.1: Firstly, we decompose the term $q_{G_1^n}(Y|X)-q_{G_1^*}(Y|X)$ as

$$\begin{aligned}
q_{G_1^n}(Y|X)-q_{G_1^*}(Y|X)&=\sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|=1}\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n[\pi(Y|h_1(X,\kappa_i^n),\tau_i^n)-\pi(Y|h_1(X,\kappa_j^*),\tau_j^*)]\\
&\quad+\sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|>1}\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n[\pi(Y|h_1(X,\kappa_i^n),\tau_i^n)-\pi(Y|h_1(X,\kappa_j^*),\tau_j^*)]\\
&\quad+\sum_{j=1}^{k_1^*}\Big(\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n-\omega_j^*\Big)\pi(Y|h_1(X,\kappa_j^*),\tau_j^*)\\
&:=A_{n,1}(Y|X)+A_{n,2}(Y|X)+A_{n,0}(Y|X).
\end{aligned}$$

By using the same arguments as in Stage 1.1 in Appendix D.1, the term $A_{n,1}(Y|X)$ is rewritten as

$$A_{n,1}(Y|X) = \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|=1}\sum_{\rho=1}^{2} A^{(j)}_{n,1,\rho}(X)\frac{\partial^{\rho}\pi}{\partial h_1^{\rho}}(Y|h_1(X,\kappa_j^*),\tau_j^*) + R_{n,1}(Y|X),$$

where $R_{n,1}(Y|X)$ is a Taylor remainder such that $R_{n,1}(Y|X)/\mathcal{D}_{3n} \to$ as $n\to\infty$, and

$$\begin{aligned}
A^{(j)}_{n,1,1}(X) &:= \sum_{i\in\mathcal{V}_{1,j}}\omega_i^n\sum_{u_1=1}^{d_1}(\Delta\kappa_{ij}^n)^{(u_1)}\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*),\\
A^{(j)}_{n,1,2}(X) &:= \sum_{i\in\mathcal{V}_{1,j}}\omega_i^n\frac{1}{2}(\Delta\tau_{ij}^n),
\end{aligned}$$

for all $j\in[k_1^*]$ such that $|\mathcal{V}_{1,j}|=1$. Meanwhile, we can represent $A_{n,2}(Y|X)$ as

$$A_{n,2}(Y|X) = \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|>1}\sum_{\rho=1}^{4} A^{(j)}_{n,1,\rho}(X)\frac{\partial^{\rho}\pi}{\partial h_1^{\rho}}(Y|h_1(X,\kappa_j^*),\tau_j^*) + R_{n,2}(Y|X),$$

where $R_{n,2}(Y|X)$ is a Taylor remainder such that $R_{n,2}(Y|X)/\mathcal{D}_{3n} \to$ as $n\to\infty$, and

$$\begin{aligned}
A^{(j)}_{n,2,1}(X) &:= \sum_{i\in\mathcal{V}_{1,j}}\omega_i^n\Big(\sum_{u_1=1}^{d_1}(\Delta\kappa_{ij}^n)^{(u_1)}\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*) + \sum_{u_1,v_1=1}^{d_1}\frac{(\Delta\kappa_{ij}^n)^{(u_1)}(\Delta\kappa_{ij}^n)^{(v_1)}}{1+1_{\{u_1=v_1\}}}\frac{\partial^2 h_1}{\partial\kappa^{(u_1)}\partial\kappa^{(v_1)}}(X,\kappa_j^*)\Big),\\
A^{(j)}_{n,2,2}(X) &:= \sum_{i\in\mathcal{V}_{1,j}}\omega_i^n\Big(\frac{1}{2}(\Delta\tau_{ij}^n) + \sum_{u_1,v_1=1}^{d_1}\frac{(\Delta\kappa_{ij}^n)^{(u_1)}(\Delta\kappa_{ij}^n)^{(v_1)}}{1+1_{\{u_1=v_1\}}}\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*)\frac{\partial h_1}{\partial\kappa^{(v_1)}}(X,\kappa_j^*)\Big),\\
A^{(j)}_{n,2,3}(X) &:= \sum_{i\in\mathcal{V}_{1,j}}\omega_i^n\sum_{u_1=1}^{d_1}\frac{1}{2}(\Delta\kappa_{ij}^n)^{(u_1)}(\Delta\tau_{ij}^n)\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*),\\
A^{(j)}_{n,2,4}(X) &:= \sum_{i\in\mathcal{V}_{1,j}}\omega_i^n\frac{1}{8}(\Delta\tau_{ij}^n)^2,
\end{aligned}$$

for all $j\in[k_1^*]$ such that $|\mathcal{V}_{1,j}|>1$.

Stage 1.2: Next, we attempt to decompose the term $Q_n(Y|X) := \Big[\sum_{j=1}^{k_2^*}\sigma((\beta_{1j}^*)^\top X+\beta_{0j}^*)\Big]\cdot[p_{G_2^n}(Y|X)-p_{G_2^*}(Y|X)]$ as

$$\begin{aligned}
Q_n(Y|X) &= \sum_{j=1}^{k_2^*}\Big[\sum_{i\in\mathcal{V}_{2,j}}\sigma((\beta_{1i}^n)^\top X+\beta_{0i}^n)\pi(Y|h_2(X,\eta_i^n),\nu_i^n) - \sigma((\beta_{1j}^*)^\top X+\beta_{0j}^*)\pi(Y|h_2(X,\eta_j^*),\nu_j^*)\Big]\\
&\quad-\sum_{j=1}^{k_2^*}\Big[\sum_{i\in\mathcal{V}_{2,j}}\sigma((\beta_{1i}^n)^\top X+\beta_{0i}^n)-\sigma((\beta_{1j}^*)^\top X+\beta_{0j}^*)\Big]p_{G_2^n}(Y|X)\\
&:= B_n(Y|X) - C_n(Y|X).
\end{aligned}$$

Stage 1.2.1: In this step, we decompose the term $B_n(Y|X)$ with a note that $\beta^*_{1j} = 0_d$ for all $j \in [k^*_2] : |\mathcal{V}_{2,j}| > 1$:

$$\begin{aligned}
B_n(Y|X) = &\sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|=1} \Big[\sum_{i\in\mathcal{V}_{2,j}} \sigma((\beta^n_{1i})^\top X + \beta^n_{0i})\pi(Y|h_2(X,\eta^n_i),\nu^n_i) - \sigma((\beta^*_{1j})^\top X + \beta^*_{0j})\pi(Y|h_2(X,\eta^*_j),\nu^*_j)\Big] \\
&+ \sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|>1} \Big[\sum_{i\in\mathcal{V}_{2,j}} \sigma((\beta^n_{1i})^\top X + \beta^n_{0i})\pi(Y|h_2(X,\eta^n_i),\nu^n_i) - \sigma(\beta^*_{0j})\pi(Y|h_2(X,\eta^*_j),\nu^*_j)\Big] \\
&= \sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|=1} \Big[\sum_{i\in\mathcal{V}_{2,j}} \sigma((\beta^n_{1i})^\top X + \beta^n_{0i})\pi(Y|h_2(X,\eta^n_i),\nu^n_i) - \sigma((\beta^*_{1j})^\top X + \beta^*_{0j})\pi(Y|h_2(X,\eta^*_j),\nu^*_j)\Big] \\
&+ \sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|>1} \sum_{i\in\mathcal{V}_{2,j}} \Big[\sigma((\beta^n_{1i})^\top X + \beta^n_{0i})\pi(Y|h_2(X,\eta^n_i),\nu^n_i) - \sigma(\beta^n_{0i})\pi(Y|h_2(X,\eta^*_j),\nu^*_j)\Big] \\
&+ \sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|>1} \Big[\sum_{i\in\mathcal{V}_{2,j}} \sigma(\beta^n_{0i}) - \sigma(\beta^*_{0j})\Big]\pi(Y|h_2(X,\eta^*_j),\nu^*_j) \\
&:= B_{n,1}(Y|X) + B_{n,2}(Y|X) + B_{n,0}(Y|X).
\end{aligned}$$

Denote $\psi(X;\beta_1,\beta_0) := \sigma(\beta_1^\top X + \beta_0)$. By applying the first-order Taylor expansion to the function $\psi(X,\beta^n_{1i},\beta^n_{0i})\pi(Y|h_2(X,\eta^n_i),\nu^n_i)$ around the point $(\beta^*_{1j},\beta^*_{0j},\eta^n_i,\nu^n_i)$, we have

$$\begin{aligned}
B_{n,1}(Y|X) = &\sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|=1} \sum_{i\in\mathcal{V}_{2,j}} \sum_{|\alpha|=1} \frac{1}{\alpha!}(\Delta\beta^n_{1ij})^{\alpha_1}(\Delta\beta^n_{0ij})^{\alpha_2}(\Delta\eta^n_{ij})^{\alpha_3}(\Delta\nu^n_{ij})^{\alpha_4} \\
&\qquad\qquad \times \frac{\partial^{|\alpha_1|+\alpha_2}\psi}{\partial\beta_1^{\alpha_1}\partial\beta_0^{\alpha_2}}(X;\beta^*_{1j},\beta^*_{0j})\frac{\partial^{|\alpha_3|+\alpha_4}\pi}{\partial\eta^{\alpha_3}\partial\nu^{\alpha_4}}(Y|h_2(X,\eta^*_j),\nu^*_j) + R_{n,3}(Y|X) \\
= &\sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|=1} \sum_{\rho=0}^{2} B^{(j)}_{n,1,\rho}(X)\cdot\frac{\partial^\rho\pi}{\partial h_2^\rho}(Y|h_2(X,\eta^*_j),\nu^*_j) + R_{n,3}(Y|X),
\end{aligned}$$

where $R_{n,3}(Y|X)$ is a Taylor remainder such that $R_{n,3}(Y|X)/\mathcal{D}_{3n} \to 0$ as $n \to \infty$ and

$$\begin{aligned}
B^{(j)}_{n,1,0}(X) &:= \sum_{i\in\mathcal{V}_{2,j}} \Big[\sum_{u=1}^{d}(\Delta\beta^n_{1ij})^{(u)}\frac{\partial\psi}{\partial\beta_1^{(u)}}(X;\beta^*_{1j},\beta^*_{0j}) + (\Delta\beta^n_{0ij})\frac{\partial\psi}{\partial\beta_0}(X;\beta^*_{1j},\beta^*_{0j})\Big], \\
B^{(j)}_{n,1,1}(X) &:= \sum_{i\in\mathcal{V}_{2,j}} \sum_{u_2=1}^{d_2}(\Delta\eta^n_{ij})^{(u_2)}\frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\eta^*_j)\psi(X;\beta^*_{1j},\beta^*_{0j}), \\
B^{(j)}_{n,1,2}(X) &:= \sum_{i\in\mathcal{V}_{2,j}} \frac{1}{2}(\Delta\nu^n_{ij})\psi(X;\beta^*_{1j},\beta^*_{0j}),
\end{aligned}$$

for all $j \in [k^*_2]$ such that $|\mathcal{V}_{2,j}| = 1$. Next, by means of the second-order Taylor expansion to the function $\psi(X;\beta^*_{1j},\beta^n_{0i})\pi(Y|h_2(X,\eta^*_j),\nu^*_j)$ around the point $(\beta^*_{1j},\eta^*_j,\nu^*_j)$ with a note that $\beta^*_{1j} = 0_d$ for all $j \in [k^*_2] : |\mathcal{V}_{2,j}| > 1$, we decompose the term $B_{n,2}(Y|X)$ as

$$B_{n,2}(Y|X) = \sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|>1} \sum_{i\in\mathcal{V}_{2,j}} \sum_{|\alpha|=1}^{2} \frac{1}{\alpha!}(\Delta\beta^n_{1ij})^{\alpha_1}(\Delta\eta^n_{ij})^{\alpha_2}(\Delta\nu^n_{ij})^{\alpha_3}$$

$$\times \frac{\partial^{|\alpha_1|}\psi}{\partial \beta_1^{\alpha_1}}(X;0_d,\beta^n_{0i})\frac{\partial^{|\alpha_2|+\alpha_3}\pi}{\partial\eta^{\alpha_2}\partial\nu^{\alpha_3}}(Y|h_2(X,\eta_j^*),\nu_j^*)+R_{n,4}(Y|X)$$

$$=\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}\sum_{\rho=0}^{4}B^{(j)}_{n,2,\rho}(X)\cdot\frac{\partial^{\rho}\pi}{\partial h_2^{\rho}}(Y|h_2(X,\eta_j^*),\nu_j^*)+R_{n,4}(Y|X),$$

where $R_{n,4}(Y|X)$ is a Taylor remainder such that $R_{n,4}(Y|X)/\mathcal{D}_{3n}\to 0$ as $n\to\infty$ and

$$B^{(j)}_{n,2,0}(X):=\sum_{i\in\mathcal{V}_{2,j}}\Big[\sum_{u=1}^{d}(\Delta\beta^n_{1ij})^{(u)}\frac{\partial\psi}{\partial\beta_1^{(u)}}(X;0_d,\beta^n_{0i})+\sum_{u,v=1}^{d}\frac{(\Delta\beta^n_{1ij})^{(u)}(\Delta\beta^n_{1ij})^{(v)}}{1+1_{\{u=v\}}}\frac{\partial^2\psi}{\partial\beta_1^{(u)}\partial\beta_1^{(v)}}(X;0_d,\beta^n_{0i})\Big],$$

$$B^{(j)}_{n,2,1}(X):=\sum_{i\in\mathcal{V}_{2,j}}\Big[\sum_{u_2=1}^{d_2}(\Delta\eta^n_{ij})^{(u_2)}\frac{\partial h_2}{\partial\eta^{(u_2)}}(X.\eta_j^*)\psi(X;0_d,\beta^n_{0i})+\sum_{u_2,v_2=1}^{d_2}\frac{(\Delta\eta^n_{ij})^{(u_2)}(\Delta\eta^n_{ij})^{(v_2)}}{1+1_{\{u_2=v_2\}}}\frac{\partial^2 h_2}{\partial\eta^{(u_2)}\partial\eta^{(v_2)}}(X,\eta_j^*)$$

$$\times\psi(X;0_d,\beta^n_{0i})+\sum_{u=1}^{d}\sum_{u_2=1}^{d_2}(\Delta\beta^n_{1ij})^{(u)}(\Delta\eta^n_{ij})^{(u_2)}\frac{\partial h_2}{\partial\eta^{(u_2)}}(X.\eta_j^*)\frac{\partial\psi}{\partial\beta_1^{(u)}}(X;0_d,\beta^n_{0i})\Big],$$

$$B^{(j)}_{n,2,2}(X):=\sum_{i\in\mathcal{V}_{2,j}}\Big[\frac{1}{2}(\Delta\nu^n_{ij})\psi(X;0_d,\beta^n_{0i})+\sum_{u=1}^{d}(\Delta\beta^n_{1ij})^{(u)}\frac{1}{2}(\Delta\nu^n_{ij})\frac{\partial\psi}{\partial\beta_1^{(u)}}(X;0_d,\beta^n_{0i})$$

$$+\sum_{u_2,v_2=1}^{d_2}\frac{(\Delta\eta^n_{ij})^{(u_2)}(\Delta\eta^n_{ij})^{(v_2)}}{1+1_{\{u_2=v_2\}}}\frac{\partial^2 h_2}{\partial\eta^{(u_2)}\partial\eta^{(v_2)}}(X,\eta_j^*)\psi(X;0_d,\beta^n_{0i})\Big],$$

$$B^{(j)}_{n,2,3}(X):=\sum_{i\in\mathcal{V}_{2,j}}\Big[\sum_{u_2=1}^{d_2}(\Delta\eta^n_{ij})^{(u_2)}\frac{1}{2}(\Delta\nu^n_{ij})\frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\eta_j^*)\psi(X;0_d,\beta^n_{0i})\Big],$$

$$B^{(j)}_{n,2,4}(X):=\sum_{i\in\mathcal{V}_{2,j}}\frac{1}{8}(\Delta\nu^n_{ij})^2\psi(X;0_d,\beta^n_{0i}),$$

for all $j\in[k_2^*]$ such that $|\mathcal{V}_{2,j}|>1$.

Stage 1.2.2: In this step, we decompose the term $C_n(Y|X)$ as

$$C_n(Y|X)=\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\Big[\sum_{i\in\mathcal{V}_{2,j}}\psi(X;\beta^n_{1i},\beta^n_{0i})-\psi(X;\beta^*_{1j},\beta^*_{0j})\Big]p_{G_2^n}(Y|X)$$

$$+\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}\Big[\sum_{i\in\mathcal{V}_{2,j}}\psi(X;\beta^n_{1i},\beta^n_{0i})-\psi(X;\beta^*_{1j},\beta^*_{0j})\Big]p_{G_2^n}(Y|X)$$

$$=\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\Big[\sum_{i\in\mathcal{V}_{2,j}}\psi(X;\beta^n_{1i},\beta^n_{0i})-\psi(X;\beta^*_{1j},\beta^*_{0j})\Big]p_{G_2^n}(Y|X)$$

$$+\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}\sum_{i\in\mathcal{V}_{2,j}}\Big[\psi(X;\beta^n_{1i},\beta^n_{0i})-\psi(X;0_d,\beta^n_{0i})\Big]p_{G_2^n}(Y|X)$$

$$+\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}\Big[\sum_{i\in\mathcal{V}_{2,j}}\psi(X;0_d,\beta^n_{0i})-\psi(X;0_d,\beta^*_{0j})\Big]p_{G_2^n}(Y|X)$$

$$:=C_{n,1}(Y|X)+C_{n,2}(Y|X)+C_{n,0}(Y|X).$$

By applying the first-order Taylor expansion to the function $\psi(X;\beta_{1i}^n,\beta_{0i}^n)$ around the point $(\beta_{1j}^*,\beta_{0j}^*)$, we have

$$C_{n,1}(Y|X)=\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\sum_{i\in\mathcal{V}_{2,j}}\Big[\sum_{u=1}^{d}(\Delta\beta_{1ij}^n)^{(u)}\frac{\partial\psi}{\beta_1^{(u)}}(X;\beta_{1j}^*,\beta_{0j}^*)+(\Delta\beta_{0ij}^n)\frac{\partial\psi}{\partial\beta_0}(X;\beta_{1j}^*,\beta_{0j}^*)\Big]p_{G_2^n}(Y|X)$$
$$+R_{n,5}(Y|X),$$

where $R_{n,5}(Y|X)$ is a Taylor remainder such that $R_{n,5}(Y|X)/\mathcal{D}_{3n}\to 0$ as $n\to\infty$. Next, by means of the second-order Taylor expansion to the function $\psi(X;\beta_{1i}^n,\beta_{0i}^n)$ around the point $\beta_{1j}^*=0_d$ for $j\in[k_2^*]:|\mathcal{V}_{2,j}|>1$, we have

$$\begin{aligned}C_{n,2}(Y|X)=\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}\sum_{i\in\mathcal{V}_{2,j}}\Big[&\sum_{u=1}^{d}(\Delta\beta_{1ij}^n)^{(u)}\frac{\partial\psi}{\partial\beta_1^{(u)}}(X;0_d,\beta_{0i}^n)\\&+\sum_{u,v=1}^{d}\frac{(\Delta\beta_{1ij}^n)^{(u)}(\Delta\beta_{1ij}^n)^{(v)}}{1+1_{\{u=v\}}}\frac{\partial^2\psi}{\partial\beta_1^{(u)}\partial\beta_1^{(v)}}(X;0_d,\beta_{0i}^n)\Big]p_{G_2^n}(Y|X)+R_{n,6}(Y|X),\end{aligned}$$

where $R_{n,6}(Y|X)$ is a Taylor remainder such that $R_{n,6}(Y|X)/\mathcal{D}_{3n}\to 0$ as $n\to\infty$.

Combining the above decompositions, we can view $A_{n,0}(Y|X)/\mathcal{D}_{3n}$, $[A_{n,1}(Y|X)-R_{n,1}(Y|X)]/\mathcal{D}_{3n}$, $[A_{n,2}(Y|X)-R_{n,2}(Y|X)]/\mathcal{D}_{3n}$, $B_{n,0}(Y|X)/\mathcal{D}_{3n}$, $[B_{n,1}(Y|X)-R_{n,3}(Y|X)]/\mathcal{D}_{3n}$, $[B_{n,2}(Y|X)-R_{n,4}(Y|X)]/\mathcal{D}_{3n}$, $C_{n,0}(Y|X)/\mathcal{D}_{3n}$, $[C_{n,1}(Y|X)-R_{n,5}(Y|X)]/\mathcal{D}_{3n}$ and $[C_{n,2}(Y|X)-R_{n,6}(Y|X)]/\mathcal{D}_{3n}$ as a combination of elements from the following sets

$$\begin{aligned}\mathcal{S}_{0,j}&:=\{\pi(Y|h_1(X,\kappa_j^*),\tau_j^*)\},\\
\mathcal{S}_{1,j}&:=\left\{\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*)\frac{\partial\pi}{\partial h_1}(Y|h_1(X,\kappa_j^*),\tau_j^*),\ \frac{\partial^2 h_1}{\partial\kappa^{(u_1)}\partial\kappa^{(v_1)}}(X,\kappa_j^*)\frac{\partial\pi}{\partial h_1}(Y|h_1(X,\kappa_j^*),\tau_j^*):u_1,v_1\in[d_1]\right\},\\
\mathcal{S}_{2,j}&:=\left\{\frac{\partial^2\pi}{\partial h_1^2}(Y|h_1(X,\kappa_j^*),\tau_j^*),\ \frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*)\frac{\partial h_1}{\partial\kappa^{(v_1)}}(X,\kappa_j^*)\frac{\partial^2\pi}{\partial h_1^2}(Y|h_1(X,\kappa_j^*),\tau_j^*):u_1,v_1\in[d_1]\right\},\\
\mathcal{S}_{3,j}&:=\left\{\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*)\frac{\partial^3\pi}{\partial h_1^3}(Y|h_1(X,\kappa_j^*),\tau_j^*):u_1,v_1\in[d_1]\right\},\\
\mathcal{S}_{4,j}&:=\left\{\frac{\partial^4\pi}{\partial h_1^4}(Y|h_1(X,\kappa_j^*),\tau_j^*):u_1,v_1\in[d_1]\right\},\end{aligned}$$

for all $j\in[k_1^*]$, and

$$\begin{aligned}\mathcal{T}_{0,j}:=\Bigg\{&\pi(Y|h_2(X,\eta_j^*),\nu_j^*),\ \frac{\partial\psi}{\partial\beta_1^{(u)}}(X;0_d,\beta_{0i}^n)\pi(Y|h_2(X,\eta_j^*),\nu_j^*),\\
&\frac{\partial^2\psi}{\partial\beta_1^{(u)}\partial\beta_1^{(v)}}(X;0_d,\beta_{0i}^n)\pi(Y|h_2(X,\eta_j^*),\nu_j^*):u,v\in[d]\Bigg\},\\
\mathcal{T}_{1,j}:=\Bigg\{&\frac{\partial h_2}{\partial\eta^{(u_2)}}(X.\eta_j^*)\psi(X;0_d,\beta_{0i}^n)\frac{\partial\pi}{\partial h_2}(Y|h_2(X,\eta_j^*),\nu_j^*),\end{aligned}$$

$$\frac{\partial h_2}{\partial \eta^{(u_2)}}(X.\eta^*_j)\frac{\partial \psi}{\partial \beta_1^{(u)}}(X;0_d,\beta^n_{0i})\frac{\partial \pi}{\partial h_2}(Y|h_2(X,\eta^*_j),\nu^*_j) : u \in [d],\ u_2 \in [d_2]\Bigg\},$$

$$\mathcal{T}_{2,j} := \Bigg\{\psi(X;0_d,\beta^n_{0i})\frac{\partial^2 \pi}{\partial h_2^2}(Y|h_2(X,\eta^*_j),\nu^*_j),\ \frac{\partial \psi}{\partial \beta_1^{(u)}}(X;0_d,\beta^n_{0i})\frac{\partial^2 \pi}{\partial h_2^2}(Y|h_2(X,\eta^*_j),\nu^*_j),$$

$$\frac{\partial^2 h_2}{\partial \eta^{(u_2)}\partial \eta^{(v_2)}}(X,\eta^*_j)\psi(X;0_d,\beta^n_{0i})\frac{\partial^2 \pi}{\partial h_2^2}(Y|h_2(X,\eta^*_j),\nu^*_j) : u \in [d],\ u_2, v_2 \in [d_2]\Bigg\},$$

$$\mathcal{T}_{3,j} := \Bigg\{\frac{\partial h_2}{\partial \eta^{(u_2)}}(X,\eta^*_j)\psi(X;0_d,\beta^n_{0i})\frac{\partial^3 \pi}{\partial h_2^3}(Y|h_2(X,\eta^*_j),\nu^*_j) : u_2 \in [d_2]\Bigg\},$$

$$\mathcal{T}_{4,j} := \Bigg\{\psi(X;0_d,\beta^n_{0i})\frac{\partial^4 \pi}{\partial h_2^4}(Y|h_2(X,\eta^*_j),\nu^*_j)\Bigg\},$$

$$\mathcal{T}_{5,j} := \Bigg\{\frac{\partial \psi}{\beta_1^{(u)}}(X;\beta^*_{1j},\beta^*_{0j})p_{G^n_2}(Y|X),\ \frac{\partial \psi}{\partial \beta_0}(X;\beta^*_{1j},\beta^*_{0j})p_{G^n_2}(Y|X),\ \frac{\partial \psi}{\partial \beta_1^{(u)}}(X;0_d,\beta^n_{0i})p_{G^n_2}(Y|X),$$

$$\frac{\partial^2 \psi}{\partial \beta_1^{(u)}\partial \beta_1^{(v)}}(X;0_d,\beta^n_{0i})p_{G^n_2}(Y|X) : u \in [d]\Bigg\},$$

for all $j \in [k^*_2]$.

Stage 2 - Non-vanishing coefficients: In this stage, we demonstrate that not all the coefficients in the representations of $A_{n,0}(Y|X)/\mathcal{D}_{3n}$, $[A_{n,1}(Y|X)-R_{n,1}(Y|X)]/\mathcal{D}_{3n}$, $[A_{n,2}(Y|X)-R_{n,2}(Y|X)]/\mathcal{D}_{3n}$, $B_{n,0}(Y|X)/\mathcal{D}_{3n}$, $[B_{n,1}(Y|X)-R_{n,3}(Y|X)]/\mathcal{D}_{3n}$, $[B_{n,2}(Y|X)-R_{n,4}(Y|X)]/\mathcal{D}_{3n}$, $C_{n,0}(Y|X)/\mathcal{D}_{3n}$, $[C_{n,1}(Y|X)-R_{n,5}(Y|X)]/\mathcal{D}_{3n}$ and $[C_{n,2}(Y|X)-R_{n,6}(Y|X)]/\mathcal{D}_{3n}$ go to zero when $n \to \infty$. Assume by contrary that all these coefficients converge to zero. By using the same arguments as in Stage 2 in Appendix D.1, we have

$$\frac{1}{\mathcal{D}_{3n}}\Big[\sum_{j=1}^{k^*_1}\Big|\sum_{i\in\mathcal{V}_{1,j}}\omega^n_i-\omega^*_j\Big|+\sum_{j\in[k^*_1]:|\mathcal{V}_{1,j}|=1}\sum_{i\in\mathcal{V}_{1,j}}\omega^n_i(\|\Delta\kappa^n_{ij}\|+|\Delta\tau^n_{ij}|)$$
$$+\sum_{j\in[k^*_1]:|\mathcal{V}_{1,j}|>1}\sum_{i\in\mathcal{V}_{1,j}}\omega^n_i(\|\Delta\kappa^n_{ij}\|^2+|\Delta\tau^n_{ij}|^2)\Big]\to 0,$$

as $n \to \infty$. Additionally, by considering the coefficients of the terms:

- $\frac{\partial \psi}{\partial \beta_1^{(u)}}(X;\beta^*_{1j},\beta^*_{0j})\pi(Y|h_2(X,\eta^*_j),\nu^*_j)$ for $j \in [k^*_2] : |\mathcal{V}_{2,j}| = 1$, we get

$$\frac{1}{\mathcal{D}_{3n}}\sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|=1}\sum_{i\in\mathcal{V}_{2,j}}\|\Delta\beta^n_{1ij}\|\to 0;$$

- $\frac{\partial \psi}{\partial \beta_0}(X;\beta^*_{1j},\beta^*_{0j})\pi(Y|h_2(X,\eta^*_j),\nu^*_j)$ for $j \in [k^*_2] : |\mathcal{V}_{2,j}| = 1$, we get

$$\frac{1}{\mathcal{D}_{3n}}\sum_{j\in[k^*_2]:|\mathcal{V}_{2,j}|=1}\sum_{i\in\mathcal{V}_{2,j}}|\Delta\beta^n_{0ij}|\to 0;$$

- $\frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\eta_j^*)\psi(X;\beta_{1j}^*,\beta_{0j}^*)\frac{\partial\pi}{\partial h_2}(Y|h_2(X,\eta_j^*),\nu_j^*)$ for $j\in[k_2^*]:|\mathcal{V}_{2,j}|=1$, we get

$$\frac{1}{\mathcal{D}_{3n}}\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\sum_{i\in\mathcal{V}_{2,j}}\|\Delta\eta_{ij}^n\|\to 0;$$

- $\psi(X;\beta_{1j}^*,\beta_{0j}^*)\frac{\partial\pi}{\partial h_2}(Y|h_2(X,\eta_j^*),\nu_j^*)$ for $j\in[k_2^*]:|\mathcal{V}_{2,j}|=1$, we get

$$\frac{1}{\mathcal{D}_{3n}}\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\sum_{i\in\mathcal{V}_{2,j}}|\Delta\nu_{ij}^n|\to 0;$$

- $\pi(Y|h_2(X,\eta_j^*),\nu_j^*)$ for $j\in[k_2^*]:|\mathcal{V}_{2,j}|>1$, we get

$$\frac{1}{\mathcal{D}_{3n}}\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}\Big|\sum_{i\in\mathcal{V}_{2,j}}\sigma(\beta_{0i}^n)-\sigma(\beta_{0j}^*)\Big|\to 0;$$

- $\frac{\partial^2\psi}{\partial\beta_1^{(u)}\partial\beta_1^{(v)}}(X;0_d,\beta_{0i}^n)\pi(Y|h_2(X,\eta_j^*),\nu_j^*)$ for $j\in[k_2^*]:|\mathcal{V}_{2,j}|>1$, we get

$$\frac{1}{\mathcal{D}_{3n}}\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}\sum_{i\in\mathcal{V}_{2,j}}\|\Delta\beta_{1ij}^n\|^2\to 0;$$

- $\frac{\partial^2 h_2}{\partial\eta^{(u_2)}\partial\eta^{(v_2)}}(X,\eta_j^*)\psi(X;0_d,\beta_{0i}^n)\frac{\partial^2\pi}{\partial h_2^2}(Y|h_2(X,\eta_j^*),\nu_j^*)$ for $j\in[k_2^*]:|\mathcal{V}_{2,j}|>1$, we get

$$\frac{1}{\mathcal{D}_{3n}}\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}\sum_{i\in\mathcal{V}_{2,j}}\|\Delta\eta_{ij}^n\|^2\to 0;$$

- $\psi(X;0_d,\beta_{0i}^n)\frac{\partial^4\pi}{\partial h_2^4}(Y|h_2(X,\eta_j^*),\nu_j^*)$ for $j\in[k_2^*]:|\mathcal{V}_{2,j}|>1$, we get

$$\frac{1}{\mathcal{D}_{3n}}\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}\sum_{i\in\mathcal{V}_{2,j}}|\Delta\nu_{ij}^n|^2\to 0.$$

Putting the above limits together, we deduce $1=\frac{\mathcal{D}_{3n}}{\mathcal{D}_{3n}}\to 0$ as $n\to\infty$, which is a contradiction. Therefore, at least one among the coefficients in the representations of $A_{n,0}(Y|X)/\mathcal{D}_{3n}$, $[A_{n,1}(Y|X)-R_{n,1}(Y|X)]/\mathcal{D}_{3n}$, $[A_{n,2}(Y|X)-R_{n,2}(Y|X)]/\mathcal{D}_{3n}$, $B_{n,0}(Y|X)/\mathcal{D}_{3n}$, $[B_{n,1}(Y|X)-R_{n,3}(Y|X)]/\mathcal{D}_{3n}$, $[B_{n,2}(Y|X)-R_{n,4}(Y|X)]/\mathcal{D}_{3n}$, $C_{n,0}(Y|X)/\mathcal{D}_{3n}$, $[C_{n,1}(Y|X)-R_{n,5}(Y|X)]/\mathcal{D}_{3n}$ and $[C_{n,2}(Y|X)-R_{n,6}(Y|X)]/\mathcal{D}_{3n}$ does not go to zero.

Stage 3 - Fatou's lemma contradiction: In this stage, we use the Fatou's lemma to show a contradiction to the result of Stage 2. For that purpose, let us denote m_n as the maximum of the absolute values of the coefficients in the representations of $A_{n,0}(Y|X)/\mathcal{D}_{3n}$, $[A_{n,1}(Y|X)-R_{n,1}(Y|X)]/\mathcal{D}_{3n}$, $[A_{n,2}(Y|X)-R_{n,2}(Y|X)]/\mathcal{D}_{3n}$, $B_{n,0}(Y|X)/\mathcal{D}_{3n}$, $[B_{n,1}(Y|X)-R_{n,3}(Y|X)]/\mathcal{D}_{3n}$, $[B_{n,2}(Y|X)-R_{n,4}(Y|X)]/\mathcal{D}_{3n}$, $C_{n,0}(Y|X)/\mathcal{D}_{3n}$, $[C_{n,1}(Y|X)-R_{n,5}(Y|X)]/\mathcal{D}_{3n}$ and $[C_{n,2}(Y|X)-$

$R_{n,6}(Y|X)]/\mathcal{D}_{3n}$. It follows from the result of Stage 2 that $1/m_n \not\to \infty$ as $n \to \infty$. In addition, we also denote

$$\frac{1}{m_n\mathcal{D}_{3n}} \cdot \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n (\Delta\kappa_{ij}^n)^{(u_1)} \to s_{1,j}^{(u_1)}, \quad \frac{1}{m_n\mathcal{D}_{3n}} \cdot \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n (\Delta\tau_{ij}^n) \to s_{2,j},$$

$$\frac{1}{m_n\mathcal{D}_{3n}} \cdot \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n (\Delta\kappa_{ij}^n)^{(u_1)} (\Delta\kappa_{ij}^n)^{(v_1)} \to s_{3,j}^{(u_1 v_1)}, \quad \frac{1}{m_n\mathcal{D}_{3n}} \cdot \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n (\Delta\tau_{ij}^n)^2 \to s_{4,j},$$

$$\frac{1}{m_n\mathcal{D}_{3n}} \cdot \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n (\Delta\kappa_{ij}^n)^{(u_1)} (\Delta\tau_{ij}^n) \to s_{5,j}^{(u_1)}, \quad \frac{1}{m_n\mathcal{D}_{3n}} \cdot \Big(\sum_{i \in \mathcal{V}_{1,j}} \omega_i^n - \omega_j^* \Big) \to s_{0,j},$$

for all $j \in [k_1^*]$ and

$$\frac{1}{m_n\mathcal{D}_{3n}} \cdot \sum_{i \in \mathcal{V}_{2,j}} (\Delta\beta_{0ij}^n) \to t_{0,j}, j \in [k_2^*] : |\mathcal{V}_{2,j}| = 1, \quad \frac{1}{m_n\mathcal{D}_{3n}} \cdot \sum_{i \in \mathcal{V}_{2,j}} (\Delta\beta_{1ij}^n)^{(u)} \to t_{1,j}^{(u)},$$

$$\frac{1}{m_n\mathcal{D}_{3n}} \cdot \sum_{i \in \mathcal{V}_{2,j}} (\Delta\eta_{ij}^n)^{(u_2)} \to t_{2,j}^{(u_2)}, \quad \frac{1}{m_n\mathcal{D}_{3n}} \cdot \sum_{i \in \mathcal{V}_{2,j}} (\Delta\nu_{ij}^n) \to t_{3,j},$$

for all $j \in [k_2^*] : |\mathcal{V}_{2,j}| = 1$, and

$$\frac{1}{m_n\mathcal{D}_{3n}} \cdot \Big(\sum_{i \in \mathcal{V}_{2,j}} \sigma(\beta_{0i}^n) - \sigma(\beta_{0j}^*) \Big) \to t_{0,j}, \quad \frac{1}{m_n\mathcal{D}_{3n}} \cdot (\Delta\beta_{1ij}^n)^{(u)} \to t_{1,j,i}^{(u)},$$

$$\frac{1}{m_n\mathcal{D}_{3n}} \cdot (\Delta\eta_{ij}^n)^{(u_2)} \to t_{2,j,i}^{(u_2)}, \quad \frac{1}{m_n\mathcal{D}_{3n}} \cdot (\Delta\nu_{ij}^n) \to t_{3,j,i},$$

$$\frac{1}{m_n\mathcal{D}_{3n}} \cdot (\Delta\beta_{1ij}^n)^{(u)} (\Delta\beta_{1ij}^n)^{(v)} \to t_{4,j,i}^{(uv)}, \quad \frac{1}{m_n\mathcal{D}_{3n}} \cdot = (\Delta\eta_{ij}^n)^{(u_2)} (\Delta\eta_{ij}^n)^{(v_2)} \to t_{5,j,i}^{(u_2 v_2)},$$

$$\frac{1}{m_n\mathcal{D}_{3n}} \cdot (\Delta\nu_{ij}^n)^2 \to t_{6,j,i}, \quad \frac{1}{m_n\mathcal{D}_{3n}} \cdot (\Delta\beta_{1ij}^n)^{(u)} (\Delta\eta_{ij}^n)^{(v_2)} \to t_{7,j,i}^{(uv_2)},$$

$$\frac{1}{m_n\mathcal{D}_{3n}} \cdot (\Delta\beta_{1ij}^n)^{(u)} (\Delta\nu_{ij}^n) \to t_{8,j,i}^{(u)}, \quad \frac{1}{m_n\mathcal{D}_{3n}} \cdot (\Delta\eta_{ij}^n)^{(u_2)} (\Delta\nu_{ij}^n) \to t_{9,j,i}^{(u_2)},$$

for all $j \in [k_2^*] : |\mathcal{V}_{2,j}| > 1$ as $n \to \infty$. Due to the result of Stage 2, at least one among the above limits is non-zero. Recall from equation (32) that we get

$$\mathbb{E}_X[V(g_{G_1^n,G_2^n}(\cdot|X), g_{G_1^*,G_2^*}(\cdot|X))]/\mathcal{D}_{3n} \to 0,$$

Moreover, by means of the Fatou's lemma, we have

$$\lim_{n\to\infty} \frac{\mathbb{E}_X[V(g_{G_1^n,G_2^n}(\cdot|X), g_{G_1^*,G_2^*}(\cdot|X))]}{m_n\mathcal{D}_{3n}} \geq \int \liminf_{n\to\infty} \frac{|g_{G_1^n,G_2^n}(Y|X) - g_{G_1^*,G_2^*}(Y|X)|}{2m_n\mathcal{D}_{3n}} \mathrm{d}(X,Y).$$

Then, we deduce $[g_{G_1^n,G_2^n}(Y|X) - g_{G_1^*,G_2^*}(Y|X)]/[m_n\mathcal{D}_{3n}] \to 0$ as $n \to \infty$ for almost surely (X,Y). Since the input space is bounded and the parameter space is compact, the quantity $\sum_{j=1}^{k_2^*} \sigma((\beta_{1j}^*)^\top X +$

$\beta_{0j}^*)$ is bounded. Thus, we also have

$$\Big[\sum_{j=1}^{k_2^*}\sigma((\beta_{1j}^*)^\top X+\beta_{0j}^*)\Big][g_{G_1^n,G_2^n}(Y|X)-g_{G_1^*,G_2^*}(Y|X)]/[m_n\mathcal{D}_{3n}]\to 0,$$

implying that

$$\frac{1}{2}\Big[\sum_{j=1}^{k_2^*}\sigma((\beta_{1j}^*)^\top X+\beta_{0j}^*)\Big]\cdot\frac{q_{G_1^n}(Y|X)-q_{G_1^*}(Y|X)}{m_n\mathcal{D}_{3n}}+\frac{1}{2}\frac{Q_n(Y|X)}{m_n\mathcal{D}_{3n}}\to 0.$$

as $n\to\infty$ for almost surely (X,Y). From the decomposition of the terms $q_{G_1^n}(Y|X)-q_{G_1^*}(Y|X)$ and $Q_n(Y|X)$ in Stage 1, we have

$$\begin{aligned}\frac{1}{2}\Big[\sum_{j=1}^{k_2^*}\sigma((\beta_{1j}^*)^\top X+\beta_{0j}^*)\Big]\cdot\frac{A_{n,2}(Y|X)+A_{n,1}(Y|X)+A_{n,0}(Y|X)}{m_n\mathcal{D}_{3n}}\\+\frac{1}{2}\frac{B_{n,1}(Y|X)+B_{n,2}(Y|X)+B_{n,3}(Y|X)-C_{n,1}(Y|X)-C_{n,2}(Y|X)-C_{n,3}(Y|X)}{m_n\mathcal{D}_{3n}}\to 0.\end{aligned}\tag{34}$$

We have

$$\begin{aligned}\lim_{n\to\infty}\frac{A_{n,0}(Y|X)}{m_n\mathcal{D}_{3n}}&=\sum_{j=1}^{k_1^*}s_{0,j}\pi(Y|h_1(X,\kappa_j^*),\tau_j^*),\\
\lim_{n\to\infty}\frac{A_{n,1}(Y|X)}{m_n\mathcal{D}_{3n}}&=\sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|=1}\Big[\sum_{u_1=1}^{d_1}s_{1,j}^{(u_1)}\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*)\frac{\partial\pi}{\partial h_1}(Y|h_1(X,\kappa_j^*),\tau_j^*)\\
&\qquad+\frac{1}{2}s_{2,j}\frac{\partial^2\pi}{\partial h_1^2}(Y|h_1(X,\kappa_j^*),\tau_j^*)\Big],\\
\lim_{n\to\infty}\frac{A_{n,2}(Y|X)}{m_n\mathcal{D}_{3n}}&=\sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|>1}\Big[\Big(\sum_{u_1=1}^{d_1}s_{1,j}^{(u_1)}\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*)+\sum_{u_1,v_1=1}^{d_1}\frac{s_{3,j}^{(u_1v_1)}}{1+1_{\{u_1=v_1\}}}\frac{\partial^2 h_1}{\partial\kappa^{(u_1)}\partial\kappa^{(v_1)}}(X,\kappa_j^*)\Big)\\
\times\frac{\partial\pi}{\partial h_1}(Y|h_1(X,\kappa_j^*),\tau_j^*)&+\Big(\frac{1}{2}s_{2,j}+\sum_{u_1,v_1=1}^{d_1}\frac{s_{3,j}^{(u_1v_1)}}{1+1_{\{u_1=v_1\}}}\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*)\frac{\partial h_1}{\partial\kappa^{(v_1)}}(X,\kappa_j^*)\Big)\frac{\partial^2\pi}{\partial h_1^2}(Y|h_1(X,\kappa_j^*),\tau_j^*)\\
&+\Big(\frac{1}{2}\sum_{u_1=1}^{d_1}s_{5,j}^{(u_1)}\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*)\Big)\frac{\partial^3\pi}{\partial h_1^3}(Y|h_1(X,\kappa_j^*),\tau_j^*)+\frac{1}{8}s_{4,j}\frac{\partial^4\pi}{\partial h_1^4}(Y|h_1(X,\kappa_j^*),\tau_j^*)\Big],\end{aligned}$$

and

$$\begin{aligned}\lim_{n\to\infty}\frac{B_{n,0}(Y|X)}{m_n\mathcal{D}_{3n}}&=\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1}t_{0,j}\pi(Y|h_2(X,\eta_j^*),\nu_j^*),\\
\lim_{n\to\infty}\frac{B_{n,1}(Y|X)}{m_n\mathcal{D}_{3n}}&=\sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1}\Big[\Big(\sum_{u=1}^{d}t_{1,j}^{(u)}\frac{\partial\psi}{\partial\beta_1^{(u)}}(X;\beta_{1j}^*,\beta_{0j}^*)+t_{0,j}\frac{\partial\psi}{\partial\beta_0}(X;\beta_{1j}^*,\beta_{0j}^*)\Big)\pi(Y|h_2(X,\eta_j^*),\nu_j^*)\end{aligned}$$

$$+ \sum_{u_2=1}^{d_2} t_{2,j}^{(u_2)} \frac{\partial h_2}{\partial \eta^{(u_2)}}(X, \eta_j^*) \psi(X; \beta_{1j}^*, \beta_{0j}^*) \frac{\partial \pi}{\partial h_2}(Y|h_2(X, \eta_j^*), \nu_j^*) + \frac{1}{2} t_{3,j} \psi(X; \beta_{1j}^*, \beta_{0j}^*) \frac{\partial^2 \pi}{\partial h_2^2}(Y|h_2(X, \eta_j^*), \nu_j^*)\Big],$$

$$\lim_{n\to\infty} \frac{B_{n,2}(Y|X)}{m_n \mathcal{D}_{3n}} = \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1} \sum_{i\in\mathcal{V}_{2,j}} \Big[\Big(\sum_{u,v=1}^{d} \frac{t_{4,j,i}^{(uv)}}{1 + 1_{\{u=v\}}} \frac{\partial^2 \psi}{\partial \beta_1^{(u)} \partial \beta_1^{(v)}}(X; 0_d, \bar{\beta}_{0i})$$

$$+ \sum_{u=1}^{d} t_{1,j,i}^{(u)} \frac{\partial \psi}{\partial \beta_1^{(u)}}(X; 0_d, \bar{\beta}_{0i})\Big) \pi(Y|h_2(X, \eta_j^*), \nu_j^*) + \Big(\sum_{u=1}^{d} \sum_{u_2=1}^{d_2} t_{7,j,i}^{(uu_2)} \frac{\partial h_2}{\partial \eta^{(u_2)}}(X.\eta_j^*) \frac{\partial \psi}{\partial \beta_1^{(u)}}(X; 0_d, \bar{\beta}_{0i})$$

$$+ \sum_{u_2=1}^{d_2} t_{2,j,i}^{(u_2)} \frac{\partial h_2}{\partial \eta^{(u_2)}}(X.\eta_j^*) \psi(X; 0_d, \bar{\beta}_{0i}) + \sum_{u_2,v_2=1}^{d_2} t_{5,j,i}^{(u_2 v_2)} \frac{\partial^2 h_2}{\partial \eta^{(u_2)} \partial \eta^{v_2)}}(X, \eta_j^*) \psi(X; 0_d, \bar{\beta}_{0i})\Big) \frac{\partial \pi}{\partial h_2}(Y|h_2(X, \eta_j^*), \nu_j^*)$$

$$+ \Big(\sum_{u=1}^{d} \frac{1}{2} t_{8,j,i}^{(u)} \frac{\partial \psi}{\partial \beta_1^{(u)}}(X; 0_d, \bar{\beta}_{0i}) + \frac{1}{2} t_{3,j,i} \psi(X; 0_d, \bar{\beta}_{0i}) + \sum_{u_2,v_2=1}^{d_2} \frac{t_{5,j,i}^{(u_2 v_2)}}{1 + 1_{\{u_2=v_2\}}} \frac{\partial h_2}{\partial \eta^{(u_2)}}(X, \eta_j^*) \frac{\partial h_2}{\partial \eta^{(v_2)}}(X, \eta_j^*)$$

$$\times \psi(X; 0_d, \bar{\beta}_{0i})\Big) \frac{\partial^2 \pi}{\partial h_2^2}(Y|h_2(X, \eta_j^*), \nu_j^*) + \sum_{u_2=1}^{d_2} \frac{1}{2} t_{9,j,i}^{(u_2)} \frac{\partial h_2}{\partial \eta^{(u_2)}}(X, \eta_j^*) \psi(X; 0_d, \bar{\beta}_{0i}) \frac{\partial^3 \pi}{\partial h_2^3}(Y|h_2(X, \eta_j^*), \nu_j^*)$$

$$+ \frac{1}{8} t_{6,j,i} \psi(X; 0_d, \bar{\beta}_{0i}) \frac{\partial^4 \pi}{\partial h_2^4}(Y|h_2(X, \eta_j^*), \nu_j^*)\Big],$$

and

$$\lim_{n\to\infty} \frac{C_{n,0}(Y|X)}{m_n \mathcal{D}_{3n}} = \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1} t_{0,j} p_{G_2^*}(Y|X),$$

$$\lim_{n\to\infty} \frac{C_{n,1}(Y|X)}{m_n \mathcal{D}_{3n}} = \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|=1} \Big[\sum_{u=1}^{d} t_{1,j}^{(u)} \frac{\partial \psi}{\beta_1^{(u)}}(X; \beta_{1j}^*, \beta_{0j}^*) + t_{0,j} \frac{\partial \psi}{\partial \beta_0}(X; \beta_{1j}^*, \beta_{0j}^*)\Big] p_{G_2^*}(Y|X),$$

$$\lim_{n\to\infty} \frac{C_{n,2}(Y|X)}{m_n \mathcal{D}_{3n}} = \sum_{j\in[k_2^*]:|\mathcal{V}_{2,j}|>1} \sum_{i\in\mathcal{V}_{2,j}} \Big[\sum_{u=1}^{d} t_{1,j,i}^{(u)} \frac{\partial \psi}{\beta_1^{(u)}}(X; 0_d, \bar{\beta}_{0i})$$

$$+ \sum_{u,v=1}^{d} \frac{t_{4,j,i}^{(uv)}}{1 + 1_{\{u=v\}}} \frac{\partial^2 \psi}{\partial \beta_1^{(u)} \partial \beta_1^{(v)}}(X; 0_d, \bar{\beta}_{0i})\Big] p_{G_2^*}(Y|X).$$

Note that for almost every X, the set

$$\left\{ \Big[\sum_{j=1}^{k_2^*} \sigma((\beta_{1j}^*)^\top X + \beta_{0j}^*)\Big] \frac{\partial^\rho \pi}{\partial h_1^\rho}(Y|h_1(X, \kappa_j^*), \tau_j^*) : 0 \le \rho \le 4,\ j \in [k_1^*] \right\}$$

$$\cup \left\{ \frac{\partial \psi}{\partial \beta_1^{(u)}}(X; \beta_{1j}^*, \beta_{0j}^*) \pi(Y|h_2(X, \eta_j^*), \nu_j^*),\ \frac{\partial \psi}{\partial \beta_0}(X; \beta_{1j}^*, \beta_{0j}^*) \pi(Y|h_2(X, \eta_j^*), \nu_j^*), \right.$$

$$\frac{\partial \psi}{\partial \beta_1^{(u)}}(X; \beta_{1j}^*, \beta_{0j}^*) p_{G_2^*}(Y|X),\ \frac{\partial \psi}{\partial \beta_0}(X; \beta_{1j}^*, \beta_{0j}^*) p_{G_2^*}(Y|X),\ \psi(X; \beta_{1j}^*, \beta_{0j}^*) \frac{\partial \pi}{\partial h_2}(Y|h_2(X, \eta_j^*), \nu_j^*),$$

$$\left. \psi(X; \beta_{1j}^*, \beta_{0j}^*) \frac{\partial^2 \pi}{\partial h_2^2}(Y|h_2(X, \eta_j^*), \nu_j^*) : u \in [d],\ j \in [k_2^*] : |\mathcal{V}_{2,j}| = 1 \right\}$$

$$\cup \Bigg\{ \frac{\partial \psi}{\partial \beta_1^{(u)}}(X; 0_d, \bar{\beta}_{0i}) \pi(Y | h_2(X, \eta_j^*), \nu_j^*), \ \frac{\partial^2 \psi}{\partial \beta_1^{(u)} \partial \beta_1^{(v)}}(X; 0_d, \bar{\beta}_{0i}) \pi(Y | h_2(X, \eta_j^*), \nu_j^*), \ \pi(Y | h_2(X, \eta_j^*), \nu_j^*)$$

$$\frac{\partial \psi}{\partial \beta_1^{(u)}}(X; 0_d, \bar{\beta}_{0i}) \frac{\partial \pi}{\partial h_2}(Y | h_2(X, \eta_j^*), \nu_j^*), \ \psi(X; 0_d, \bar{\beta}_{0i}) \frac{\partial \pi}{\partial h_2}(Y | h_2(X, \eta_j^*), \nu_j^*),$$

$$\frac{\partial \psi}{\partial \beta_1^{(u)}}(X; 0_d, \bar{\beta}_{0i}) \frac{\partial^2 \pi}{\partial h_2^2}(Y | h_2(X, \eta_j^*), \nu_j^*), \ \psi(X; 0_d, \bar{\beta}_{0i}) \frac{\partial^2 \pi}{\partial h_2^2}(Y | h_2(X, \eta_j^*), \nu_j^*),$$

$$\psi(X; 0_d, \bar{\beta}_{0i}) \frac{\partial^3 \pi}{\partial h_2^3}(Y | h_2(X, \eta_j^*), \nu_j^*), \ \psi(X; 0_d, \bar{\beta}_{0i}) \frac{\partial^4 \pi}{\partial h_2^4}(Y | h_2(X, \eta_j^*), \nu_j^*),$$

$$\frac{\partial \psi}{\partial \beta_1^{(u)}}(X; 0_d, \bar{\beta}_{0i}) p_{G_2^*}(Y|X), \ \frac{\partial^2 \psi}{\partial \beta_1^{(u)} \partial \beta_1^{(v)}}(X; 0_d, \bar{\beta}_{0i}) p_{G_2^*}(Y|X) : u, v \in [d], j \in [k_2^*] : |\mathcal{V}_{2,j}| > 1, i \in \mathcal{V}_{2,j} \Bigg\}$$

is linearly independent w.r.t Y, implying that the coefficients of those terms in the limit in equation (34) are equal to zero.

For $j \in [k_1^*]$, by looking at the coefficient of the term $\left[\sum_{j=1}^{k_2^*} \sigma((\beta_{1j}^*)^\top X + \beta_{0j}^*)\right] \pi(Y | h_1(X, \kappa_j^*), \tau_j^*)$, we have $s_{0,j} = 0$.

For $j \in [k_1^*]$ such that $|\mathcal{V}_{1,j}| = 1$, by considering the coefficients of

- $\left[\sum_{j=1}^{k_2^*} \sigma((\beta_{1j}^*)^\top X + \beta_{0j}^*)\right] \frac{\partial \pi}{\partial h_1}(Y | h_1(X, \kappa_j^*), \tau_j^*)$, we have $\sum_{u_1=1}^{d_1} s_{1,j}^{(u_1)} \frac{\partial h_1}{\partial \kappa^{(u_1)}}(X, \kappa_j^*) = 0$ for almost every X. Since the expert function h_1 is strongly identifiable, we get $s_{1,j}^{(u_1)} = 0$ for all $u_1 \in [d_1]$;

- $\left[\sum_{j=1}^{k_2^*} \sigma((\beta_{1j}^*)^\top X + \beta_{0j}^*)\right] \frac{\partial^2 \pi}{\partial h_1^2}(Y | h_1(X, \kappa_j^*), \tau_j^*)$, we have $s_{2,j} = 0$.

For $j \in [k_1^*]$ such that $|\mathcal{V}_{1,j}| > 1$, by taking into account the coefficients of

- $\left[\sum_{j=1}^{k_2^*} \sigma((\beta_{1j}^*)^\top X + \beta_{0j}^*)\right] \frac{\partial \pi}{\partial h_1}(Y | h_1(X, \kappa_j^*), \tau_j^*)$, we have

$$\sum_{u_1=1}^{d_1} s_{1,j} \frac{\partial h_1}{\partial \kappa^{(u_1)}}(X, \kappa_j^*) + \sum_{u_1, v_1=1}^{d_1} \frac{s_{3,j}^{(u_1 v_1)}}{1 + 1_{\{u_1 = v_1\}}} \frac{\partial^2 h_1}{\partial \kappa^{(u_1)} \partial \kappa^{(v_1)}}(X, \kappa_j^*) = 0,$$

  for almost every X. Since the expert function h_1 satisfies the strong identifiability condition, we get $s_{1,j}^{(u_1)} = s_{3,j}^{(u_1 v_1)} = 0$ for all $u_1, v_1 \in [d_1]$;

- $\left[\sum_{j=1}^{k_2^*} \sigma((\beta_{1j}^*)^\top X + \beta_{0j}^*)\right] \frac{\partial^2 \pi}{\partial h_1^2}(Y | h_1(X, \kappa_j^*), \tau_j^*)$, we have

$$\frac{1}{2} s_{2,j} + \sum_{u_1, v_1=1}^{d_1} \frac{s_{3,j}^{(u_1 v_1)}}{1 + 1_{\{u_1 = v_1\}}} \frac{\partial h_1}{\partial \kappa^{(u_1)}}(X, \kappa_j^*) \frac{\partial h_1}{\partial \kappa^{(v_1)}}(X, \kappa_j^*) = 0,$$

  for almost every X. Since $s_{3,j}^{(u_1 v_1)} = 0$ for all $u_1, v_1 \in [d_1]$, we deduce $s_{2,j} = 0$;

- $\Big[\sum_{j=1}^{k_2^*}\sigma((\beta_{1j}^*)^\top X+\beta_{0j}^*)\Big]\frac{\partial^3\pi}{\partial h_1^3}(Y|h_1(X,\kappa_j^*),\tau_j^*)$, we have $\frac{1}{2}\sum_{u_1=1}^{d_1}s_{5,j}^{(u_1)}\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*)=0$, for almost every X. As the expert function h_1 meets the strong identifiability condition, we get $s_{5,j}^{(u_1)}=0$ for all $u_1\in[d_1]$;
- $\Big[\sum_{j=1}^{k_2^*}\sigma((\beta_{1j}^*)^\top X+\beta_{0j}^*)\Big]\frac{\partial^4\pi}{\partial h_1^4}(Y|h_1(X,\kappa_j^*),\tau_j^*)$, we have $s_{4,j}=0$.

For $j\in[k_2^*]$ such that $|\mathcal{V}_{2,j}|=1$, by considering the coefficients of

- $\frac{\partial\psi}{\partial\beta_1^{(u)}}(X;\beta_{1j}^*,\beta_{0j}^*)\pi(Y|h_2(X,\eta_j^*),\nu_j^*)$, we have $t_{1,j}^{(u)}=0$ for all $u\in[d]$;
- $\frac{\partial\psi}{\partial\beta_0}(X;\beta_{1j}^*,\beta_{0j}^*)\pi(Y|h_2(X,\eta_j^*),\nu_j^*)$, we have $t_{0,j}=0$;
- $\psi(X;\beta_{1j}^*,\beta_{0j}^*)\frac{\partial\pi}{\partial h_2}(Y|h_2(X,\eta_j^*),\nu_j^*)$, we have $\sum_{u_2=1}^{d_2}t_{2,j}^{(u_2)}\frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\eta_j^*)=0$. Since the expert function h_2 is strongly identifiable, we deduce $t_{2,j}^{(u_2)}=0$ for all $u_2\in[d_2]$;
- $\psi(X;\beta_{1j}^*,\beta_{0j}^*)\frac{\partial^2\pi}{\partial h_2^2}(Y|h_2(X,\eta_j^*),\nu_j^*)$, we have $t_{3,j}=0$.

For $j\in[k_2^*]$ such that $|\mathcal{V}_{2,j}|>1$, by considering the coefficients of

- $\pi(Y|h_2(X,\eta_j^*),\nu_j^*)$, we have $t_{0,j}=0$;
- $\frac{\partial\psi}{\partial\beta_1^{(u)}}(X;0_d,\bar{\beta}_{0i})\pi(Y|h_2(X,\eta_j^*),\nu_j^*)$, we have $t_{1,j,i}^{(u)}=0$ for all $u\in[d]$ and $i\in\mathcal{V}_{2,j}$;
- $\frac{\partial^2\psi}{\partial\beta_1^{(u)}\partial\beta_1^{(v)}}(X;0_d,\bar{\beta}_{0i})\pi(Y|h_2(X,\eta_j^*),\nu_j^*)$, we have $t_{4,j,i}^{(uv)}=0$ for all $u,v\in[d]$ and $i\in\mathcal{V}_{2,j}$;
- $\psi(X;0_d,\bar{\beta}_{0i})\frac{\partial\pi}{\partial h_2}(Y|h_2(X,\eta_j^*),\nu_j^*)$, we have

$$\sum_{u_2=1}^{d_2}t_{2,j,i}^{(u_2)}\frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\eta_j^*)+\sum_{u_2,v_2=1}^{d_2}t_{5,j,i}^{(u_2v_2)}\frac{\partial^2h_2}{\partial\eta^{(u_2)}\partial\eta^{(v_2)}}(X,\eta_j^*)=0.$$

  As the expert function h_2 satisfies the strong identifiability condition, we deduce $t_{2,j,i}^{(u_2)}=t_{5,j,i}^{(u_2v_2)}=0$ for all $u_2,v_2\in[d_2]$ and $i\in\mathcal{V}_{2,j}$;
- $\frac{\partial\psi}{\partial\beta_1^{(u)}}(X;0_d,\bar{\beta}_{0i})\frac{\partial\pi}{\partial h_2}(Y|h_2(X,\eta_j^*),\nu_j^*)$, we have $\sum_{u_2=1}^{d_2}t_{7,j,i}^{(uu_2)}\frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\eta_j^*)=0$. Since the expert function h_2 is strongly identifiable, we deduce $t_{7,j,i}^{(uu_2)}=0$ for all $u\in[d]$, $u_2\in[d_2]$ and $i\in\mathcal{V}_{2,j}$;
- $\psi(X;0_d,\bar{\beta}_{0i})\frac{\partial^2\pi}{\partial h_2^2}(Y|h_2(X,\eta_j^*),\nu_j^*)$, we have

$$\frac{1}{2}t_{3,j,i}+\sum_{u_2,v_2=1}^{d_2}\frac{t_{5,j,i}^{(u_2v_2)}}{1+1_{\{u_2=v_2\}}}\frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\eta_j^*)\frac{\partial h_2}{\partial\eta^{(v_2)}}(X,\eta_j^*)=0.$$

  Note that $t_{5,j,i}^{(u_2v_2)}=0$ for all $u_2,v_2\in[d_2]$ and $i\in\mathcal{V}_{2,j}$, we deduce $t_{3,j,i}=0$ for all $i\in\mathcal{V}_{2,j}$;
- $\frac{\partial\psi}{\partial\beta_1^{(u)}}(X;0_d,\bar{\beta}_{0i})\frac{\partial^2\pi}{\partial h_2^2}(Y|h_2(X,\eta_j^*),\nu_j^*)$, we have $t_{8,j,i}^{(u)}=0$ for all $u\in[d]$ and $i\in\mathcal{V}_{2,j}$;

- $\psi(X; 0_d, \bar{\beta}_{0i})\frac{\partial^3 \pi}{\partial h_2^3}(Y|h_2(X, \eta_j^*), \nu_j^*)$, we have $\sum_{u_2=1}^{d_2} \frac{1}{2} t_{9,j,i}^{(u_2)} \frac{\partial h_2}{\partial \eta^{(u_2)}}(X, \eta_j^*) = 0$. Since the expert function h_2 meets the strong identifiability, we deduce $t_{9,j,i}^{(u_2)}$ for all $u_2 \in [d_2]$ and $i \in \mathcal{V}_{2,j}$;
- $\psi(X; 0_d, \bar{\beta}_{0i})\frac{\partial^4 \pi}{\partial h_2^4}(Y|h_2(X, \eta_j^*), \nu_j^*)$, we have $t_{6,j,i} = 0$ for all $i \in \mathcal{V}_{2,j}$.

Putting the above results together, we have (i) $s_{0,j} = s_{1,j}^{(u_1)} = s_{2,j} = s_{3,j}^{(u_1 v_1)} = s_{4,j} = s_{5,j}^{(u_1)} = 0$ for all $j \in [k_1^*]$ and $u_1, v_1 \in [d_1]$; (ii) $t_{0,j} = t_{1,j}^{(u)} = t_{2,j}^{(u_2)} = t_{3,j} = 0$ for all $j \in [k_2^*] : |\mathcal{V}_{2,j}| = 1$, $u \in [d]$ and $u_2 \in [d_2]$; (iii) $t_{0,j} = t_{1,j,i}^{(u)} = t_{2,j,i}^{(u_2)} = t_{3,j,i} = t_{4,j,i}^{(uv)} = t_{5,j,i}^{(u_2 v_2)} = t_{6,j,i} = t_{7,j,i}^{uv_2} = t_{8,j,i}^{(u)} = t_{9,j,i}^{(u_2)}$ for all $j \in [k_2^*] : |\mathcal{V}_{2,j}| > 1$, $u, v \in [d]$ and $u_2, v_2 \in [d_2]$. This contradicts to the fact that at least one among them is non-zero. Consequently, we achieve the local part in equation (30) and complete the proof.

D.4 Proof of Theorem 4

Note that it is sufficient to demonstrate that

$$\inf_{(G_1,G_2)\in\mathcal{G}_{k_1,k_2}(\Theta)} \frac{\mathbb{E}_X[V(g_{G_1,G_2}(\cdot|X), g_{G_1^*,\check{G}_2}(\cdot|X))]}{\mathcal{D}_4((G_1,G_2),(G_1^*,\check{G}_2))} > 0,$$

for any pair of mixing measures $(G_1^*, \check{G}_2) \in \check{\mathcal{G}}_{k_1^*,k_2}(\Theta)$. For that purpose, given an arbitrary mixing measure $\check{G}_2 := \sum_{i=1}^{k_2} \sigma(\check{\beta}_{0i})\delta_{(\check{\beta}_{1i},\check{\eta}_i,\check{\nu}_i)}$, we need to establish its local part

$$\lim_{\varepsilon\to 0} \inf_{(G_1,G_2)\in\mathcal{G}_{k_1,k_2}(\Theta):\mathcal{D}_4((G_1,G_2),(G_1^*,\check{G}_2))\leq\varepsilon} \frac{\mathbb{E}_X[V(g_{G_1,G_2}(\cdot|X), g_{G_1^*,\check{G}_2}(\cdot|X))]}{\mathcal{D}_4((G_1,G_2),(G_1^*,\check{G}_2))} > 0, \tag{35}$$

and its global part

$$\inf_{(G_1,G_2)\in\mathcal{G}_{k_1,k_2}(\Theta):\mathcal{D}_4((G_1,G_2),(G_1^*,\check{G}_2))>\varepsilon'} \frac{\mathbb{E}_X[V(g_{G_1,G_2}(\cdot|X), g_{G_1^*,\check{G}_2}(\cdot|X))]}{\mathcal{D}_4((G_1,G_2),(G_1^*,\check{G}_2))} > 0. \tag{36}$$

Since the global part (36) can be demonstrated analogously to that in Appendix D.1, we will focus only on proving the local part (35) in this appendix. Assume by contrary that the above local part is not true. Then, we can find a sequence (G_1^n, G_2^n) of the form $G_1^n := \sum_{i=1}^{k_1^n} \omega_i^n \delta_{(\kappa_i^n, \tau_i^n)}$, $G_2^n := \sum_{i=1}^{k_2^n} \sigma(\beta_{0i}^n)\delta_{(\beta_{1i}^n,\eta_i^n,\nu_i^n)}$ for $n \in \mathbb{N}$ satisfying $\mathcal{D}_{4n} := \mathcal{D}_4((G_1^n, G_2^n),(G_1^*,\check{G}_2)) \to 0$ and

$$\mathbb{E}_X[V(g_{G_1^n,G_2^n}(\cdot|X), g_{G_1^*,\check{G}_2}(\cdot|X))]/\mathcal{D}_{4n} \to 0, \tag{37}$$

as $n \to \infty$. Moreover, we may assume WLOG that the number of shared experts k_1^n, the number of routed experts k_2^n, and Voronoi cells $\mathcal{V}_{1,j} = \mathcal{V}_{1,j}(G_1^n)$, $\mathcal{V}_{2,j} = \mathcal{V}_{2,j}(G_2^n)$ are independent of the sample size n. In addition, since G_2^n and $\check{G}_2$ have the same number of atoms k_2, we may assume WLOG that the Voronoi cell $\mathcal{V}_{2,j}$ admits only one element, that is, $\mathcal{V}_{2,j} = \{j\}$ for all $j \in [k_2]$. Thus, we can represent the Voronoi loss $\mathcal{D}_{4n}$ as

$$\mathcal{D}_{4n} = \sum_{j=1}^{k_1^*} \Big| \sum_{i\in\mathcal{V}_{1,j}} \omega_i^n - \omega_j^* \Big| + \sum_{i=1}^{k_2^*} (\|\Delta\check{\beta}_{1i}^n\| + |\Delta\check{\beta}_{0i}^n| + \|\Delta\check{\eta}_i^n\| + |\Delta\check{\nu}_i^n|)$$

$$+\sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|=1}\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n(\|\Delta\kappa_{ij}^n\|+|\Delta\tau_{ij}^n|)+\sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|>1}\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n(\|\Delta\kappa_{ij}^n\|^2+|\Delta\tau_{ij}^n|^2), \quad (38)$$

where we denote $\Delta\check{\beta}_{1i}^n := \beta_{1i}^n - \check{\beta}_{1i}$, $\Delta\check{\beta}_{0i}^n := \beta_{0i}^n - \check{\beta}_{0i}$, $\Delta\check{\eta}_i^n := \eta_i^n - \check{\eta}_i$, and $\Delta\check{\nu}_i^n := \nu_i^n - \check{\nu}_i$ for all $i \in [k_2]$. Recall that $\mathcal{D}_{4n} \to 0$ as $n \to \infty$, then equation (38) implies that as $n \to \infty$, we have

- For $j \in [k_1^*]$ and $i \in \mathcal{V}_{1,j}$: $\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n \to \omega_j^*$, $(\kappa_i^n, \tau_i^n) \to (\kappa_j^*, \tau_j^*)$;
- For $i \in [k_2^*]$: $(\beta_{1i}^n, \beta_{0i}^n, \eta_i^n, \nu_i^n) \to (\check{\beta}_{1i}, \check{\beta}_{0i}, \check{\eta}_i, \check{\nu}_i)$.

Now, we divide the proof into three main stages:

Stage 1 - Density Decomposition: In this step, we reuse the following decomposition of the density discrepancy $g_{G_1^n,G_2^n}(Y|X) - g_{G_1^*,G_2^*}(Y|X)$ in Appendix D.3

$$g_{G_1^n,G_2^n}(Y|X) - g_{G_1^*,G_2^*}(Y|X) = \frac{1}{2}\left[(q_{G_1^n}(Y|X) - q_{G_1^*}(Y|X)) + (p_{G_2^n}(Y|X) - p_{G_2^*}(Y|X))\right],$$

where we denote

$$\begin{aligned}
q_{G_1^n}(Y|X) &:= \sum_{i=1}^{k_1^n}\omega_i^n\pi(Y|h_1(X,\kappa_i^n),\tau_i^n),\\
q_{G_1^*}(Y|X) &:= \sum_{i=1}^{k_1^*}\omega_i^*\pi(Y|h_1(X,\kappa_i^*),\tau_i^*),\\
p_{G_2^n}(Y|X) &:= \sum_{i=1}^{k_2^n}\frac{\sigma((\beta_{1i}^n)^\top X+\beta_{0i}^n)}{\sum_{j=1}^{k_2^n}\sigma((\beta_{1j}^n)^\top X+\beta_{0j}^n)}\cdot\pi(Y|h_2(X,\eta_i^n),\nu_i^n),\\
p_{\check{G}_2}(Y|X) &:= \sum_{i=1}^{k_2}\frac{\sigma((\check{\beta}_{1i})^\top X+\check{\beta}_{0i})}{\sum_{j=1}^{k_2}\sigma((\check{\beta}_{1j})^\top X+\check{\beta}_{0j})}\cdot\pi(Y|h_2(X,\check{\eta}_i),\check{\nu}_i).
\end{aligned}$$

Stage 1.1: We also utilize the decomposition of the term $q_{G_1^n}(Y|X) - q_{G_1^*}(Y|X)$ in Appendix D.3 as follows:

$$\begin{aligned}
q_{G_1^n}(Y|X) - q_{G_1^*}(Y|X) &= \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|=1}\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n[\pi(Y|h_1(X,\kappa_i^n),\tau_i^n) - \pi(Y|h_1(X,\kappa_j^*),\tau_j^*)]\\
&+\sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|>1}\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n[\pi(Y|h_1(X,\kappa_i^n),\tau_i^n) - \pi(Y|h_1(X,\kappa_j^*),\tau_j^*)]\\
&+\sum_{j=1}^{k_1^*}\Big(\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n - \omega_j^*\Big)\pi(Y|h_1(X,\kappa_j^*),\tau_j^*)\\
&:= A_{n,1}(Y|X) + A_{n,2}(Y|X) + A_{n,0}(Y|X).
\end{aligned}$$

Above, the quantity $A_{n,1}(Y|X)$ is expanded as

$$A_{n,1}(Y|X) = \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|=1}\sum_{\rho=1}^{2}A_{n,1,\rho}^{(j)}(X)\frac{\partial^\rho\pi}{\partial h_1^\rho}(Y|h_1(X,\kappa_j^*),\tau_j^*) + R_{n,1}(Y|X),$$

where $R_{n,1}(Y|X)$ is a Taylor remainder such that $R_{n,1}(Y|X)/\mathcal{D}_{4n} \to$ as $n \to \infty$, and

$$A^{(j)}_{n,1,1}(X) := \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n \sum_{u_1=1}^{d_1} (\Delta \kappa_{ij}^n)^{(u_1)} \frac{\partial h_1}{\partial \kappa^{(u_1)}}(X, \kappa_j^*),$$
$$A^{(j)}_{n,1,2}(X) := \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n \frac{1}{2}(\Delta \tau_{ij}^n),$$

for all $j \in [k_1^*]$ such that $|\mathcal{V}_{1,j}| = 1$. In addition, we can rewrite $A_{n,2}(Y|X)$ as

$$A_{n,2}(Y|X) = \sum_{j \in [k_1^*]: |\mathcal{V}_{1,j}| > 1} \sum_{\rho=1}^{4} A^{(j)}_{n,1,\rho}(X) \frac{\partial^\rho \pi}{\partial h_1^\rho}(Y|h_1(X, \kappa_j^*), \tau_j^*) + R_{n,2}(Y|X),$$

where $R_{n,2}(Y|X)$ is a Taylor remainder such that $R_{n,2}(Y|X)/\mathcal{D}_{4n} \to$ as $n \to \infty$, and

$$A^{(j)}_{n,2,1}(X) := \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n \Big(\sum_{u_1=1}^{d_1} (\Delta \kappa_{ij}^n)^{(u_1)} \frac{\partial h_1}{\partial \kappa^{(u_1)}}(X, \kappa_j^*) + \sum_{u_1,v_1=1}^{d_1} \frac{(\Delta \kappa_{ij}^n)^{(u_1)} (\Delta \kappa_{ij}^n)^{(v_1)}}{1 + 1_{\{u_1 = v_1\}}} \frac{\partial^2 h_1}{\partial \kappa^{(u_1)} \partial \kappa^{(v_1)}}(X, \kappa_j^*) \Big),$$
$$A^{(j)}_{n,2,2}(X) := \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n \Big(\frac{1}{2}(\Delta \tau_{ij}^n) + \sum_{u_1,v_1=1}^{d_1} \frac{(\Delta \kappa_{ij}^n)^{(u_1)} (\Delta \kappa_{ij}^n)^{(v_1)}}{1 + 1_{\{u_1 = v_1\}}} \frac{\partial h_1}{\partial \kappa^{(u_1)}}(X, \kappa_j^*) \frac{\partial h_1}{\partial \kappa^{(v_1)}}(X, \kappa_j^*) \Big),$$
$$A^{(j)}_{n,2,3}(X) := \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n \sum_{u_1=1}^{d_1} \frac{1}{2} (\Delta \kappa_{ij}^n)^{(u_1)} (\Delta \tau_{ij}^n) \frac{\partial h_1}{\partial \kappa^{(u_1)}}(X, \kappa_j^*),$$
$$A^{(j)}_{n,2,4}(X) := \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n \frac{1}{8} (\Delta \tau_{ij}^n)^2,$$

for all $j \in [k_1^*]$ such that $|\mathcal{V}_{1,j}| > 1$.

Stage 1.2: Next, we decompose the term $Q_n(Y|X) := \Big[\sum_{j=1}^{k_2} \sigma((\check{\beta}_{1j})^\top X + \check{\beta}_{0j})\Big] \cdot [p_{G_2^n}(Y|X) - p_{\check{G}_2}(Y|X)]$ as

$$\begin{aligned}
Q_n(Y|X) &= \sum_{i=1}^{k_2} \Big[\sigma((\beta_{1i}^n)^\top X + \beta_{0i}^n) \pi(Y|h_2(X, \eta_i^n), \nu_i^n) - \sigma((\check{\beta}_{1i})^\top X + \check{\beta}_{0i}) \pi(Y|h_2(X, \check{\eta}_i), \check{\nu}_i) \Big] \\
&\quad - \sum_{i=1}^{k_2} \Big[\sigma((\beta_{1i}^n)^\top X + \beta_{0i}^n) - \sigma((\check{\beta}_{1i})^\top X + \check{\beta}_{0i}) \Big] p_{G_2^n}(Y|X) \\
&= \sum_{i=1}^{k_2} \Big[\psi(X; \beta_{1i}^n, \beta_{0i}^n) \pi(Y|h_2(X, \eta_i^n), \nu_i^n) - \psi(X; \check{\beta}_{1i}, \check{\beta}_{0i}) \pi(Y|h_2(X, \check{\eta}_i), \check{\nu}_i) \Big] \\
&\quad - \sum_{i=1}^{k_2} \Big[\psi(X; \beta_{1i}^n, \beta_{0i}^n) - \psi(X; \check{\beta}_{1i}, \check{\beta}_{0i}) \Big] p_{G_2^n}(Y|X) \\
&:= B_n(Y|X) - C_n(Y|X),
\end{aligned}$$

where we denote $\psi(X;\beta_1,\beta_0) := \sigma(\beta_1^\top X + \beta_0)$.

Stage 1.2.1: In this step, we decompose $B_n(Y|X)$ by applying the first-order Taylor expansion to the function $\psi(X;\beta_{1i}^n,\beta_{0i}^n)\pi(Y|h_2(X,\eta_i^n),\nu_i^n)$ around the point $(\check{\beta}_{1i},\check{\beta}_{0i},\check{\eta}_i,\check{\nu}_i)$ as follows:

$$\begin{aligned}B_n(Y|X) &= \sum_{i=1}^{k_2}\sum_{|\alpha|=1}(\Delta\check{\beta}_{1i}^n)^{\alpha_1}(\Delta\check{\beta}_{0i}^n)^{\alpha_2}(\Delta\check{\eta}_i^n)^{\alpha_3}(\Delta\check{\nu}_i^n)^{\alpha_4}\\
&\qquad\qquad\times\frac{\partial^{|\alpha_1|+\alpha_2}\psi}{\partial\beta_1^{\alpha_1}\partial\beta_0^{\alpha_2}}(X;\check{\beta}_{1i},\check{\beta}_{0i})\frac{\partial^{|\alpha_3|+\alpha_4}\pi}{\partial\eta^{\alpha_3}\partial\nu^{\alpha_4}}(Y|h_2(X,\check{\eta}_i),\check{\nu}_i) + R_{n,3}(Y|X)\\
&= \sum_{i=1}^{k_2}\sum_{\rho=0}^{2}B_{n,\rho}^{(i)}(X)\frac{\partial^\rho\pi}{\partial h_2^\rho}(Y|h_2(X,\check{\eta}_i),\check{\nu}_i) + R_{n,3}(Y|X),\end{aligned}$$

where $R_{n,3}(Y|X)$ is a Taylor remainder such that $R_{n,3}(Y|X)/\mathcal{D}_{4n} \to$ as $n\to\infty$, and

$$\begin{aligned}B_{n,0}^{(i)} &:= \sum_{u=1}^{d}(\Delta\check{\beta}_{1i}^n)^{(u)}\frac{\partial\psi}{\partial\beta_1^{(u)}}(X;\check{\beta}_{1i},\check{\beta}_{0i}) + (\Delta\check{\beta}_{0i}^n)\frac{\partial\psi}{\partial\beta_0}(X;\check{\beta}_{1i},\check{\beta}_{0i}),\\
B_{n,1}^{(i)} &:= \sum_{u_2=1}^{d_2}(\Delta\check{\eta}_i^n)^{(u_2)}\frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\check{\eta}_i)\psi(X;\check{\beta}_{1i},\check{\beta}_{0i}),\\
B_{n,2}^{(i)} &:= \frac{1}{2}(\Delta\check{\nu}_i^n)\psi(X;\check{\beta}_{1i},\check{\beta}_{0i}),\end{aligned}$$

for all $i\in[k_2]$.

Stage 1.2.2: Next, we proceed to decompose $C_n(Y|X)$ by applying the first-order Taylor expansion to the function $\psi(X;\beta_{1i}^n,\beta_{0i}^n)$ around the point $(\check{\beta}_{1i},\check{\beta}_{0i})$ as

$$\begin{aligned}C_n(Y|X) &= \sum_{i=1}^{k_2}\sum_{|\alpha|=1}(\Delta\check{\beta}_{1i}^n)^{\alpha_1}(\Delta\check{\beta}_{0i}^n)^{\alpha_2}\frac{\partial^{|\alpha_1|+\alpha_2}\psi}{\partial\beta_1^{\alpha_1}\partial\beta_0^{\alpha_2}}(X;\check{\beta}_{1i},\check{\beta}_{0i})p_{G_2^n}(Y|X) + R_{n,4}(Y|X)\\
&= \sum_{i=1}^{k_2}\Big[\sum_{u=1}^{d}(\Delta\check{\beta}_{1i}^n)^{(u)}\frac{\partial\psi}{\partial\beta_1^{(u)}}(X;\check{\beta}_{1i},\check{\beta}_{0i}) + (\Delta\check{\beta}_{0i}^n)\frac{\partial\psi}{\partial\beta_0}(X;\check{\beta}_{1i},\check{\beta}_{0i})\Big]p_{G_2^n}(Y|X) + R_{n,4}(Y|X),\end{aligned}$$

where $R_{n,4}(Y|X)$ is a Taylor remainder such that $R_{n,4}(Y|X)/\mathcal{D}_{4n}\to$ as $n\to\infty$.

Combining the above decompositions, we can view $A_{n,0}(Y|X)/\mathcal{D}_{4n}$, $[A_{n,1}(Y|X) - R_{n,1}(Y|X)]/\mathcal{D}_{4n}$, $[A_{n,2}(Y|X) - R_{n,2}(Y|X)]/\mathcal{D}_{4n}$, $[B_n(Y|X) - R_{n,3}(Y|X)]/\mathcal{D}_{4n}$, $[C_n(Y|X) - R_{n,4}(Y|X)]/\mathcal{D}_{4n}$ as a combination of elements from the following sets

$$\begin{aligned}&\mathcal{S}_{0,j} := \{\pi(Y|h_1(X,\kappa_j^*),\tau_j^*)\},\\
&\mathcal{S}_{1,j} := \left\{\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*)\frac{\partial\pi}{\partial h_1}(Y|h_1(X,\kappa_j^*),\tau_j^*),\ \frac{\partial^2 h_1}{\partial\kappa^{(u_1)}\partial\kappa^{(v_1)}}(X,\kappa_j^*)\frac{\partial\pi}{\partial h_1}(Y|h_1(X,\kappa_j^*),\tau_j^*) : u_1,v_1\in[d_1]\right\},\\
&\mathcal{S}_{2,j} := \left\{\frac{\partial^2\pi}{\partial h_1^2}(Y|h_1(X,\kappa_j^*),\tau_j^*),\ \frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*)\frac{\partial h_1}{\partial\kappa^{(v_1)}}(X,\kappa_j^*)\frac{\partial^2\pi}{\partial h_1^2}(Y|h_1(X,\kappa_j^*),\tau_j^*) : u_1,v_1\in[d_1]\right\},\end{aligned}$$

$$\mathcal{S}_{3,j} := \left\{ \frac{\partial h_1}{\partial \kappa^{(u_1)}}(X, \kappa_j^*) \frac{\partial^3 \pi}{\partial h_1^3}(Y|h_1(X, \kappa_j^*), \tau_j^*) : u_1, v_1 \in [d_1] \right\},$$

$$\mathcal{S}_{4,j} := \left\{ \frac{\partial^4 \pi}{\partial h_1^4}(Y|h_1(X, \kappa_j^*), \tau_j^*) : u_1, v_1 \in [d_1] \right\},$$

for all $j \in [k_1^*]$, and

$$\begin{aligned}
\mathcal{T}_{0,j} := \Bigg\{ & \frac{\partial \psi}{\partial \beta_1^{(u)}}(X; \check{\beta}_{1i}, \check{\beta}_{0i}) \pi(Y|h_2(X, \check{\eta}_i), \check{\nu}_i), \ \frac{\partial \psi}{\partial \beta_0}(X; \check{\beta}_{1i}, \check{\beta}_{0i}) \pi(Y|h_2(X, \check{\eta}_i), \check{\nu}_i), \\
& \frac{\partial \psi}{\partial \beta_1^{(u)}}(X; \check{\beta}_{1i}, \check{\beta}_{0i}) p_{G_2^n}(Y|X), \ \frac{\partial \psi}{\partial \beta_0}(X; \check{\beta}_{1i}, \check{\beta}_{0i}) p_{G_2^n}(Y|X) : u \in [d] \Bigg\}, \\
\mathcal{T}_{1,j} := \Bigg\{ & \frac{\partial h_2}{\partial \eta^{(u_2)}}(X, \eta_j^*) \psi(X; \check{\beta}_{1i}, \check{\beta}_{0i}) \frac{\partial \pi}{\partial h_2}(Y|h_2(X, \check{\eta}_i), \check{\nu}_i) : u \in [d], \ u_2 \in [d_2] \Bigg\}, \\
\mathcal{T}_{2,j} := \Bigg\{ & \psi(X; \check{\beta}_{1i}, \check{\beta}_{0i}) \frac{\partial^2 \pi}{\partial h_2^2}(Y|h_2(X, \check{\eta}_i), \check{\nu}_i) \Bigg\},
\end{aligned}$$

for all $j \in [k_2^*]$.

Stage 2 - Non-vanishing coefficients: In this stage, we show that at least one among the coefficients in the representations of $A_{n,0}(Y|X)/\mathcal{D}_{4n}$, $[A_{n,1}(Y|X) - R_{n,1}(Y|X)]/\mathcal{D}_{4n}$, $[A_{n,2}(Y|X) - R_{n,2}(Y|X)]/\mathcal{D}_{4n}$, $[B_n(Y|X) - R_{n,3}(Y|X)]/\mathcal{D}_{4n}$, $[C_n(Y|X) - R_{n,4}(Y|X)]/\mathcal{D}_{4n}$ does not converge to zero when $n \to \infty$. Suppose that all these coefficients go to zero. By using the same arguments as in Stage 2 in Appendix D.1, we have

$$\begin{aligned}
\frac{1}{\mathcal{D}_{4n}} \Big[\sum_{j=1}^{k_1^*} \Big| \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n - \omega_j^* \Big| + \sum_{j \in [k_1^*] : |\mathcal{V}_{1,j}| = 1} \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n (\|\Delta \kappa_{ij}^n\| + |\Delta \tau_{ij}^n|) \\
+ \sum_{j \in [k_1^*] : |\mathcal{V}_{1,j}| > 1} \sum_{i \in \mathcal{V}_{1,j}} \omega_i^n (\|\Delta \kappa_{ij}^n\|^2 + |\Delta \tau_{ij}^n|^2) \Big] \to 0,
\end{aligned}$$

as $n \to \infty$. Additionally, by considering the coefficients of the terms:

- $\frac{\partial \psi}{\partial \beta_1^{(u)}}(X; \check{\beta}_{1i}, \check{\beta}_{0i}) \pi(Y|h_2(X, \check{\eta}_i), \check{\nu}_i)$ for $i \in [k_2]$, we get $\frac{1}{\mathcal{D}_{4n}} \sum_{i=1}^{k_2} \|\Delta \beta_{1ij}^n\| \to 0$;
- $\frac{\partial \psi}{\partial \beta_0}(X; \check{\beta}_{1i}, \check{\beta}_{0i}) \pi(Y|h_2(X, \check{\eta}_i), \check{\nu}_i)$ for $i \in [k_2]$, we get $\frac{1}{\mathcal{D}_{4n}} \sum_{i=1}^{k_2} |\Delta \beta_{0ij}^n| \to 0$;
- $\frac{\partial h_2}{\partial \eta^{(u_2)}}(X, \check{\eta}_i) \psi(X; \check{\beta}_{1i}, \check{\beta}_{0i}) \frac{\partial \pi}{\partial h_2}(Y|h_2(X, \check{\eta}_i), \check{\nu}_i)$ for $i \in [k_2]$, we get $\frac{1}{\mathcal{D}_{4n}} \sum_{i=1}^{k_2} \|\Delta \eta_{ij}^n\| \to 0$;
- $\psi(X; \check{\beta}_{1i}, \check{\beta}_{0i}) \frac{\partial \pi}{\partial h_2}(Y|h_2(X, \check{\eta}_i), \check{\nu}_i)$ for $i \in [k_2]$, we get $\frac{1}{\mathcal{D}_{4n}} \sum_{i=1}^{k_2} |\Delta \nu_{ij}^n| \to 0$.

Taking the summation of the above limits, we deduce $1 = \frac{\mathcal{D}_{4n}}{\mathcal{D}_{4n}} \to 0$ as $n \to \infty$, which is a contradiction. Thus, not all the coefficients in the representations of $A_{n,0}(Y|X)/\mathcal{D}_{4n}$, $[A_{n,1}(Y|X) - R_{n,1}(Y|X)]/\mathcal{D}_{4n}$, $[A_{n,2}(Y|X) - R_{n,2}(Y|X)]/\mathcal{D}_{4n}$, $[B_n(Y|X) - R_{n,3}(Y|X)]/\mathcal{D}_{4n}$, $[C_n(Y|X) - R_{n,4}(Y|X)]/\mathcal{D}_{4n}$ converge to zero as $n \to \infty$.

Stage 3 - Fatou's lemma contradiction: In this stage, we attempto to show a contradiction to the result of Stage 2 using the Fatou's lemma. Firstly, we denote m_n as the maximum of the absolute values of the coefficients in the representations of $A_{n,0}(Y|X)/\mathcal{D}_{4n}$, $[A_{n,1}(Y|X) - R_{n,1}(Y|X)]/\mathcal{D}_{4n}$, $[A_{n,2}(Y|X) - R_{n,2}(Y|X)]/\mathcal{D}_{4n}$, $[B_n(Y|X) - R_{n,3}(Y|X)]/\mathcal{D}_{4n}$, $[C_n(Y|X) - R_{n,4}(Y|X)]/\mathcal{D}_{4n}$. The result of Stage 2 implies that $1/m_n \not\to \infty$ as $n \to \infty$. In addition, we also denote

$$\frac{1}{m_n\mathcal{D}_{4n}} \cdot \sum_{i\in\mathcal{V}_{1,j}} \omega_i^n(\Delta\kappa_{ij}^n)^{(u_1)} \to s_{1,j}^{(u_1)}, \quad \frac{1}{m_n\mathcal{D}_{4n}} \cdot \sum_{i\in\mathcal{V}_{1,j}} \omega_i^n(\Delta\tau_{ij}^n) \to s_{2,j},$$
$$\frac{1}{m_n\mathcal{D}_{4n}} \cdot \sum_{i\in\mathcal{V}_{1,j}} \omega_i^n(\Delta\kappa_{ij}^n)^{(u_1)}(\Delta\kappa_{ij}^n)^{(v_1)} \to s_{3,j}^{(u_1v_1)}, \quad \frac{1}{m_n\mathcal{D}_{4n}} \cdot \sum_{i\in\mathcal{V}_{1,j}} \omega_i^n(\Delta\tau_{ij}^n)^2 \to s_{4,j},$$
$$\frac{1}{m_n\mathcal{D}_{4n}} \cdot \sum_{i\in\mathcal{V}_{1,j}} \omega_i^n(\Delta\kappa_{ij}^n)^{(u_1)}(\Delta\tau_{ij}^n) \to s_{5,j}^{(u_1)}, \quad \frac{1}{m_n\mathcal{D}_{4n}} \cdot \Big(\sum_{i\in\mathcal{V}_{1,j}} \omega_i^n - \omega_j^*\Big) \to s_{0,j},$$

for all $j \in [k_1^*]$ and

$$\frac{1}{m_n\mathcal{D}_{4n}} \cdot (\Delta\check{\beta}_{0i}^n) \to t_{0,i}, \quad \frac{1}{m_n\mathcal{D}_{4n}} \cdot (\Delta\check{\beta}_{1i}^n)^{(u)} \to t_{1,i}^{(u)},$$
$$\frac{1}{m_n\mathcal{D}_{4n}} \cdot (\Delta\check{\eta}_i^n)^{(u_2)} \to t_{2,i}^{(u_2)}, \quad \frac{1}{m_n\mathcal{D}_{4n}} \cdot (\Delta\check{\nu}_i^n) \to t_{3,i},$$

for all $i \in [k_2]$. Due to the result of Stage 2, at least one among the above limits is different from zero. Recall from equation (37) that we have

$$\mathbb{E}_X[V(g_{G_1^n,G_2^n}(\cdot|X), g_{G_1^*,\check{G}_2}(\cdot|X))]/\mathcal{D}_{4n} \to 0,$$

Furthermore, according to the Fatou's lemma, we get

$$\lim_{n\to\infty} \frac{\mathbb{E}_X[V(g_{G_1^n,G_2^n}(\cdot|X), g_{G_1^*,\check{G}_2}(\cdot|X))]}{m_n\mathcal{D}_{4n}} \geq \int \liminf_{n\to\infty} \frac{|g_{G_1^n,G_2^n}(Y|X) - g_{G_1^*,\check{G}_2}(Y|X)|}{2m_n\mathcal{D}_{4n}} \mathrm{d}(X,Y).$$

Then, it follows that $[g_{G_1^n,G_2^n}(Y|X) - g_{G_1^*,\check{G}_2}(Y|X)]/[m_n\mathcal{D}_{4n}] \to 0$ as $n \to \infty$ for almost surely (X,Y). As the input space is bounded and the parameter space is compact, the quantity $\sum_{j=1}^{k_2} \sigma((\check{\beta}_{1j})^\top X + \check{\beta}_{0j})$ is bounded. Therefore, we deduce

$$\Big[\sum_{j=1}^{k_2} \sigma((\check{\beta}_{1j})^\top X + \check{\beta}_{0j})\Big][g_{G_1^n,G_2^n}(Y|X) - g_{G_1^*,\check{G}_2}(Y|X)]/[m_n\mathcal{D}_{4n}] \to 0,$$

as $n \to \infty$. This result indicates

$$\frac{1}{2}\Big[\sum_{j=1}^{k_2} \sigma((\check{\beta}_{1j})^\top X + \check{\beta}_{0j})\Big] \cdot \frac{q_{G_1^n}(Y|X) - q_{G_1^*}(Y|X)}{m_n\mathcal{D}_{4n}} + \frac{1}{2}\frac{Q_n(Y|X)}{m_n\mathcal{D}_{4n}} \to 0.$$

as $n \to \infty$ for almost surely (X, Y). From the decomposition of the terms $q_{G_1^n}(Y|X) - q_{G_1^*}(Y|X)$ and $Q_n(Y|X)$ in Stage 1, we have

$$\frac{1}{2}\Big[\sum_{j=1}^{k_2}\sigma((\check{\beta}_{1j})^\top X + \check{\beta}_{0j})\Big] \cdot \frac{A_{n,2}(Y|X) + A_{n,1}(Y|X) + A_{n,0}(Y|X)}{m_n\mathcal{D}_{4n}} + \frac{1}{2}\frac{B_n(Y|X) - C_n(Y|X)}{m_n\mathcal{D}_{4n}} \to 0. \tag{39}$$

We have

$$\lim_{n\to\infty}\frac{A_{n,0}(Y|X)}{m_n\mathcal{D}_{4n}} = \sum_{j=1}^{k_1^*} s_{0,j}\pi(Y|h_1(X,\kappa_j^*),\tau_j^*),$$

$$\lim_{n\to\infty}\frac{A_{n,1}(Y|X)}{m_n\mathcal{D}_{4n}} = \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|=1}\Big[\sum_{u_1=1}^{d_1} s_{1,j}^{(u_1)}\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*)\frac{\partial\pi}{\partial h_1}(Y|h_1(X,\kappa_j^*),\tau_j^*) + \frac{1}{2}s_{2,j}\frac{\partial^2\pi}{\partial h_1^2}(Y|h_1(X,\kappa_j^*),\tau_j^*)\Big],$$

$$\lim_{n\to\infty}\frac{A_{n,2}(Y|X)}{m_n\mathcal{D}_{4n}} = \sum_{j\in[k_1^*]:|\mathcal{V}_{1,j}|>1}\Big[\Big(\sum_{u_1=1}^{d_1} s_{1,j}^{(u_1)}\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*) + \sum_{u_1,v_1=1}^{d_1}\frac{s_{3,j}^{(u_1v_1)}}{1+1_{\{u_1=v_1\}}}\frac{\partial^2 h_1}{\partial\kappa^{(u_1)}\partial\kappa^{(v_1)}}(X,\kappa_j^*)\Big)$$
$$\times\frac{\partial\pi}{\partial h_1}(Y|h_1(X,\kappa_j^*),\tau_j^*) + \Big(\frac{1}{2}s_{2,j} + \sum_{u_1,v_1=1}^{d_1}\frac{s_{3,j}^{(u_1v_1)}}{1+1_{\{u_1=v_1\}}}\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*)\frac{\partial h_1}{\partial\kappa^{(v_1)}}(X,\kappa_j^*)\Big)\frac{\partial^2\pi}{\partial h_1^2}(Y|h_1(X,\kappa_j^*),\tau_j^*)$$
$$+\Big(\frac{1}{2}\sum_{u_1=1}^{d_1} s_{5,j}^{(u_1)}\frac{\partial h_1}{\partial\kappa^{(u_1)}}(X,\kappa_j^*)\Big)\frac{\partial^3\pi}{\partial h_1^3}(Y|h_1(X,\kappa_j^*),\tau_j^*) + \frac{1}{8}s_{4,j}\frac{\partial^4\pi}{\partial h_1^4}(Y|h_1(X,\kappa_j^*),\tau_j^*)\Big],$$

and

$$\lim_{n\to\infty}\frac{B_n(Y|X)}{m_n\mathcal{D}_{4n}} = \sum_{i=1}^{k_2}\Big[\Big(\sum_{u=1}^{d} t_{1,i}^{(u)}\frac{\partial\psi}{\partial\beta_1^{(u)}}(X;\check{\beta}_{1i},\check{\beta}_{0i}) + t_{0,i}\frac{\partial\psi}{\partial\beta_0}(X;\check{\beta}_{1i},\check{\beta}_{0i})\Big)\pi(Y|h_2(X,\check{\eta}_i),\check{\nu}_i)$$
$$+\sum_{u_2=1}^{d_2} t_{2,i}^{(u_2)}\frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\check{\eta}_i)\psi(X;\check{\beta}_{1i},\check{\beta}_{0i})\frac{\partial\pi}{\partial h_2}(Y|h_2(X,\check{\eta}_i),\check{\nu}_i)$$
$$+\frac{1}{2}(\Delta\check{\nu}_i^n)\psi(X;\check{\beta}_{1i},\check{\beta}_{0i})\frac{\partial^2\pi}{\partial h_2^2}(Y|h_2(X,\check{\eta}_i),\check{\nu}_i)\Big],$$

$$\lim_{n\to\infty}\frac{C_n(Y|X)}{m_n\mathcal{D}_{4n}} = \sum_{i=1}^{k_2}\Big[\sum_{u=1}^{d} t_{1,i}^{(u)}\frac{\partial\psi}{\partial\beta_1^{(u)}}(X;\check{\beta}_{1i},\check{\beta}_{0i}) + t_{0,i}\frac{\partial\psi}{\partial\beta_0}(X;\check{\beta}_{1i},\check{\beta}_{0i})\Big]p_{\check{G}_2}(Y|X).$$

Note that for almost every X, the set

$$\left\{\Big[\sum_{j=1}^{k_2}\sigma((\check{\beta}_{1j})^\top X + \check{\beta}_{0j})\Big]\frac{\partial^\rho\pi}{\partial h_1^\rho}(Y|h_1(X,\kappa_j^*),\tau_j^*) : 0 \le \rho \le 4,\ j \in [k_1^*]\right\}$$
$$\cup\left\{\frac{\partial\psi}{\partial\beta_1^{(u)}}(X;\check{\beta}_{1i},\check{\beta}_{0i})\pi(Y|h_2(X,\check{\eta}_i),\check{\nu}_i),\ \frac{\partial\psi}{\partial\beta_0}(X;\check{\beta}_{1i},\check{\beta}_{0i})\pi(Y|h_2(X,\check{\eta}_i),\check{\nu}_i),\right.$$

$$\frac{\partial\psi}{\partial\beta_1^{(u)}}(X;\check{\beta}_{1i},\check{\beta}_{0i})p_{\check{G}_2}(Y|X),\ \frac{\partial\psi}{\partial\beta_0}(X;\check{\beta}_{1i},\check{\beta}_{0i})p_{\check{G}_2}(Y|X),$$
$$\psi(X;\check{\beta}_{1i},\check{\beta}_{0i})\frac{\partial\pi}{\partial h_2}(Y|h_2(X,\check{\eta}_i),\check{\nu}_i),\ \psi(X;\check{\beta}_{1i},\check{\beta}_{0i})\frac{\partial^2\pi}{\partial h_2^2}(Y|h_2(X,\check{\eta}_i),\check{\nu}_i) : u\in[d],\ i\in[k_2]\Bigg\}$$

is linearly independent w.r.t Y, implying that the coefficients of those terms in the limit in equation (39) are equal to zero.

Since the expert function h_1 is strongly identifiable, then by employing the same arguments as in the Stage 3 of Appendix D.3, we get $s_{0,j}=s_{1,j}^{(u_1)}=s_{2,j}=s_{3,j}^{(u_1v_1)}=s_{4,j}=s_{5,j}^{(u_1)}=0$ for all $j\in[k_1^*]$ and $u_1,v_1\in[d_1]$. For $i\in[k_2]$, by considering the coefficients of

- $\frac{\partial\psi}{\partial\beta_1^{(u)}}(X;\check{\beta}_{1i},\check{\beta}_{0i})\pi(Y|h_2(X,\check{\eta}_i),\check{\nu}_i)$, we get $t_{1,i}^{(u)}=0$ for all $u\in[d]$;

- $\frac{\partial\psi}{\partial\beta_0}(X;\check{\beta}_{1i},\check{\beta}_{0i})\pi(Y|h_2(X,\check{\eta}_i),\check{\nu}_i)$, we get $t_{0,i}=0$;

- $\psi(X;\check{\beta}_{1i},\check{\beta}_{0i})\frac{\partial\pi}{\partial h_2}(Y|h_2(X,\check{\eta}_i),\check{\nu}_i)$, we get $\sum_{u_2=1}^{d_2}t_{2,i}^{(u_2)}\frac{\partial h_2}{\partial\eta^{(u_2)}}(X,\check{\eta}_i)=0$. Since the expert function h_2 is weakly identifiable, we deduce $t_{2,i}^{(u_2)}=0$ for all $u_2\in[d_2]$;

- $\psi(X;\check{\beta}_{1i},\check{\beta}_{0i})\frac{\partial^2\pi}{\partial h_2^2}(Y|h_2(X,\check{\eta}_i),\check{\nu}_i)$, we get $t_{3,i}=0$.

From the above results, it follows that (i) $s_{0,j}=s_{1,j}^{(u_1)}=s_{2,j}=s_{3,j}^{(u_1v_1)}=s_{4,j}=s_{5,j}^{(u_1)}=0$ for all $j\in[k_1^*]$ and $u_1,v_1\in[d_1]$; (ii) $t_{0,i}=t_{1,i}^{(u)}=t_{2,i}^{(u_2)}=t_{3,i}=0$ for all $i\in[k_2]$, $u\in[d]$ and $u_2\in[d_2]$. This contradicts to the fact that not all of them equal to zero. As a consequence, we obtain the local part in equation (35). Hence, the proof is completed.

E Proof of Auxiliary Results

E.1 Proof of Proposition 1

In this proof, we will leverage fundamental results on density estimation for M-estimators in [72]. Before streamlining our arguments, let us introduce some concepts from the empirical process theory adapted to the setting of the model (1).

Firstly, we denote by $\mathcal{F}_{k_1,k_2}(\Theta):=\{f_{G_1,G_2}(Y|X):(G_1,G_2)\in\mathcal{G}_{k_1,k_2}(\Theta)\}$ the set of conditional density functions of interest. Furthermore, we also consider two variants of this set defined as

$$\widetilde{\mathcal{F}}_{k_1,k_2}(\Theta):=\Big\{\frac{1}{2}f_{(G_1,G_2)}(Y|X)+\frac{1}{2}f_{(G_1,G_2)}(Y|X):(G_1^*,G_2^*)\in\mathcal{G}_{k_1,k_2}(\Theta)\Big\},$$
$$\widetilde{\mathcal{F}}^{1/2}_{k_1,k_2}(\Theta):=\{\tilde{f}^{1/2}:\tilde{f}\in\widetilde{\mathcal{F}}_{k_1,k_2}(\Theta)\}.$$

For any $\delta>0$, the Hellinger ball centered around the the true density $f_{G_1^*,G_2^*}(Y|X)$ and intersected with $\widetilde{\mathcal{F}}_{k_1,k_2}(\Theta)$ is defined as

$$\widetilde{\mathcal{F}}^{1/2}_{k_1,k_2}(\Theta,\delta):=\{p^{1/2}\in\widetilde{\mathcal{F}}^{1/2}_{k_1,k_2}(\Theta):h(p,f_{G_1^*,G_2^*})\leq\delta\}.$$

The size of the above Hellinger ball is determined by the quantity [72]

$$\mathcal{J}_B(\delta, \widetilde{\mathcal{F}}^{1/2}_{k_1,k_2}(\Theta,\delta), \|\cdot\|_2) := \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(t, \widetilde{\mathcal{F}}^{1/2}_{k_1,k_2}(\Theta,t), \|\cdot\|_2)\mathrm{d}t \vee \delta, \tag{40}$$

where $H_B(t, \widetilde{\mathcal{F}}^{1/2}_{k_1,k_2}(\Theta,t), \|\cdot\|_2)$ stands for the bracketing entropy of $\widetilde{\mathcal{F}}^{1/2}_{k_1,k_2}(\Theta,t)$ under the $L^2(m)$-norm with m being the Lebesgue measure, and $t \vee \delta := \max\{t,\delta\}$. Equipped with these notations, we are ready to present a standard result on density estimation for M-estimators in the following lemma:

Lemma 1 (Theorem 7.4, [72]). *Let $\delta \in (0,1)$ and take $\Psi(\delta) \geq \mathcal{J}_B(\delta, \widetilde{\mathcal{F}}^{1/2}_{k_1,k_2}(\Theta,\delta))$ such that $\Psi(\delta)/\delta^2$ is a non-increasing function of δ. Then, for a universal constant c and for some sequence (δ_n) satisfying $\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n)$, the following holds for all $\delta \geq \delta_n$:*

$$\mathbb{P}\Big(\mathbb{E}_X\Big[h(f_{\widetilde{G}_1^n,\widetilde{G}_2^n}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X)) > \delta\Big]\Big) \leq c\exp\Big(-\frac{n\delta^2}{c^2}\Big).$$

Given the above result, we will provide below the proof for Proposition 1.

Main proof of Proposition 1. Since $\widetilde{\mathcal{F}}^{1/2}_{k_1,k_2}(\Theta,t) \subset \widetilde{\mathcal{F}}^{1/2}_{k_1,k_2}(\Theta)$ for any $t > 0$, we have

$$H_B(t, \widetilde{\mathcal{F}}^{1/2}_{k_1,k_2}(\Theta,t), \|\cdot\|_2) \leq H_B(t, \widetilde{\mathcal{F}}^{1/2}_{k_1,k_2}(\Theta), \|\cdot\|_2) = H_B(t/\sqrt{2}, \widetilde{\mathcal{F}}_{k_1,k_2}(\Theta), h), \tag{41}$$

where the last equality is due to the relationship between the Hellinger distance h and the L^2-norm. Note that for any two mixing measure pairs (G_1,G_2) and (G_1',G_2'), Lemma 4.2 in [72] shows that

$$h^2\Big(\frac{1}{2}f_{G_1,G_2} + \frac{1}{2}f_{G_1^*,G_2^*}, \frac{1}{2}f_{G_1',G_2'} + \frac{1}{2}f_{G_1^*,G_2^*}\Big) \leq \frac{1}{2}h^2(f_{G_1,G_2}, f_{G_1',G_2'}),$$

which yields that $H_B(t/\sqrt{2}, \widetilde{\mathcal{F}}_{k_1,k_2}(\Theta), h) \leq H_B(t, \mathcal{F}_{k_1,k_2}(\Theta), h)$. This result together with equation (41) implies that

$$H_B(t, \widetilde{\mathcal{F}}^{1/2}_{k_1,k_2}(\Theta,t), \|\cdot\|_2) \leq H_B(t, \mathcal{F}_{k_1,k_2}(\Theta), h).$$

From the definition of the Hellinger ball size in equation (40), we have that

$$\begin{aligned}\mathcal{J}_B(\delta, \widetilde{\mathcal{F}}^{1/2}_{k_1,k_2}(\Theta,\delta), \|\cdot\|_2) &= \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(t, \widetilde{\mathcal{F}}^{1/2}_{k_1,k_2}(\Theta,t), \|\cdot\|_2)\mathrm{d}t \vee \delta \\ &\leq \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(t, \mathcal{F}_{k_1,k_2}(\Theta), h)\mathrm{d}t \vee \delta \\ &\lesssim \int_{\delta^2/2^{13}}^{\delta} [\log(1/t)]^{1/2}\mathrm{d}t \vee \delta,\end{aligned}$$

where the last inequality is due to Lemma 2 below. Let $\Psi(\delta) := \delta\sqrt{\log(1/\delta)}$, it can be verified that $\Psi(\delta)/\delta^2$ is a non-increasing function of δ. Furthermore, the above result indicates that $\Psi(\delta) \geq \mathcal{J}_B(\delta, \widetilde{\mathcal{F}}^{1/2}_{k_1,k_2}(\Theta,\delta), \|\cdot\|_2)$. By considering the sequence (δ_n) defined as $\delta_n := \sqrt{\log(n)/n}$, we have $\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n)$ for some universal constant $c > 0$. Then, according to Lemma 1, we get

$$\mathbb{P}\Big(\mathbb{E}_X\Big[h(f_{\widetilde{G}_1^n,\widetilde{G}_2^n}(\cdot|X), f_{G_1^*,G_2^*}(\cdot|X)) > C\sqrt{\log(n)/n}\Big]\Big) \lesssim \exp(-c\log(n)),$$

for some universal constant C depending on Θ. □

Lemma 2. *The following holds for any* $0 < \epsilon < 1/2$*:*

$$H_B(\epsilon, \mathcal{F}_{k_1,k_2}(\Theta), h) \lesssim \log(1/\epsilon).$$

Proof of Lemma 2. Recall that for any mixing measure pair (G_1, G_2), we have

$$f_{G_1,G_2}(Y|X) = \frac{1}{2}\sum_{i=1}^{k_1} \omega_i \pi(Y|h_1(X,\kappa_i),\tau_i) + \frac{1}{2}\sum_{i=1}^{k_2} \frac{\exp((\beta_{1i})^\top X + \beta_{0i})}{\sum_{j=1}^{k_2}\exp((\beta_{1j})^\top X + \beta_{0j})} \cdot \pi(Y|h_2(X,\eta_i),\nu_i).$$

Firstly, we will establish upper bounds for the Gaussian densities $\pi(Y|h_1(X,\kappa),\tau)$ and $\pi(Y|h_2(X,\eta),\nu)$, respectively. Indeed, since the expert function h_1 is bounded and the parameter space is compact, we have $|h_1(X,\kappa)| \leq M_1$ for all $X \in \mathcal{X}$ for some constant $M_1 > 0$, and $\ell_1 \leq \tau \leq u_1$ for some $\ell_1, u_1 > 0$. Therefore, for $|Y| \geq 2M_1$, since $\frac{(Y-h_1(X,\kappa))^2}{2\tau} \geq \frac{Y^2}{8u_1}$ for all $X \in \mathcal{X}$, we have

$$\pi(Y|h_1(X,\kappa),\tau) = \frac{1}{\sqrt{2\pi\tau}}\exp\Big(-\frac{(Y-h_1(X,\kappa))^2}{2\tau}\Big) \leq \frac{1}{\sqrt{2\pi\ell_1}}\exp\Big(-\frac{Y^2}{8u_1}\Big).$$

Next, for $|Y| < 2M_1$, it follows that

$$\pi(Y|h_1(X,\kappa),\tau) = \frac{1}{\sqrt{2\pi\tau}}\exp\Big(-\frac{(Y-h_1(X,\kappa))^2}{2\tau}\Big) \leq \frac{1}{\sqrt{2\pi\tau}} \leq \frac{1}{\sqrt{2\pi\ell_1}}.$$

Combine the above results together, we deduce $\pi(Y|h_1(X,\kappa),\tau) \leq E_1(Y|X)$ for all (X,Y) where

$$E_1(Y|X) := \begin{cases} \frac{1}{\sqrt{2\pi\ell_1}}\exp\Big(-\frac{Y^2}{8u_1}\Big), & \text{for } |Y| \geq 2M_1 \\ \frac{1}{\sqrt{2\pi\ell_1}}, & \text{for } |Y| < 2M_1. \end{cases}$$

By arguing in similar fashion based on the assumptions that $|h_2(X,\eta)| \leq M_2$ for all $X \in \mathcal{X}$ for some constant $M_2 > 0$, and $\ell_2 \leq \nu \leq u_2$ for some $\ell_2, u_2 > 0$, we also get $\pi(Y|h_2(X,\eta),\nu) \leq E_2(Y|X)$, where

$$E_2(Y|X) := \begin{cases} \frac{1}{\sqrt{2\pi\ell_2}}\exp\Big(-\frac{Y^2}{8u_2}\Big), & \text{for } |Y| \geq 2M_2 \\ \frac{1}{\sqrt{2\pi\ell_2}}, & \text{for } |Y| < 2M_2. \end{cases}$$

Now, let $\lambda \leq \epsilon$ be some constant that we will choose later, we denote $p_1, p_2, \ldots, p_N$ as an λ-cover of the set $\mathcal{F}_{k_1,k_2}(\Theta)$, where $N := N(\lambda, \mathcal{F}_{k_1,k_2}(\Theta), \|\cdot\|_\infty)$ stands for the λ-covering number of the set $\mathcal{F}_{k_1,k_2}(\Theta)$ under the L^∞-norm. Then, we take into account the brackets $[p_i^L, p_i^U]$ given by

$$p_i^L(Y|X) := \max\{p_i(Y|X) - \lambda, 0\},$$
$$p_i^U(Y|X) := \max\{p_i(Y|X) + \lambda, E(Y|X)\},$$

for all $i \in [N]$, where $E(Y|X) := \frac{1}{2}E_1(Y|X) + \frac{1}{2}E_2(Y|X)$. It can be justified that $\mathcal{F}_{k_1,k_2}(\Theta) \subseteq \cup_{i=1}^N [p_i^L, p_i^U]$ and $p_i^U(Y|X) - p_i^L(Y|X) \leq \min\{2\lambda, E(Y|X)\}$. Furthermore, we have

$$\|p_i^U - p_i^L\|_2 = \Big(\int [p_i^U(Y|X) - p_i^L(Y|X)]^2 \mathrm{d}(X,Y)\Big)^{1/2} \leq 2\lambda.$$

By definition of the bracketing entropy, we get

$$H_B(2\lambda, \mathcal{F}_{k_1,k_2}(\Theta), \|\cdot\|_2) \leq \log N = \log N(\lambda, \mathcal{F}_{k_1,k_2}(\Theta), \|\cdot\|_\infty).$$

Thus, we need to derive an upper bound for the covering number $N(\lambda, \mathcal{F}_{k_1,k_2}(\Theta), \|\cdot\|_\infty)$. Let us denote $\Delta := \Delta_1 \times \Delta_2$ and $\Omega := \Omega_1 \times \Omega_2$, where

$$\begin{aligned}
&\Delta_1 := \{\omega_i \in \mathbb{R}_+ : (\omega, \kappa, \tau) \in \Theta_1\}, \\
&\Delta_2 := \{(\kappa, \tau) \in \mathbb{R}^{d_1} \times \mathbb{R}_+ : (\omega, \kappa, \tau) \in \Theta_1\}, \\
&\Omega_1 := \{(\beta_0, \beta_1) \in \mathbb{R} \times \mathbb{R}^d : (\beta_0, \beta_1, \eta, \nu) \in \Theta_2\}, \\
&\Omega_2 := \{(\eta, \nu) \in \mathbb{R}^{d_2} \times \mathbb{R}_+ : (\beta_0, \beta_1, \eta, \nu) \in \Theta_2\}.
\end{aligned}$$

Since Θ_1 and Θ_2 are compact, the sets Δ_1, Δ_2 and Ω_1, Ω_2 are also compact. Thus, there exist λ-covers $\Delta_{1,\lambda}, \Delta_{2,\lambda}$ and $\Omega_{1,\lambda}, \Omega_{2,\lambda}$ for those sets, respectively. Moreover, the cardinalities of those λ-covers are bounded as follows:

$$\begin{aligned}
|\Delta_{1,\lambda}| \leq \mathcal{O}(\lambda^{-k_1}), &\quad |\Delta_{2,\lambda}| \leq \mathcal{O}(\lambda^{-(d_1+1)k_1}), \\
|\Omega_{1,\lambda}| \leq \mathcal{O}(\lambda^{-(d+1)k_2}), &\quad |\Omega_{2,\lambda}| \leq \mathcal{O}(\lambda^{-(d_2+1)k_2}).
\end{aligned}$$

For each pair of mixing measure $(G_1, G_2) \in \mathcal{G}_{k_1,k_2}(\Theta)$, we consider two other mixing measure pairs (G'_1, G'_2) and $(\overline{G}_1, \overline{G}_2)$ given by

$$\begin{aligned}
G'_1 &:= \sum_{i=1}^{k_1} \omega_i \delta_{(\bar{\kappa}_i, \bar{\tau}_i)}, & G'_2 &:= \sum_{i=1}^{k_2} \bar{\omega}_i \delta_{(\bar{\kappa}_i, \bar{\tau}_i)}, \\
\overline{G}_1 &:= \sum_{i=1}^{k_2} \exp(\beta_{0i}) \delta_{(\beta_{1i}, \bar{\eta}_i, \bar{\tau}_i)}, & \overline{G}_2 &:= \sum_{i=1}^{k_2} \exp(\bar{\beta}_{0i}) \delta_{(\bar{\beta}_{1i}, \bar{\eta}_i, \bar{\nu}_i)}.
\end{aligned}$$

Above, $\bar{\omega}_i \in \Delta_{1,\lambda}$ is the closest point to ω_i in that set, $(\bar{\kappa}_i, \bar{\tau}_i) \in \Delta_{2,\lambda}$ is the closest point to (κ_i, τ_i) in that set, $(\bar{\beta}_{0i}, \bar{\beta}_{1i}) \in \Omega_{1,\lambda}$ is the closest point to (β_{0i}, β_{1i}) in that set, $(\bar{\eta}_i, \bar{\nu}_i) \in \Omega_{2,\lambda}$ is the closest point to (η_i, ν_i) in that set. Subsequently, we aim to upper bound the term $\|f_{G_1,G_2} - f_{\overline{G}_1,\overline{G}_2}\|_\infty$. By the triangle inequality, we have

$$\|f_{G_1,G_2} - f_{\overline{G}_1,\overline{G}_2}\|_\infty \leq \|f_{G_1,G_2} - f_{G'_1,G'_2}\|_\infty + \|f_{G'_1,G'_2} - f_{\overline{G}_1,\overline{G}_2}\|_\infty.$$

We aim to upper bound the two terms in the above right hand sides, respectively. For ease of presentation, for any mixing measure pair (G_1, G_2), we denote

$$\begin{aligned}
q_{G_1}(Y|X) &:= \sum_{i=1}^{k_1} \omega_i \pi(Y|h_1(X, \kappa_i), \tau_i), \\
p_{G_2}(Y|X) &:= \sum_{i=1}^{k_2} \frac{\exp(\beta_{1i}^\top X + \beta_{0i})}{\sum_{j=1}^{k_2} \exp(\beta_{1j}^\top X + \beta_{0j})} \pi(Y|h_2(X, \eta_i), \nu_i).
\end{aligned}$$

We start with bounding the term $\|f_{G_1,G_2} - f_{G'_1,G'_2}\|_\infty$ as follows:

$$\|f_{G_1,G_2} - f_{G'_1,G'_2}\|_\infty \leq \frac{1}{2}\|q_{G_1} - q_{G'_1}\|_\infty + \frac{1}{2}\|p_{G_2} - p_{G'_2}\|_\infty.$$

In particular, we have

$$\begin{aligned}
\|q_{G_1} - q_{G_1'}\|_\infty &= \sup_{(X,Y)\in\mathcal{X}\times\mathcal{Y}} \left| \sum_{i=1}^{k_1} \omega_i \Big[\pi(Y|h_1(X,\kappa_i),\tau_i) - \pi(Y|h_1(X,\bar{\kappa}_i),\bar{\tau}_i)\Big] \right| \\
&\leq \sum_{i=1}^{k_1} \omega_i \sup_{(X,Y)\in\mathcal{X}\times\mathcal{Y}} \Big|\pi(Y|h_1(X,\kappa_i),\tau_i) - \pi(Y|h_1(X,\bar{\kappa}_i),\bar{\tau}_i)\Big| \\
&\lesssim \sum_{i=1}^{k_1} \omega_i (\|\kappa_i - \bar{\kappa}_i\| + |\tau_i - \bar{\tau}_i|) \\
&\leq \sum_{i=1}^{k_1} \omega_i (\lambda + \lambda) = 2\lambda \lesssim \lambda,
\end{aligned}$$

and

$$\begin{aligned}
&\|p_{G_2} - p_{G_2'}\|_\infty \\
&= \sup_{(X,Y)\in\mathcal{X}\times\mathcal{Y}} \left| \sum_{i=1}^{k_2} \frac{\exp(\beta_{1i}^\top X + \beta_{0i})}{\sum_{j=1}^{k_2}\exp(\beta_{1j}^\top X + \beta_{0j})} \Big[\pi(Y|h_2(X,\eta_i),\nu_i) - \pi(Y|h_1(X,\bar{\eta}_i),\bar{\nu}_i)\Big] \right| \\
&\leq \sum_{i=1}^{k_2} \sup_{(X,Y)\in\mathcal{X}\times\mathcal{Y}} \frac{\exp(\beta_{1i}^\top X + \beta_{0i})}{\sum_{j=1}^{k_2}\exp(\beta_{1j}^\top X + \beta_{0j})} \Big|\pi(Y|h_2(X,\eta_i),\nu_i) - \pi(Y|h_1(X,\bar{\eta}_i),\bar{\nu}_i)\Big| \\
&\leq \sum_{i=1}^{k_2} \sup_{(X,Y)\in\mathcal{X}\times\mathcal{Y}} \Big|\pi(Y|h_2(X,\eta_i),\nu_i) - \pi(Y|h_1(X,\bar{\eta}_i),\bar{\nu}_i)\Big| \\
&\lesssim \sum_{i=1}^{k_2} (\|\eta_i - \bar{\eta}_i\| + |\nu_i - \bar{\nu}_i|) \leq \sum_{i=1}^{k_2} (\lambda + \lambda) \lesssim \lambda,
\end{aligned}$$

which implies that

$$\|f_{G_1,G_2} - f_{G_1',G_2'}\|_\infty \lesssim \frac{1}{2}\lambda + \frac{1}{2}\lambda = \lambda. \tag{42}$$

Next, we continue with bounding the term $\|f_{G_1',G_2'} - f_{\overline{G}_1,\overline{G}_2}\|_\infty$ as

$$\|f_{G_1',G_2'} - f_{\overline{G}_1,\overline{G}_2}\|_\infty \leq \frac{1}{2}\|q_{G_1'} - q_{\overline{G}_1}\|_\infty + \frac{1}{2}\|p_{G_2'} - p_{\overline{G}_2}\|_\infty.$$

By looking into each term in the above right hand side, we have

$$\begin{aligned}
\|q_{G_1'} - q_{\overline{G}_1}\|_\infty &= \sup_{(X,Y)\in\mathcal{X}\times\mathcal{Y}} \left| \sum_{i=1}^{k_1} [\omega_i - \bar{\omega}_i] \pi(Y|h_1(X,\bar{\kappa}_i),\bar{\tau}_i) \right| \\
&\leq \sum_{i=1}^{k_1} |\omega_i - \bar{\omega}_i| \sup_{(X,Y)\in\mathcal{X}\times\mathcal{Y}} |\pi(Y|h_1(X,\bar{\kappa}_i),\bar{\tau}_i)| \\
&\lesssim \sum_{i=1}^{k_1} |\omega_i - \bar{\omega}_i| \leq \sum_{i=1}^{k_1} \lambda \lesssim \lambda,
\end{aligned}$$

and

$$\begin{aligned}
&\|p_{G'_2} - p_{\overline{G}_2}\|_\infty \\
&= \sup_{(X,Y)\in\mathcal{X}\times\mathcal{Y}} \left| \sum_{i=1}^{k_2} \left[\frac{\exp(\beta_{1i}^\top X + \beta_{0i})}{\sum_{j=1}^{k_2}\exp(\beta_{1j}^\top X + \beta_{0j})} - \frac{\exp(\bar{\beta}_{1i}^\top X + \bar{\beta}_{0i})}{\sum_{j=1}^{k_2}\exp(\bar{\beta}_{1j}^\top X + \bar{\beta}_{0j})} \right] \pi(Y|h_1(X,\bar{\kappa}_i),\bar{\tau}_i) \right| \\
&\leq \sum_{i=1}^{k_2} \sup_{(X,Y)\in\mathcal{X}\times\mathcal{Y}} \left| \frac{\exp(\beta_{1i}^\top X + \beta_{0i})}{\sum_{j=1}^{k_2}\exp(\beta_{1j}^\top X + \beta_{0j})} - \frac{\exp(\bar{\beta}_{1i}^\top X + \bar{\beta}_{0i})}{\sum_{j=1}^{k_2}\exp(\bar{\beta}_{1j}^\top X + \bar{\beta}_{0j})} \right| \cdot |\pi(Y|h_1(X,\bar{\kappa}_i),\bar{\tau}_i)| \\
&\lesssim \sum_{i=1}^{k_2} \sup_{(X,Y)\in\mathcal{X}\times\mathcal{Y}} \left| \frac{\exp(\beta_{1i}^\top X + \beta_{0i})}{\sum_{j=1}^{k_2}\exp(\beta_{1j}^\top X + \beta_{0j})} - \frac{\exp(\bar{\beta}_{1i}^\top X + \bar{\beta}_{0i})}{\sum_{j=1}^{k_2}\exp(\bar{\beta}_{1j}^\top X + \bar{\beta}_{0j})} \right| \\
&\lesssim \sum_{i=1}^{k_2} \sup_{X\in\mathcal{X}} \Big(\|\beta_{1i} - \bar{\beta}_{1i}\| \cdot \|X\| + \|\beta_{0i} - \bar{\beta}_{0i}\| \Big) \\
&\lesssim \sum_{i=1}^{k_2} \Big(\lambda \cdot \sup_{X\in\mathcal{X}} \|X\| + \lambda \Big) \lesssim \lambda.
\end{aligned}$$

Putting these bounds together, we deduce

$$\|f_{G'_1,G'_2} - f_{\overline{G}_1,\overline{G}_2}\|_\infty \lesssim \frac{1}{2}\lambda + \frac{1}{2}\lambda = \lambda. \tag{43}$$

From equations (42) and (43), we obtain

$$\|f_{G_1,G_2} - f_{\overline{G}_1,\overline{G}_2}\|_\infty \leq \lambda + \lambda \lesssim \lambda.$$

By definition of the covering number, we get

$$\begin{aligned}
N(\lambda, \mathcal{F}_{k_1,k_2}(\Theta), \|\cdot\|_\infty) &\leq |\Delta_{1,\lambda}| \cdot |\Delta_{2,\lambda}| \cdot |\Omega_{1,\lambda}| \cdot |\Omega_{2,\lambda}| \\
&\leq \mathcal{O}(\lambda^{-k_1}) \cdot \mathcal{O}(\lambda^{-(d_1+1)k_1}) \cdot \mathcal{O}(\lambda^{-(d+1)k_2}) \cdot \mathcal{O}(\lambda^{-(d_2+1)k_2}) \\
&\leq \mathcal{O}(\lambda^{-(d_1+2)k_1 - (d_2+d+2)k_2}).
\end{aligned}$$

As a result, we deduce

$$H_B(2\lambda, \mathcal{F}_{k_1,k_2}(\Theta), \|\cdot\|_2) \leq \log N(\lambda, \mathcal{F}_{k_1,k_2}(\Theta), \|\cdot\|_\infty) \lesssim \log(1/\lambda).$$

Let $\lambda = \epsilon/2$, we achieve the desired result that $H_B(\epsilon, \mathcal{F}_{k_1,k_2}(\Theta), \|\cdot\|_2) \lesssim \log(1/\epsilon)$. Hence, the proof is completed. □

E.2 Identifiability of DeepSeekMoE

Proposition 5 (Identifiability). *For any pair of mixing measures (G_1, G_2), if the equation $f_{G_1,G_2}(Y|X) = f_{G_1^*,G_2^*}(Y|X)$ holds for almost surely (X,Y), then we obtain $(G_1,G_2) \equiv (G_1^*, G_2^*)$.*

Proof of Proposition 5. First of all, we expand the equation $f_{G_1,G_2}(Y|X) = f_{G_1^*,G_2^*}(Y|X)$ for almost surely (X,Y) as follows:

$$\frac{1}{2}\sum_{i=1}^{k_1} \omega_i \pi(Y|h_1(X,\kappa_i),\tau_i) + \frac{1}{2}\sum_{i=1}^{k_2} \frac{\exp(\beta_{1i}^\top X + \beta_{0i})}{\sum_{j=1}^{k_2}\exp(\beta_{1j}^\top X + \beta_{0j})} \pi(Y|h_2(X,\eta_i),\nu_i)$$

$$= \frac{1}{2}\sum_{i=1}^{k_1^*} \omega_i^* \pi(Y|h_1(X,\kappa_i^*),\tau_i^*) + \frac{1}{2}\sum_{i=1}^{k_2^*} \frac{\exp((\beta_{1i}^*)^\top X + \beta_{0i}^*)}{\sum_{j=1}^{k_2^*}\exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)} \cdot \pi(Y|h_2(X,\eta_i^*),\nu_i^*). \quad (44)$$

Since the location-scale Gaussian mixtures are identifiable [70], the above equation implies that $k_1 + k_2 = k_1^* + k_2^*$ and

$$\left\{\omega_{i'}, \frac{\exp(\beta_{1i}^\top X + \beta_{0i})}{\sum_{j=1}^{k_2}\exp(\beta_{1j}^\top X + \beta_{0j})} : i' \in [k_1],\ i \in [k_2]\right\}$$
$$= \left\{\omega_{i'}^*, \frac{\exp((\beta_{1i}^*)^\top X + \beta_{0i}^*)}{\sum_{j=1}^{k_2^*}\exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)} : i' \in [k_1^*],\ i \in [k_2^*]\right\},$$

for almost surely X. As the weights $\omega_{i'}$ and $\omega_{i'}^*$ are independent of X for all $i' \in [k_1^*]$, we deduce $k_1 = k_1^*$ and $\{\omega_{i'} : i' \in [k_1^*]\} = \{\omega_{i'}^* : i' \in [k_1^*]\}$. For simplicity, we assume WLOG that $\omega_{i'} = \omega_{i'}^*$ for all $i' \in [k_1^*]$. Furthermore, we also get $k_2 = k_2^*$ and

$$\left\{\frac{\exp(\beta_{1i}^\top X + \beta_{0i})}{\sum_{j=1}^{k_2^*}\exp(\beta_{1j}^\top X + \beta_{0j})} : i \in [k_2^*]\right\} = \left\{\frac{\exp((\beta_{1i}^*)^\top X + \beta_{0i}^*)}{\sum_{j=1}^{k_2^*}\exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)} : i \in [k_2^*]\right\},$$

for almost surely X. Again, we assume WLOG that $\frac{\exp(\beta_{1i}^\top X+\beta_{0i})}{\sum_{j=1}^{k_2^*}\exp(\beta_{1j}^\top X+\beta_{0j})} = \frac{\exp((\beta_{1i}^*)^\top X+\beta_{0i}^*)}{\sum_{j=1}^{k_2^*}\exp((\beta_{1j}^*)^\top X+\beta_{0j}^*)}$ for almost surely X for all $i \in [k_2^*]$. Due to the invariance to translation of the softmax function, this result indicates that $\beta_{1i} = \beta_{1i}^* + c_1$ and $\beta_{0i} = \beta_{0i}^* + c_0$ for some $c_1 \in \mathbb{R}^d$ and $c_0 \in \mathbb{R}$. Then, it follows from the assumption $\beta_{1k_2^*} = \beta_{1k_2^*}^* = 0_d$ and $\beta_{0k_2^*} = \beta_{0k_2^*}^* = 0$ that $c_1 = 0_d$ and $c_0 = 0$. Therefore, we obtain $\beta_{1i} = \beta_{1i}^*$ and $\beta_{0i} = \beta_{0i}^*$ for all $i \in [k_2^*]$.

Subsequently, we partition the index set $[k_1^*]$ into disjoint subsets $U_1, U_2, \ldots, U_{m_1}$ such that for each $\ell \in [m_1]$, we have (i) $\omega_i = \omega_{i'}^*$ for $i, i' \in U_\ell$ and (ii) $\omega_i \neq \omega_{i'}^*$ if i and i' dot not belong to the same set U_ℓ. Similarly, we also partition the index set $[k_2^*]$ into disjoint subsets $V_1, V_2, \ldots, V_{m_2}$ such that for each $\ell \in [m_2]$, we have (i) $\exp(\beta_{0i}) = \exp(\beta_{0i'}^*)$ for $i, i' \in V_\ell$ and (ii) $\exp(\beta_{0i}) \neq \exp(\beta_{0i'}^*)$ if i and i' dot not belong to the same set V_ℓ. As a consequence, we can rewrite equation (44) as

$$\frac{1}{2}\sum_{\ell=1}^{m_1}\sum_{i\in U_\ell} \omega_i \pi(Y|h_1(X,\kappa_i),\tau_i) + \frac{1}{2S}\sum_{\ell=1}^{m_2}\sum_{i\in V_\ell} \exp(\beta_{0i})\exp(\beta_{1i}^\top X)\pi(Y|h_2(X,\eta_i),\nu_i)$$
$$= \frac{1}{2}\sum_{\ell=1}^{m_1}\sum_{i\in U_\ell} \omega_i^* \pi(Y|h_1(X,\kappa_i^*),\tau_i^*) + \frac{1}{2S}\sum_{\ell=1}^{m_2}\sum_{i\in V_\ell} \exp(\beta_{0i}^*)\exp((\beta_{1i}^*)^\top X)\pi(Y|h_2(X,\eta_i^*),\nu_i^*),$$

for almost surely (X, Y), where we denote $S := \sum_{j=1}^{k_2^*}\exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)$. The above equation implies that

$$\{(h_1(X,\kappa_i),\tau_i) : i \in U_\ell\} = \{(h_1(X,\kappa_i^*),\tau_i^*) : i \in U_\ell\}, \quad \forall \ell \in [m_1]$$
$$\{(h_2(X,\eta_i),\nu_i) : i \in V_\ell\} = \{(h_2(X,\eta_i^*),\nu_i^*) : i \in V_\ell\}, \quad \forall \ell \in [m_2],$$

for almost surely X. As the expert functions h_1 and h_2 are identifiable, we deduce

$$\{(\kappa_i,\tau_i) : i \in U_\ell\} = \{(\kappa_i^*,\tau_i^*) : i \in U_\ell\}, \quad \forall \ell \in [m_1]$$

$$\{(\eta_i, \nu_i) : i \in V_\ell\} = \{(\eta_i^*, \nu_i^*) : i \in V_\ell\}, \quad \forall \ell \in [m_2].$$

Therefore, we obtain

$$G_1 = \sum_{\ell=1}^{m_1} \sum_{i \in U_\ell} \omega_i \delta_{(\kappa_i, \tau_i)} = \sum_{\ell=1}^{m_1} \sum_{i \in U_\ell} \omega_i^* \delta_{(\kappa_i^*, \tau_i^*)} = G_1^*,$$
$$G_2 = \sum_{\ell=1}^{m_2} \sum_{i \in V_\ell} \exp(\beta_{0i}) \delta_{(\beta_{1i}, \eta_i, \nu_i)} = \sum_{\ell=1}^{m_2} \sum_{i \in V_\ell} \exp(\beta_{0i}^*) \delta_{(\beta_{1i}^*, \eta_i^*, \nu_i^*)} = G_2^*.$$

Hence, the proof is completed. □

F Extended Theoretical Results for Sparse Gating MoE

In this appendix, we extend the convergence analysis of parameter and expert estimations presented in Theoreom 1 to the setting of a Top-K sparse gating function. Our main arguments rely on fundamental techniques for dealing with the sparse gating function proposed in [54]. Since the results of Theorems 2, 3, and 4 can be extended in a similar fashion, we will omit their extension here.

Problem setting: Assume that $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ are i.i.d. samples drawn from the softmax gating Gaussian mixture of experts of order k_* whose conditional density function $s_{G_1^*, G_2^*}(y|x)$ is given by:

$$\begin{aligned} s_{G_1^*, G_2^*}(y|x) := \frac{1}{2} \sum_{i=1}^{k_1^*} \omega_i^* \pi(y|h_1(x, \kappa_i^*), \tau_i^*) \\ + \frac{1}{2} \sum_{i=1}^{k_2^*} \text{softmax}(\text{Top}_K((\beta_{1i}^*)^\top x; \beta_{0i}^*)) \pi(y|h_2(x, \eta_i^*), \nu_i^*), \end{aligned} \tag{45}$$

where the pair of ground-truth mixing measures (G_1^*, G_2^*) are given by $G_1^* := \sum_{i=1}^{k_1^*} \omega_i^* \delta_{(\kappa_i^*, \tau_i^*)}$ and $G_2^* := \sum_{i=1}^{k_2^*} \exp(\beta_{0i}^*) \delta_{(\beta_{1i}^*, \eta_i^*, \nu_i^*)}$. Additionally, for any natural number k and vectors $(v_i)_{i=1}^k$ and $(u_i)_{i=1}$ in $\mathbb{R}^k$, the Top$_K$ sparse function is defined as

$$\text{Top}_K(v_i, K; u_i) := \begin{cases} v_i + u_i, & \text{if } v_i \text{ is in the top } K \text{ elements of } v; \\ -\infty, & \text{otherwise,} \end{cases}$$

while the softmax function is formulated as $\text{softmax}(v_i) := \exp(v_i) / \sum_{j=1}^k \exp(v_j)$.

In practice, since the number of shared experts k_1^* and routed experts k^*2 are typically unknown, we have to fit the ground-truth model (45) with $k_1 > k_1^*$ shared experts and $k_2 > k_2^*$ routed experts. Thus, some ground-truth shared experts and routed experts will be fitted by more than one estimated expert. As a result, since there are K routed experts activated per input in the ground-truth density $s_{G_1^*, G_2^*}$, it is necessary to activate $\bar{K} > K$ experts in the density estimation in order to ensure its convergence to the true density. For that purpose, let us introduce the formulation of the density estimation as follows:

$$\bar{s}_{G_1^n,G_2^n}(Y|X) := \frac{1}{2}\sum_{i=1}^{k_1^n}\omega_i^n\pi(y|h_1(x,\kappa_i^n),\tau_i^n) \\ + \frac{1}{2}\sum_{i=1}^{k_2^n}\mathrm{softmax}(\mathrm{Top}_{\bar{K}}((\beta_{1i}^n)^\top x;\beta_{0i}^n))\pi(y|h_2(x,\eta_i^n),\nu_i^n),$$

where $K < \bar{K} \leq k_2$ and the pair of mixing measure estimations (G_1^n, G_2^n) are defined as

$$(\widehat{G}_1^n, \widehat{G}_2^n) \in \underset{(G_1,G_2)\in\mathcal{G}_{k_1,k_2}(\Theta)}{\arg\max} \frac{1}{n}\sum_{i=1}^{n}\log(\bar{s}_{G_1,G_2}(Y_i|X_i)), \tag{46}$$

where the set of mixing measures $\mathcal{G}_{k_1,k_2}(\Theta) := \mathcal{G}_{k_1}(\Theta_1) \times \mathcal{G}_{k_2}(\Theta)$ is defined below equation (2).

Input space partition w.r.t the true density. In order that the density estimation $s_{G_1^n,G_2^n}$ converges to the true density $s_{G_1^*,G_2^*}$, we must ensure that for each input, the $\bar{K}$ routed experts activated in the density estimation converge to the K routed experts activated in the true density. Since the activated experts vary with the input value, we need to partition the input space $\mathcal{X}$ into $M := \binom{k_2^*}{K}$ regions $\mathcal{X}_m^*$ corresponding to $\binom{k_2^*}{K}$ choices of activated experts in the true density. For each $m \in [M]$, let us denote $\{m_1, m_2, \ldots, m_K\}$ as an K-element subset of the index set $[k_2^*]$, and $\{m_{K+1}, \ldots, m_{k_2^*}\} := [k_2^*] \setminus \{m_1, m_2, \ldots, m_K\}$. Then, the m-th region of the input space is defined as

$$\mathcal{X}_m^* := \Big\{x \in \mathcal{X} : (\beta_{1i}^*)^\top x \geq (\beta_{1i'}^*)^\top x, \ \forall i \in \{m_1, m_2, \ldots, m_K\}, i' \in \{m_{K+1}, \ldots, m_{k_2^*}\}\Big\},$$

for any $m \in [M]$. For example, suppose that $X \in \mathcal{X}_m^*$ where $m \in [M]$ such that $\{m_1, m_2, \ldots, m_K\} = \{1, 2, \ldots, K\}$. Then, it follows that

$$\mathrm{Top}_K((\beta_{1i}^*)^\top X; \beta_{0i}^*) = (\beta_{1i}^*)^\top X + \beta_{0i}^*,$$

for all $i \in [K]$. In other words, $h_2(X,\eta_1^*), h_2(X,\eta_2^*), \ldots, h_2(X,\eta_K^*)$ are the K routed experts activated in the true density $s_{G_1^*,G_2^*}(y|x)$, which is reduced to

$$s_{G_1^*,G_2^*}(y|x) := \frac{1}{2}\sum_{i=1}^{k_1^*}\omega_i^*\pi(y|h_1(x,\kappa_i^*),\tau_i^*) \\ + \frac{1}{2}\sum_{i=1}^{K}\frac{\exp((\beta_{1i}^*)^\top x + \beta_{0i}^*)}{\sum_{j=1}^{k_2^*}\exp((\beta_{1j}^*)^\top x + \beta_{0j}^*)} \cdot \pi(y|h_2(x,\eta_i^*),\nu_i^*). \tag{47}$$

Input space partition w.r.t the density estimation. Next, with the same input $X \in \mathcal{X}_m^*$, we need to guarantee that the routed expert estimations converging to the above K routed experts activated in the true density $s_{G_1^*,G_2^*}(Y|X)$ are also activated in the density estimation $s_{G_1^n,G_2^n}(Y|X)$. For that purpose, it is necessary to partition the input space with respect to the density estimation. In particular, we partition the input space into $\bar{M} := \binom{k_2}{\bar{K}}$ regions $\bar{\mathcal{X}}_m$ corresponding to $\binom{k_2}{\bar{K}}$ choices of activated experts in the true density. For each $\bar{m} \in [\bar{M}]$, we denote $\{\bar{m}_1, \bar{m}_2, \ldots, \bar{m}_{\bar{K}}\}$ as an

$\bar{K}$-element subset of the index set $[k_2]$, and $\{\bar{m}_{\bar{K}+1}, \ldots, \bar{m}_{k_2}\} := [k_2] \setminus \{\bar{m}_1, \bar{m}_2, \ldots, \bar{m}_{\bar{K}}\}$. Given these notations, we are ready to show that the input partition w.r.t the density estimation aligns with the input space partition w.r.t the true density in the following lemma whose proof will be provided in Appendix F.1:

Lemma 3. *For any $j \in [k_2^*]$, $i \in \mathcal{V}_{2,j}$ and $\beta_{1i}, \beta_{1j}^* \in \mathbb{R}^d$, assume that there exist sufficiently small $\varepsilon_j > 0$ satisfying $\|\beta_{1i} - \beta_{1j}^*\| \leq \varepsilon_j$. Moreover, suppose that there exist $m \in [M]$ and $\bar{m} \in [\bar{M}]$ such that $\{\bar{m}_1, \bar{m}_2, \ldots, \bar{m}_{\bar{K}}\} = \mathcal{V}_{2,m_1} \cup \mathcal{V}_{2,m_2} \ldots \cup \mathcal{V}_{2,m_K}$. Then, for any $m \in [M]$, if the input region $\mathcal{X}_m^*$ has non-zero measure, we have $\mathcal{X}_m^* = \bar{\mathcal{X}}_{\bar{m}}$, where*

$$\bar{\mathcal{X}}_{\bar{m}} := \Big\{x \in \mathcal{X} : (\beta_{1i})^\top x \geq (\beta_{1i'})^\top x, \ \forall i \in \{\bar{m}_1, \bar{m}_2, \ldots, \bar{m}_{\bar{K}}\}, i' \in \{\bar{m}_{K+1}, \ldots, \bar{m}_{k_2}\}\Big\}.$$

Suppose that the expert estimation $h(X, \hat{\eta}_i^n)$ converges to the ground-truth expert $h(X, \eta_j^*)$ for some $j \in [k_2^*]$ and $i \in \mathcal{V}_{2,j}$. Then, Lemma 3 reveals that for almost surely X, if the expert $h(X, \eta_j^*)$ is activated in the true density, then the expert $h(X, \hat{\eta}_i^n)$ is also activated in the density estimation. Mathematically, we have $\mathrm{Top}_K((\beta_{1j}^*)^\top X; \beta_{0j}^*) = (\beta_{1j}^*)^\top X + \beta_{0j}^*$ occurs holds if and only if $\mathrm{Top}_{\bar{K}}((\hat{\beta}_{1i}^n)^\top X; \hat{\beta}_{0i}^n) = (\hat{\beta}_{1i}^n)^\top X + \hat{\beta}_{0i}^n$.

Density estimation convergence. Given the above input partition w.r.t the density estimation, we exhibit in Proposition 6 an interesting phenomenon that the density estimation $\bar{s}_{\widehat{G}_1^n, \widehat{G}_2^n}$ converges to the true density $s_{G_1^*, G_2^*}$ under the Total Variation distance only if the number of routed experts activated in the density estimation is bounded below as $\bar{K} \geq \max_{\{m_1, m_2, \ldots, m_K\} \subset [k_2^*]} \sum_{j=1}^K |\mathcal{V}_{2,m_j}|$.

Proposition 6. *If $\bar{K} < \max_{\{m_1, m_2, \ldots, m_K\} \subset [k_2^*]} \sum_{j=1}^K |\mathcal{V}_{2,m_j}|$, then the following holds:*

$$\inf_{(G_1, G_2) \in \mathcal{G}_{k_1, k_2}(\Theta)} \mathbb{E}_X[V(\bar{s}_{G_1, G_2}(\cdot|X), s_{G_1^*, G_2^*}(\cdot|X))] > 0.$$

Proof of Proposition 6 will be provided in Appendix F.2. Following from the result of Proposition 6, we will assume $\max_{\{m_1, m_2, \ldots, m_K\} \subset [k_2^*]} \sum_{j=1}^K |\mathcal{V}_{2,m_j}| \leq \bar{K} \leq k_2$ in the rest of this appendix unless stating otherwise to ensure the convergence of density estimation. Next, by combining the above results and the arguments used to prove Proposition 1, we arrive at the following density estimation rate.

Proposition 7. *The density estimation $\bar{s}_{\widehat{G}_1^n, \widehat{G}_2^n}(Y|X)$ converges to the true density $s_{G_1^*, G_2^*}(Y|X)$ at the following rate:*

$$\mathbb{E}_X[V(\bar{s}_{\widehat{G}_1^n, \widehat{G}_2^n}(\cdot|X), s_{G_1^*, G_2^*}(\cdot|X))] = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}}).$$

Voronoi loss. In align with the above input partition w.r.t the true density, we need to modify the formulation of the Voronoi loss previously defined in equation (4) as follows:

$$\mathcal{D}_5((G_1, G_2), (G_1^*, G_2^*)) := \max_{\{m_1, \ldots, m_K\} \subset [k_2^*]} \Bigg\{ \sum_{j=1}^{k_1^*} \Big| \sum_{i \in \mathcal{V}_{1,j}} \omega_i - \omega_j^* \Big| + \sum_{j=1}^{K} \Big| \sum_{i \in \mathcal{V}_{2,m_j}} \exp(\beta_{0i}) - \exp(\beta_{0j}^*) \Big|$$
$$+ \sum_{\substack{j \in [k_1^*], \\ |\mathcal{V}_{1,j}|=1}} \sum_{i \in \mathcal{V}_{1,j}} \omega_i (\|\Delta \kappa_{ij}\| + |\Delta \tau_{ij}|) + \sum_{\substack{j \in [K], \\ |\mathcal{V}_{2,m_j}|=1}} \sum_{i \in \mathcal{V}_{2,m_j}} \exp(\beta_{0i}) (\|\Delta \beta_{1im_j}\| + \|\Delta \eta_{im_j}\| + |\Delta \nu_{im_j}|)$$

$$+ \sum_{\substack{j\in[k_1^*],\\ |\mathcal{V}_{1,j}|>1}} \sum_{i\in\mathcal{V}_{1,j}} \omega_i(\|\Delta\kappa_{ij}\|^2 + |\Delta\tau_{ij}|^2) + \sum_{\substack{j\in[K],\\ |\mathcal{V}_{2,m_j}|>1}} \sum_{i\in\mathcal{V}_{2,m_j}} \exp(\beta_{0i})(\|\Delta\beta_{1im_j}\|^2 + \|\Delta\eta_{im_j}\|^2 + |\Delta\nu_{im_j}|^2)\Bigg\}.$$

The maximum operator in the above formulation helps capture the convergence behavior of the parameter estimation in different input regions partitioned w.r.t the true density. Given the loss function $\mathcal{D}_5(G, G_*)$, it is sufficient to establish parameter and expert estimation rates in the following theorem:

Theorem 5. *Suppose that the expert functions h_1 and h_2 are strongly identifiable. Then, the lower bound $\mathbb{E}_X[V(\bar{s}_{G_1,G_2}(\cdot|X), s_{G_1^*,G_2^*}(\cdot|X))] \gtrsim \mathcal{D}_5((G_1,G_2),(G_1^*,G_2^*))$ holds for any $(G_1,G_2)\in\mathcal{G}_{k_1,k_2}(\Theta)$. As a consequence, we have*

$$\mathcal{D}_5((\widehat{G}_1^n, \widehat{G}_2^n), (G_1^*, G_2^*)) = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}}).$$

Proof of Theorem 5. Analogous to Appendix D.1, it suffices to derive the local part

$$\lim_{\varepsilon\to 0} \inf_{(G_1,G_2)\in\mathcal{G}_{k_1,k_2}(\Theta):\mathcal{D}_5((G_1,G_2),(G_1^*,G_2^*))\leq\varepsilon} \frac{\mathbb{E}_X[V(\bar{s}_{G_1,G_2}(\cdot|X), s_{G_1^*,G_2^*}(\cdot|X))]}{\mathcal{D}_5((G_1,G_2),(G_1^*,G_2^*))} > 0, \tag{48}$$

and the global part

$$\inf_{(G_1,G_2)\in\mathcal{G}_{k_1,k_2}(\Theta):\mathcal{D}_5((G_1,G_2),(G_1^*,G_2^*))>\varepsilon'} \frac{\mathbb{E}_X[V(\bar{s}_{G_1,G_2}(\cdot|X), s_{G_1^*,G_2^*}(\cdot|X))]}{\mathcal{D}_5((G_1,G_2),(G_1^*,G_2^*))} > 0. \tag{49}$$

in this appendix. However, since the global part (21) can be established in the same fashion as in Appendix D.1, its proof is omitted here. Thus, we will focus on showing only the local part (48). Suppose that the local part does not hold. Then, we can find a sequence of mixing measure pairs (G_1^n, G_2^n) of the form $G_1^n := \sum_{i=1}^{k_1^n} \omega_i^n \delta_{(\kappa_{1i}^n, \kappa_{0i}^n, \tau_i^n)}$, $G_2^n := \sum_{i=1}^{k_2^n} \exp(\beta_{0i}^n)\delta_{(\beta_{1i}^n,\eta_{1i}^n,\eta_{0i}^n,\nu_i^n)}$ for $n\in\mathbb{N}$ satisfying $\mathcal{D}_{5n} := \mathcal{D}_5((G_1^n,G_2^n),(G_1^*,G_2^*)) \to 0$ and

$$\mathbb{E}_X[V(\bar{s}_{G_1^n,G_2^n}(\cdot|X), s_{G_1^*,G_2^*}(\cdot|X))]/\mathcal{D}_{5n} \to 0, \tag{50}$$

as $n\to\infty$. Here, we may assume WLOG that the number of shared experts and routed experts k_1^n, k_2^n and Voronoi cells $\mathcal{V}_{1,j} = \mathcal{V}_{1,j}(G_1^n)$, $\mathcal{V}_{2,j} = \mathcal{V}_{2,j}(G_2^n)$ do not change with the sample size n. WLOG, we may assume that the Voronoi loss $\mathcal{D}_{5n}$ is reduced to

$$\begin{aligned}
\mathcal{D}_{5n} &= \sum_{j=1}^{k_1^*}\Big|\sum_{i\in\mathcal{V}_{1,j}}\omega_i^n - \omega_j^*\Big| + \sum_{j=1}^{K}\Big|\sum_{i\in\mathcal{V}_{2,j}}\exp(\beta_{0i}^n) - \exp(\beta_{0j}^*)\Big| \\
&+ \sum_{\substack{j\in[k_1^*],\\ |\mathcal{V}_{1,j}|=1}} \sum_{i\in\mathcal{V}_{1,j}} \omega_i^n(\|\Delta\kappa_{ij}^n\| + |\Delta\tau_{ij}^n|) + \sum_{\substack{j\in[K],\\ |\mathcal{V}_{2,j}|=1}} \sum_{i\in\mathcal{V}_{2,j}} \exp(\beta_{0i}^n)(\|\Delta\beta_{1ij}^n\| + \|\Delta\eta_{ij}^n\| + |\Delta\nu_{ij}^n|) \\
&+ \sum_{\substack{j\in[k_1^*],\\ |\mathcal{V}_{1,j}|>1}} \sum_{i\in\mathcal{V}_{1,j}} \omega_i^n(\|\Delta\kappa_{ij}^n\|^2 + |\Delta\tau_{ij}^n|^2) + \sum_{\substack{j\in[K],\\ |\mathcal{V}_{2,j}|>1}} \sum_{i\in\mathcal{V}_{2,j}} \exp(\beta_{0i}^n)(\|\Delta\beta_{1ij}^n\|^2 + \|\Delta\eta_{ij}^n\|^2 + |\Delta\nu_{ij}^n|^2).
\end{aligned} \tag{51}$$

Recall that we partition the input space w.r.t the true density into $M = \binom{k_2^*}{K}$ regions. For each $m \in [M]$, we denote $\{m_1, m_2, \dots, m_K\}$ as a subset of the index set $[k_2^*]$ and $\{m_{K+1}, \dots, m_{k_2^*}\} = [k_2^*] \setminus \{m_1, m_2, \dots, m_K\}$. Then, the m-th region is given by

$$\mathcal{X}_m^* := \Big\{x \in \mathcal{X} : (\beta_{1i}^*)^\top x \geq (\beta_{1i'}^*)^\top x, \ \forall i \in \{m_1, m_2, \dots, m_K\}, i' \in \{m_{K+1}, \dots, m_{k_2^*}\}\Big\},$$

for any $m \in [M]$. Let $\bar{K} \in \mathbb{N}$ such that $\max\{\sum_{j=1}^K |\mathcal{V}_{2,j}| : \{m_1, \dots, m_K\} \subset [k_2^*]\} \leq \bar{K} \leq k_2$ and let $\bar{M} := \binom{k_2}{\bar{K}}$. Next, for any $\bar{m} \in [\bar{M}]$, we denote $\{\bar{m}_1, \bar{m}_2, \dots, \bar{m}_{\bar{K}}\}$ as a subset of the index set $[k_2]$ and $\{\bar{m}_{\bar{K}+1}, \dots, \bar{m}_{k_2}\} := [k_2] \setminus \{\bar{m}_1, \bar{m}_2, \dots, \bar{m}_{\bar{K}}\}$. Then, we partition the input space w.r.t the density estimation $s_{G_1^n, G_2^n}(Y|X)$ as $\mathcal{X} = \cup_{\bar{m}=1}^{\bar{M}} \mathcal{X}_{\bar{m}}^n$, where the $\bar{m}$-th region is defined as

$$\mathcal{X}_{\bar{m}}^n := \Big\{x \in \mathcal{X} : (\beta_{1i}^n)^\top x \geq (\beta_{1i'}^n)^\top x, \ \forall i \in \{\bar{m}_1, \bar{m}_2, \dots, \bar{m}_{\bar{K}}\}, i' \in \{\bar{m}_{\bar{K}+1}, \dots, \bar{m}_{k_2}\}\Big\}$$

for any $\bar{m} \in [\bar{M}]$. Let $X\mathcal{X}_m^*$ for $m \in [M]$ such that $\{m_1, m_2, \dots, m_K\} = \{1, 2, \dots, K\}$. If there does not exist $\bar{m} \in [\bar{M}]$ such that $\{\bar{m}_1, \bar{m}_2, \dots, \bar{m}_{\bar{K}}\} = \mathcal{V}_{2,1} \cup \mathcal{V}_{2,2} \cup \dots \cup \mathcal{V}_{2,K}$, then the ratio $\mathbb{E}_X[V(\bar{s}_{G_1^n, G_2^n}(\cdot|X), s_{G_1^*, G_2^*}(\cdot|X))]/\mathcal{D}_{5n}$ does not converge to zero, which contradicts the result in equation (50). Thus, we can find $\bar{m} \in [\bar{M}]$ such that $\{\bar{m}_1, \bar{m}_2, \dots, \bar{m}_{\bar{K}}\} = \mathcal{V}_{2,1} \cup \mathcal{V}_{2,2} \cup \dots \cup \mathcal{V}_{2,K}$.

Since the Voronoi loss $\mathcal{D}_{5n}$ converges to zero, it follows that $\beta_{1i}^n \to \beta_{1j}^*$ for all $j \in [K]$ and $i \in \mathcal{V}_{2,j}$. Then, by means of Lemma 3, we deduce $\mathcal{X}_m^* = \mathcal{X}_{\bar{m}}^n$ for sufficiently large n, implying that $X \in \mathcal{X}_{\bar{m}}^n$. Therefore, we can represent the true density and the density estimation when the sample size n is sufficiently large as follows:

$$s_{G_1^*, G_2^*}(y|x) := \frac{1}{2} \sum_{i=1}^{k_1^*} \omega_i^* \pi(y|h_1(x, \kappa_i^*), \tau_i^*) + \frac{1}{2} \sum_{i=1}^{K} \frac{\exp((\beta_{1i}^*)^\top x + \beta_{0i}^*)}{\sum_{j=1}^{k_2^*} \exp((\beta_{1j}^*)^\top x + \beta_{0j}^*)} \cdot \pi(y|h_2(x, \eta_i^*), \nu_i^*),$$

$$s_{G_1^n, G_2^n}(y|x) := \frac{1}{2} \sum_{i=1}^{k_1^n} \omega_i^* \pi(y|h_1(x, \kappa_i^n), \tau_i^n) + \frac{1}{2} \sum_{i=1}^{\bar{K}} \frac{\exp((\beta_{1i}^n)^\top x + \beta_{0i}^n)}{\sum_{j=1}^{\bar{K}} \exp((\beta_{1j}^n)^\top x + \beta_{0j}^n)} \cdot \pi(y|h_2(x, \eta_i^n), \nu_i^n).$$

Given the above formulations, we can achieve the local part (48) by employing the same arguments used in Appendix D.1. Hence, the proof is completed. □

F.1 Proof of Lemma 3

Let us consider $\varepsilon_j = N_j \eta$, where $\eta > 0$ is some fixed constant, and $N_j > 0$ will be chosen later. Since the input space $\mathcal{X}$ and the parameter space Θ are bounded, there exists a constant $c_m^* \geq 0$ such that

$$\min_{x,j,j'} \Big[(\beta_{1j}^*)^\top x - (\beta_{1j'}^*)^\top x\Big] = c_m^* \eta, \tag{52}$$

where the above minimum is subject to $x \in \mathcal{X}_m^*, j \in \{m_1, m_2, \dots, m_K\}$ and $j' \in \{m_{K+1}, \dots, m_{k_2^*}\}$. We will show by contradiction that $c_m^* > 0$. Suppose that $c_m^* = 0$. For $x \in \mathcal{X}_m^*$, we may assume for any $1 \leq i < j \leq k_2^*$ that

$$(\beta_{1m_i}^*)^\top x \geq (\beta_{1m_j}^*)^\top x.$$

As $c^*_m = 0$, the result in equation (52) indicates that $(\beta^*_{1m_K})^\top x - (\beta^*_{1m_{K+1}})^\top x = 0$, or equivalently

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

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

$$\mathcal{N} := \{x \in \mathcal{X} : (\beta^*_{1m_K} - \beta^*_{1m_{K+1}})^\top x = 0\}.$$

Since the difference $\beta_{1m_K} - \beta_{1m_{K+1}}$ is non-zero and the input X follows a continuous distribution, then the set $\mathcal{N}$ has measure zero. Furthermore, as $\mathcal{X}^*_m \subseteq \mathcal{N}$, it follows that $\mathcal{X}^*_m$ also has measure zero, which contradicts the fact that it has non-zero measure. Thus, we must have $c^*_m > 0$.

Subsequently, let $x \in \mathcal{X}^*_m$ and $\bar{m} \in [\bar{M}]$ such that $\{\bar{m}_1, \bar{m}_2, \ldots, \bar{m}_{\bar{K}}\} = \mathcal{V}_{2,m_1} \cup \mathcal{V}_{2,m_2} \cup \ldots \cup \mathcal{V}_{2,m_K}$. We will demonstrate that $x \in \bar{\mathcal{X}}_{\bar{m}}$. Indeed, recall that the input space $\mathcal{X}$ is bounded, then we may assume that $\|x\| \leq B$ for any $x \in \mathcal{X}$, where $B > 0$ is some constant. Then, for any $i \in \{\bar{m}_1, \bar{m}_2, \ldots, \bar{m}_{\bar{K}}\}$ and $i' \in \{\bar{m}_{\bar{K}+1}, \ldots, \bar{m}_{k_2}\}$, we have

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

where $j \in \{m_1, m_2, \ldots, m_K\}$ and $j' \in \{m_{K+1}, \ldots, m_{k^*_2}\}$ such that $i \in \mathcal{V}_{2,j}$ and $i' \in \mathcal{V}_{2,j'}$. Note that if $N_j \leq \dfrac{c^*_m}{2B}$, then we obtain $x \in \mathcal{X}_{\bar{m}}$, which implies that $\mathcal{X}^*_m \subseteq \bar{\mathcal{X}}_{\bar{m}}$.

Analogously, assume that there exists some constant $c_m \geq 0$ such that

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

where the above minimum is subject to $x \in \bar{\mathcal{X}}_{\bar{m}}$, $i \in \{\bar{m}_1, \bar{m}_2, \ldots, \bar{m}_{\bar{K}}\}$ and $i' \in \{\bar{m}_{\bar{K}+1}, \ldots, \bar{m}_k\}$. Then, if $N_j \leq \dfrac{c_m}{2B}$, we have $\bar{\mathcal{X}}_{\bar{m}} \subseteq \mathcal{X}^*_m$. Consequently, by setting $N_j = \dfrac{1}{2B} \min\{c^*_m, c_m\}$, we reach the conclusion that $\bar{\mathcal{X}}_{\bar{m}} = \mathcal{X}^*_m$. Hence, the proof is completed.

F.2 Proof of Proposition 6

To begin with, we show that

$$\lim_{\varepsilon \to 0} \inf_{(G_1,G_2) \in \mathcal{G}_{k_1,k_2}(\Theta): \mathcal{D}_5((G_1,G_2),(G^*_1,G^*_2)) \leq \varepsilon} \mathbb{E}_X[V(\bar{s}_{G_1,G_2}(\cdot|X), s_{G^*_1,G^*_2}(\cdot|X))] > 0. \tag{53}$$

Suppose that the above inequality does not hold, then there exist a sequence of pairs of mixing measures (G^n_1, G^n_2) in $\mathcal{G}_{k_1,k_2}(\Theta)$ given by $G^n_1 = \sum_{i=1}^{k^n_1} \omega^n_i \delta_{(\kappa^n_i, \tau^n_i)}$ and $G^n_2 = \sum_{i=1}^{k^n_2} \exp(\beta^n_{0i}) \delta_{(\beta^n_{1i}, \eta^n_i, \nu^n_i)}$ that satisfies $\mathcal{D}_5((G^n_1, G^n_2), (G^*_1, G^*_2)) \to 0$ and

$$\mathbb{E}_X[V(\bar{s}_{G^n_1,G^n_2}(\cdot|X), s_{G^*_1,G^*_2}(\cdot|X))] \to 0$$

as $n \to \infty$. According to the Fatou's lemma, we have

$$0 = \lim_{n\to\infty} \mathbb{E}_X[V(\bar{s}_{G_1^n,G_2^n}(\cdot|X), s_{G_1^*,G_2^*}(\cdot|X))] \\ \geq \frac{1}{2}\int_{\mathcal{X}\times\mathcal{Y}} \liminf_{n\to\infty} |\bar{s}_{G_1^n,G_2^n}(Y|X) - s_{G_1^*,G_2^*}(Y|X)| \mathrm{d}(X,Y), \tag{54}$$

implying that $\bar{s}_{G_1^n,G_2^n}(Y|X) - s_{G_1^*,G_2^*}(Y|X) \to 0$ as $n \to \infty$ for almost surely (X, Y). WLOG, we may assume that

$$\max_{\{m_1,m_2,\ldots,m_K\}} \sum_{j=1}^{K} |\mathcal{V}_{2,m_j}| = |\mathcal{V}_{2,1}| + |\mathcal{V}_{2,2}| + \ldots + |\mathcal{V}_{2,K}|.$$

Let $X \in \mathcal{X}_m^*$, where $m \in [M]$ such that $\{m_1, m_2, \ldots, m_K\} = \{1, 2, \ldots, K\}$. Since the Voronoi loss $\mathcal{D}_5((G_1^n, G_2^n), (G_1^*, G_2^*))$ goes to zero, it follows that $\beta_{1i}^n \to \beta_{1j}^*$ as $n \to \infty$ for any $j \in [k_2^*]$ and $i \in \mathcal{V}_{2,j}$. By means of Lemma 3, we deduce $X \in \bar{\mathcal{X}}_{\bar{m}}$, where $\bar{m} \in [\bar{q}]$ such that $\{\bar{m}_1, \bar{m}_2, \ldots, \bar{m}_{\bar{K}}\} = \mathcal{V}_{2,1} \cup \mathcal{V}_{2,2} \cup \ldots \cup \mathcal{V}_{2,K}$. However, as $\bar{K} < \sum_{j=1}^{K} |\mathcal{V}_{2,j}|$, the fact that $\{\bar{m}_1, \bar{m}_2, \ldots, \bar{m}_{\bar{K}}\} = \mathcal{V}_{2,1} \cup \mathcal{V}_{2,2} \cup \ldots \cup \mathcal{V}_{2,K}$ cannot occur. Thus, we obtain the result in equation (53). As a consequence, we can find a positive constant ε' such that

$$\inf_{(G_1,G_2)\in\mathcal{G}_{k_1,k_2}(\Theta):\mathcal{D}_5((G_1,G_2),(G_1^*,G_2^*))\leq\varepsilon'} \mathbb{E}_X[V(\bar{s}_{G_1,G_2}(\cdot|X), s_{G_1^*,G_2^*}(\cdot|X))] > 0.$$

Given the above result, it is sufficient to show that

$$\inf_{(G_1,G_2)\in\mathcal{G}_{k_1,k_2}(\Theta):\mathcal{D}_5((G_1,G_2),(G_1^*,G_2^*))>\varepsilon'} \mathbb{E}_X[V(\bar{s}_{G_1,G_2}(\cdot|X), s_{G_1^*,G_2^*}(\cdot|X))] > 0. \tag{55}$$

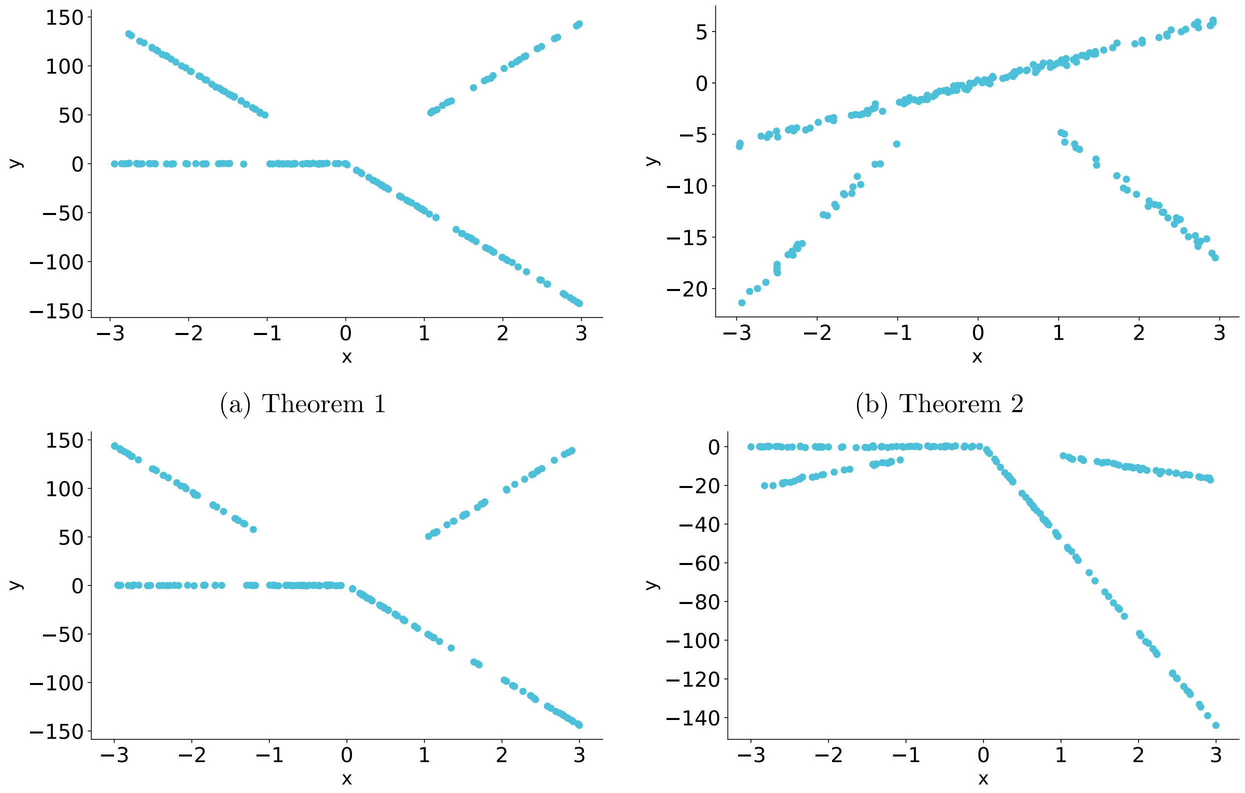
Assume by contrary that the inequality (55) does not hold, then we can find a sequence $(\tilde{G}_1^n, \tilde{G}_2^n) \in \mathcal{G}_{k_1,k_2}(\Theta)$ such that $\mathcal{D}_5((\tilde{G}_1^n, \tilde{G}_2^n), (G_1^*, G_2^*)) > \varepsilon'$ and

$$\mathbb{E}_X[V(\bar{s}_{\tilde{G}_1^n,\tilde{G}_2^n}(\cdot|X), s_{G_1^*,G_2^*}(\cdot|X))] \to 0.$$

Again, by utilizing the Fatou's lemma as in equation (54), we get $\bar{s}_{\tilde{G}_1^n,\tilde{G}_2^n}(Y|X) - s_{G_1^*,G_2^*}(Y|X) \to 0$ as $n \to \infty$ for almost surely (X, Y). Since the parameter space Θ is compact, we can substitute the sequence $(\tilde{G}_1^n, \tilde{G}_2^n)$ with its subsequence which converges to some pair of mixing measures $(\tilde{G}_1, \tilde{G}_2)$ in $\mathcal{G}_{k_1,k_2}(\Theta)$. This result leads to $\bar{s}_{\tilde{G}_1,\tilde{G}_2}(Y|X) = s_{G_1^*,G_2^*}(Y|X)$ for almost surely (X, Y). As the Top-K sparse gating MoE is identifiable, we deduce $(\tilde{G}_1, \tilde{G}_2) \equiv (G_1^*, G_2^*)$, or equivalently, $\mathcal{D}_5((\tilde{G}_1, \tilde{G}_2), (G_1^*, G_2^*)) = 0$. On the other hand, due to the fact that $\mathcal{D}_5((\tilde{G}_1^n, \tilde{G}_2^n), (G_1^*, G_2^*)) > \varepsilon'$ for any $n \in \mathbb{N}$, we obtain $\mathcal{D}_5((\tilde{G}_1, \tilde{G}_2), (G_1^*, G_2^*)) > \varepsilon' > 0$, which contradicts the previous result. Hence, we reach the result in equation (55) and complete the proof.

G Additional Experiments

In this appendix, we provide supplementary experimental results that reinforce and extend our theoretical analyses. In Appendix G.1, we illustrate the convergence properties of the maximum likelihood estimator (MLE) $(\widehat{G}_1^n, \widehat{G}_2^n)$ towards the true mixing measure (G_1^*, G_2^*) using synthetic data, explicitly evaluating four theorem-based scenarios. Appendix G.2 and Appendix G.3 provide detailed training and validation performance curves during training of each model in language modeling and vision-language modeling, respectively.



(a) Theorem 1

(b) Theorem 2

(c) Theorem 3

(d) Theorem 4

Figure 5: Empirical illustration of the input - output relationship (X, Y) under synthetic conditions for each theoretical result. Each subplot corresponds to a different theoretical setting: (a) Theorem 1, (b) Theorem 2, (c) Theorem 3, and (d) Theorem 4.

G.1 Numerical Experiments

G.1.1 Experimental Setup

Synthetic Data. For each sample size n, we generate i.i.d samples $\{(X_i, Y_i)\}_{i=1}^n$ by first sampling X_i's from the uniform distribution Uniform$[-3, 3]$ and then sampling Y_i's from the true conditional density $f_{G_1^*, G_2^*}(Y|X)$ or $g_{G_1^*, G_2^*}(Y|X)$ of Gaussian mixture of experts (MoE) model setting of each theorem configuration. Figure 5 shows the visualization of the relationship between X and Y in each experiment.

Maximum Likelihood Estimation (MLE). A popular approach to determining the MLE $(\widehat{G}_1^n, \widehat{G}_2^n)$ for each set of samples is to use the Expectation-Maximization (EM) algorithm [16]. 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 [5]. We select the convergence criterion of $\epsilon = 10^{-6}$ and run a maximum of 1000 EM iterations.

Experiment Design. Our empirical investigation systematically examines four experimental configurations, each precisely corresponding to the theoretical scenarios elaborated in our main paper.

Each configuration includes 40 independent sample generations over a comprehensive range of sample sizes n, specifically $n \in [10^2, 10^5]$. To ensure consistency and comparative clarity across experiments, we uniformly adopt an architecture consisting of one shared expert ($k_1^* = 1$) complemented by two routed experts ($k_2^* = 2$), where we fit two shared experts ($k_1 = 2$) and three routed experts ($k_2 = 3$) in our experiment settings.

G.1.2 Theorem 1

The problem setting is defined in Equation 1, where we establish h_1 and h_2 to satisfy the identifiable experts condition, specifically $h_1(x, (\kappa_2, \kappa_1, \kappa_0)) := \kappa_2 \mathrm{ReLU}(\kappa_1^\top x + \kappa_0)$ and $h_2(x, (\eta_2, \eta_1)) := \eta_2 \mathrm{ReLU}(\eta_1^\top x)$. The ground-truth parameters employed in our experiments are presented as follows:

$$\begin{array}{lllll}
\omega^* = 1.0, & \kappa_0^* = 0, & \kappa_1^* = 6, & \kappa_2^* = -8, & \tau^* = 0.25, \\
\beta_{01}^* = -0.5, & \beta_{11}^* = 5, & \eta_{11}^* = -12, & \eta_{21}^* = 4, & \nu_1^* = 0.4, \\
\beta_{02}^* = 0.5, & \beta_{12}^* = 5, & \eta_{12}^* = 12, & \eta_{22}^* = 4, & \nu_2^* = 0.4,
\end{array}$$

As illustrated in Figure 6a, the maximum likelihood estimator MLE $(\widehat{G}_1^n, \widehat{G}_2^n)$ exhibits empirical convergence to the true mixing measure G^* under the Voronoi metric $\mathcal{D}_1$ (Equation 4) at the rate of order $\mathcal{O}_P([\log(n)/n]^{0.451})$. This empirically observed rate closely matches the theoretical parametric convergence rate $\mathcal{O}_P([\log(n)/n]^{1/2})$ established in Theorem 1, thereby validating the practical applicability of the theoretical result under strongly identifiable expert assumptions.

G.1.3 Theorem 2

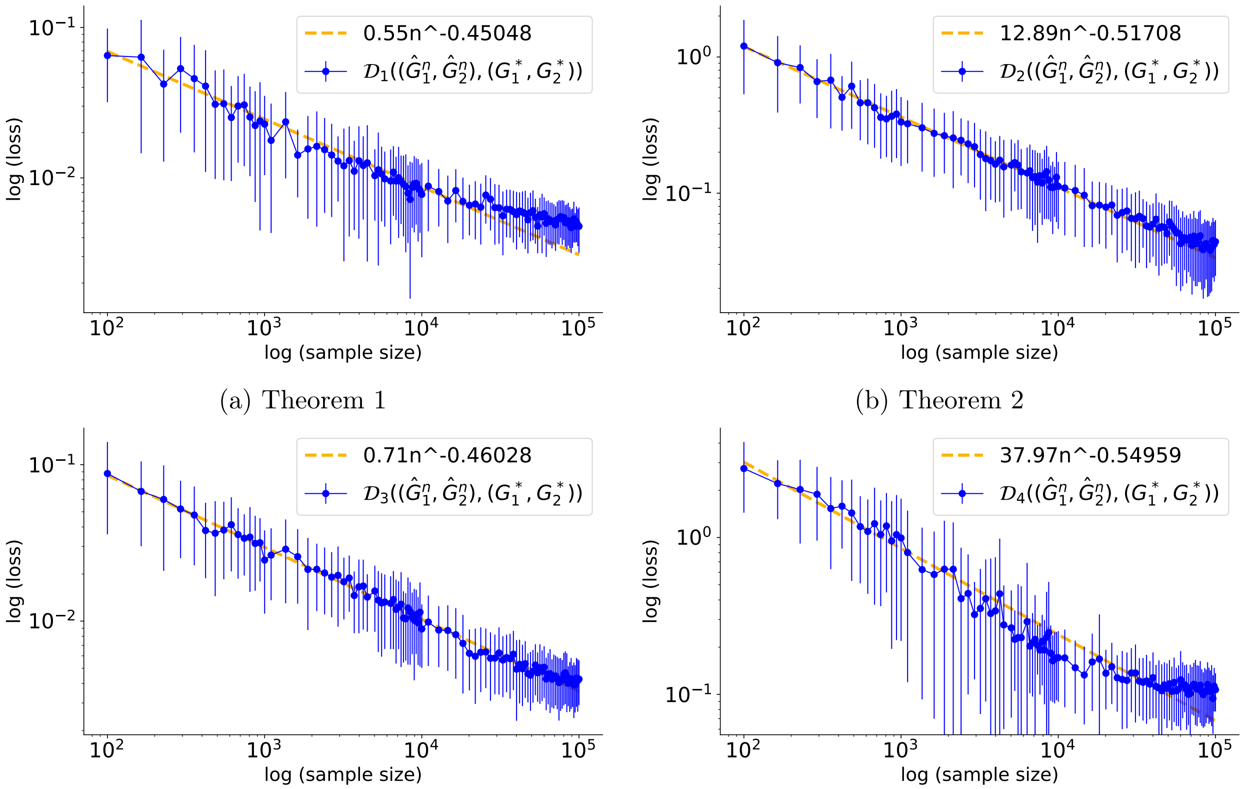
In this experiment, we adopt the problem setting outlined in Theorem 1. However, instead of using two-layer FFNs, we define h_1 and h_2 are linear experts as in Section 2.2. Specifically, we set $h_1(X, (\kappa_1, \kappa_0)) := \kappa_1^\top X + \kappa_0$ and $h_2(X, (\eta_1, \eta_0)) := \eta_1^\top X + \eta_0$, with the associated ground-truth parameters defined as follows:

$$\begin{array}{lllll}
\omega^* = 1.0, & \kappa_0^* = 0, & \kappa_1^* = 2, & \tau^* = 0.2, & \\
\beta_{01}^* = -0.5, & \beta_{11}^* = 5, & \eta_{11}^* = 8, & \eta_{01}^* = 2, & \nu_1^* = 0.4, \\
\beta_{02}^* = 0.5, & \beta_{12}^* = 5, & \eta_{12}^* = -6, & \eta_{02}^* = 1, & \nu_2^* = 0.4,
\end{array}$$

The result is shown in Figure 6b. Under linear experts settings and Voronoi metric $\mathcal{D}_2$ (Equation 6), the maximum likelihood estimator MLE has the convergence rate of $\mathcal{O}_P([\log(n)/n]^{1/2})$. Notably, the linear expert settings make a perfect result with convergence rate of $\mathcal{O}_P([\log(n)/n]^{0.517})$ where the noise in each sample size is minimal and uniform. This result strongly supports our theoretical result in Theorem 2.

G.1.4 Theorem 3

This experiment is designed to empirically validate Theorem 3 under the problem setting specified in Appendix A, which employs normalized sigmoid gating. Under the sparse regime, we set all



(a) Theorem 1

(b) Theorem 2

(c) Theorem 3

(d) Theorem 4

Figure 6: Log-log scaled plots illustrating simulation results with different model settings. The blue curves depict the mean discrepancy between the MLE $(\widehat{G}_1^n, \widehat{G}_2^n)$ and the true mixing measure (G_1^*, G_2^*) accompanied by error bars representing the standard deviation over 40 times of experiments for each sample size n. Additionally, an orange dash-dotted line represents the least-squares fitted linear regression line for these data points.

over-specified parameters β_{1i}^* equal to zero vectors. The expert functions follow the same structural assumptions as in Theorem 1 where $h_1(x, (\kappa_2, \kappa_1, \kappa_0)) := \kappa_2 \mathrm{ReLU}(\kappa_1^\top x + \kappa_0)$ and $h_2(x, (\eta_2, \eta_1)) := \eta_2 \mathrm{ReLU}(\eta_1^\top x)$. The complete set of ground-truth parameters used in this experiment is detailed below:

$$\begin{array}{lllll}
\omega^* = 1.0, & \kappa_0^* = 0, & \kappa_1^* = 6, & \kappa_2^* = -8, & \tau^* = 0.25, \\
\beta_{01}^* = -0.5, & \beta_{11}^* = 0, & \eta_{11}^* = -12, & \eta_{21}^* = 4, & \nu_1^* = 0.4, \\
\beta_{02}^* = 0.5, & \beta_{12}^* = 0, & \eta_{12}^* = 12, & \eta_{22}^* = 4, & \nu_2^* = 0.4,
\end{array}$$

Figure 6c presents the experimental results for the convergence analysis under the sparse regime utilizing normalized sigmoid gating. The maximum likelihood estimator (MLE) $(\widehat{G}_1^n, \widehat{G}_2^n)$ empirically converges to the true mixing measure (G_1^*, G_2^*) at a rate of $\mathcal{O}_P([\log(n)/n]^{0.46})$ under the Voronoi metric $\mathcal{D}_3$ (Equation 10). This empirical convergence rate is closely aligned with the theoretical prediction articulated in Theorem 3. Consistent with the theorem's implications, our experimental results suggest that under the sparse regime, normalized sigmoid gating does not exhibit significant advantages in terms of convergence speed compared to standard softmax gating mechanisms.

G.1.5 Theorem 4

In this experiment, we adopt the same problem setting with Theorem D.4 specified in Appendix A. With sigmoid gating under the dense regime, we define shared expert function h_1 is strongly identifiable while the routed expert function h_2 is weakly identifiable. Specifically, h_1 is the two-layer FFNs function $h_1(x, (\kappa_2, \kappa_1, \kappa_0)) := \kappa_2 \mathrm{ReLU}(\kappa_1^\top x + \kappa_0)$ where h_2 is the linear experts $h_2(X, (\eta_1, \eta_0)) := \eta_1^\top X + \eta_0$. The complete set of ground-truth parameters used in this experiment is detailed below:

$$\begin{aligned}
&\omega^* = 1.0, && \kappa_0^* = 0, && \kappa_1^* = 6, && \kappa_2^* = -8, && \tau^* = 0.25,\\
&\beta_{01}^* = -0.5, && \beta_{11}^* = 5, && \eta_{11}^* = 8, && \eta_{01}^* = 2, && \nu_1^* = 0.4,\\
&\beta_{02}^* = 0.5, && \beta_{12}^* = 5, && \eta_{12}^* = -6, && \eta_{02}^* = 1, && \nu_2^* = 0.4,
\end{aligned}$$

Figure 6d presents the numerical results corresponding to Theorem D.4. Under the dense regime, the Mixture-of-Experts (MoE) model achieves a convergence rate of $\mathcal{O}_P([\log(n)/n]^{0.552})$, which closely aligns with the theoretical rate of $\mathcal{O}_P([\log(n)/n]^{1/2})$. This empirical evidence substantiates Theorem D.4, suggesting that the use of normalized sigmoid gating contributes to improved sample efficiency in DeepSeekMoE.

G.2 Language Modeling

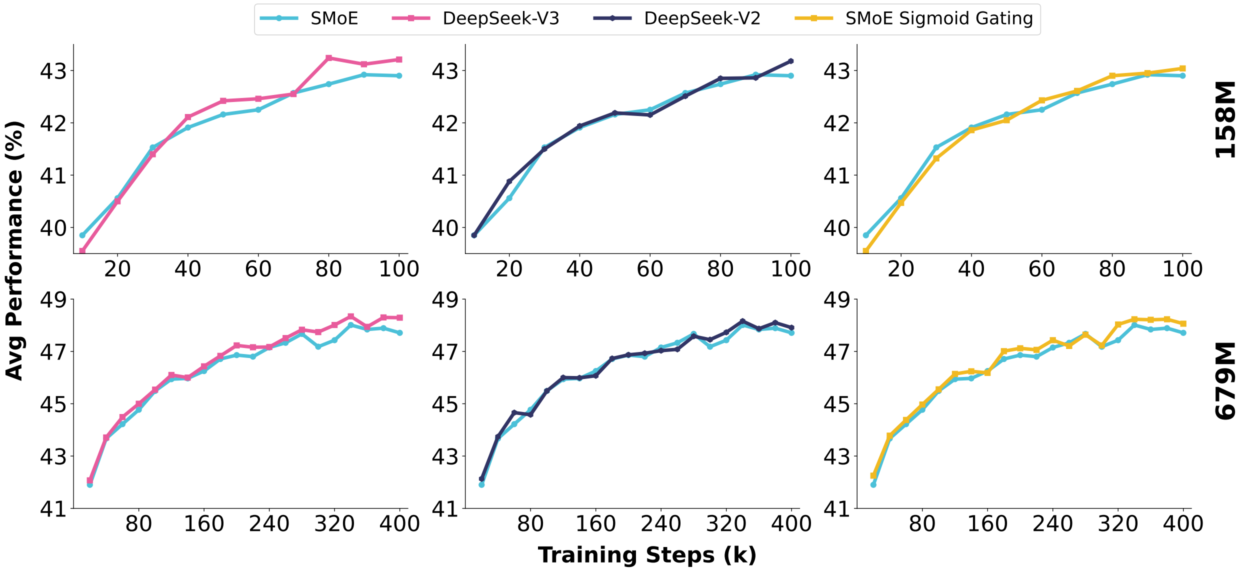


Figure 7: Average performance (%) compared in pairs with Vanilla SMoE across three model settings over training steps on language modeling tasks. **Left:** Vanilla SMoE vs. DeepSeek-V3; **Center:** Vanilla SMoE vs. DeepSeek-V2; **Right:** Vanilla SMoE vs. SMoE Sigmoid Gating.

Figure 7 presents a pairwise comparison between DeepSeek-V3, DeepSeek-V2, SMoE Sigmoid Gating, and the baseline Vanilla SMoE. Remarkably, across both model scales, by integrating normalized sigmoid gating into SMoE, SMoE Sigmoid Gating yields a substantial improvement in convergence rate compared to the softmax-gated baseline. Notably, in several training trajectories, SMoE Sigmoid Gating achieves a convergence rate comparable to that of DeepSeek-V2. For a more detailed examination, we provide the full training benchmark curves for both the 158M and 679M parameter language modeling settings in Figure 14 and Figure 15, respectively.

G.3 Vision-Language Modeling

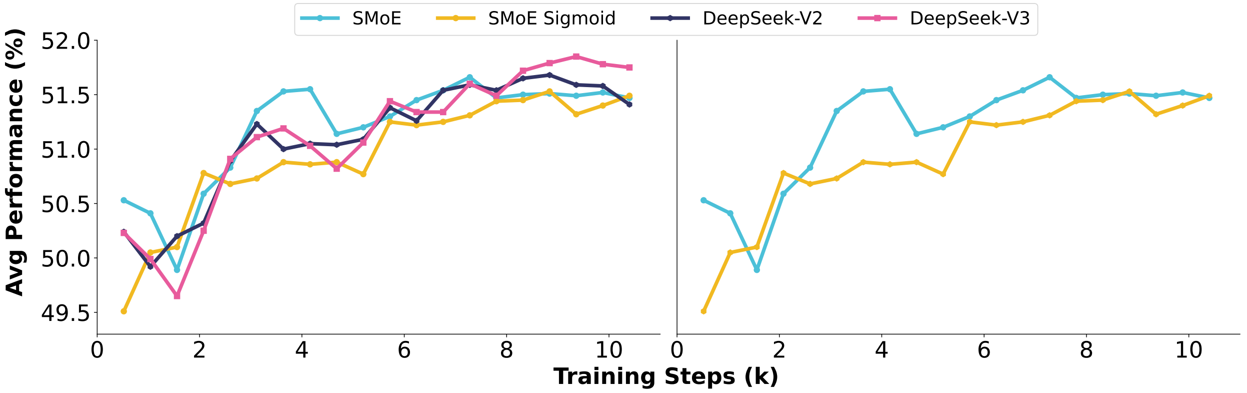


Figure 8: Average performance (%) over training steps on vision-language pretraining tasks, comparing SMoE variants across three model configurations. **Left:** Full comparison among Vanilla SMoE, SMoE with Sigmoid Gating, DeepSeek-V2, and DeepSeek-V3; **Right:** Focused comparison between Vanilla SMoE and SMoE Sigmoid Gating.

Figure 7 presents a pairwise comparison among DeepSeek-V3, DeepSeek-V2, SMoE with Sigmoid Gating, and the baseline Vanilla SMoE. On vision-language pretraining tasks, SMoE Sigmoid Gating exhibits a comparable convergence rate and final performance to the Vanilla SMoE. However, similar to the DeepSeek variants, it demonstrates faster convergence during the later stages of training and achieves greater training stability. To facilitate a finer-grained analysis, we provide benchmark-specific performance trajectories in Figure 16.

H Additional Router Analysis

In this appendix, we provide further analyses regarding router behavior. Formal definitions, equations, and detailed discussions on router saturation and router change rate are provided in Appendix H.1 and Appendix H.2, respectively. Additionally, an in-depth analysis of expert utilization is included in Appendix H.3. For consistency, all analyses utilize the same ordered set of the 6000 most frequent tokens from the validation dataset.

H.1 Router Saturation

In formal terms, router saturation is the proportion of expert activations at some intermediary checkpoint at time t that matches the expert IDs activated at some final checkpoint T over the same

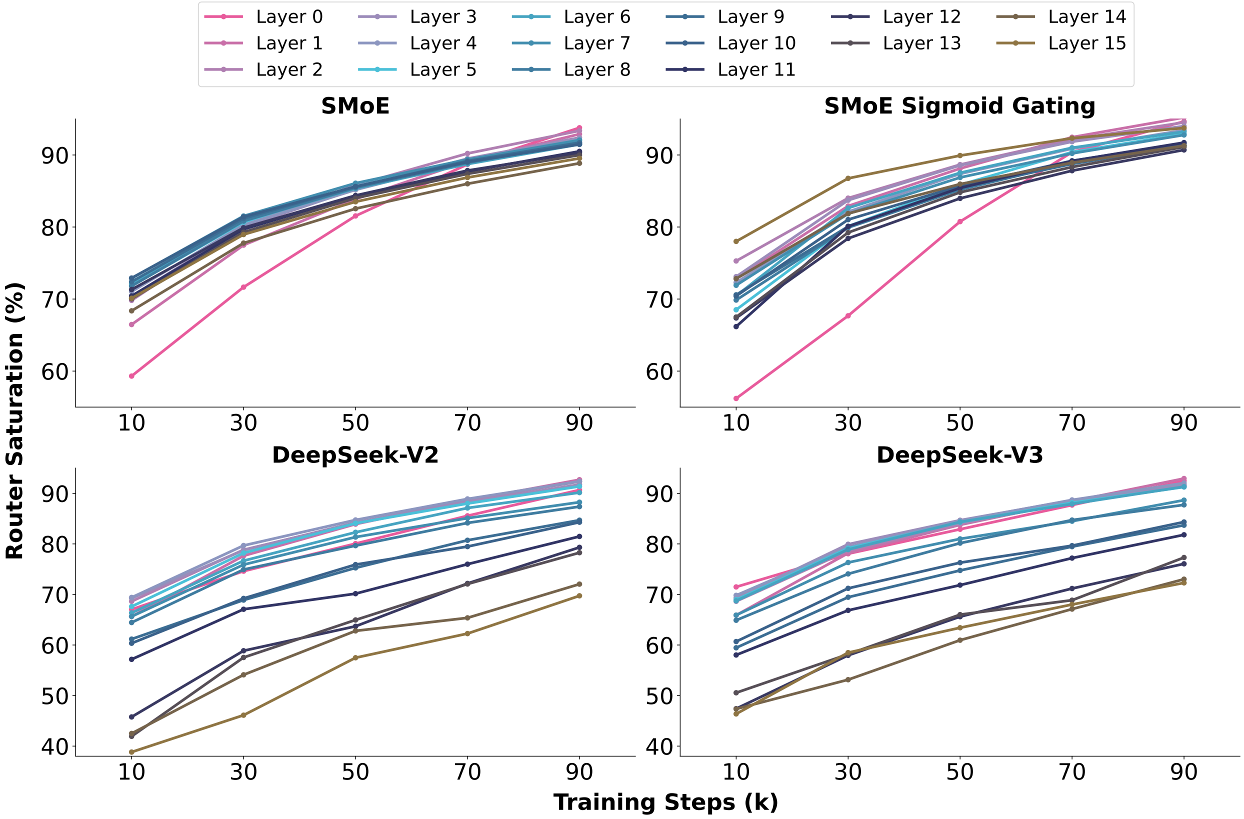


Figure 9: Router saturation across layers for 158M-parameter models in language modeling tasks. We compute saturation by comparing the routing to the top-8 experts with SMoE and SMoE Sigmoid Gating, and the top-6 experts with DeepSeek variants.

dataset:

$$\text{Router Saturation}(t) = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|\mathcal{E}_i^{(t)} \cap \mathcal{E}_i^{(T)}\right|}{k},$$

where:

- N: The total number of tokens in the dataset.
- k: The number of top-k experts activated per input token.
- $\mathcal{E}_i^{(t)}$: The set of k experts activated for the i-th token at the t-th checkpoint.
- $\mathcal{E}_i^{(T)}$: The set of k experts activated for the i-th token at the final checkpoint T.
- $\left|\mathcal{E}_i^{(t)} \cap \mathcal{E}_i^{(T)}\right|$: The number of common experts activated for the i-th token between the t-th and final checkpoints.

Router saturation provides a quantitative measure of how early the routing decisions converge during training. A saturation value of 100% indicates that the router at an intermediate checkpoint routes to the same set of experts as at the final checkpoint. High saturation values at early checkpoints reflect early convergence in expert selection, indicating that the router has rapidly settled into a stable assignment pattern. In contrast, low saturation values suggest ongoing exploration or adaptation in expert allocations, signaling that the routing mechanism is still undergoing significant adjustments.

Figure 9 and Figure 11 show the detailed router saturation for each layer with 158M and 679M parameters, respectively. The result shows that the later layer tends to saturate earlier during training, where layer 0 is an outlier and saturates significantly slower than the others. Additionally, we observe that in shared layer settings (DeepSeek-V2 and DeepSeek-V3), the gap between saturation of different layers is smaller than SMoE and SMoE Sigmoid Gating. When comparing the MoE model with normalized sigmoid gating and softmax gating, we can easily observe that the model with normalized sigmoid gating exhibits a more uniform saturation rate across layers compared to the model with softmax gating. This observation further highlights the effectiveness of normalized sigmoid gating in mixture-of-experts model.

H.2 Router Change Rate

Router Change Rate is a metric that measures the stability of the gating of mixture of experts models. This metric directly quantifies the gating fluctuation between two consecutive checkpoints:

$$\text{Router Change Rate}(t) = \frac{1}{N} \sum_{i=1}^{N} \frac{\left| \mathcal{E}_i^{(t+1)} \backslash \mathcal{E}_i^{(i)} \right|}{k},$$

Where:

- N: The total number of tokens in the dataset.
- k: The number of top-k experts activated per input token.
- $\mathcal{E}_i^{(t)}$: The set of k experts activated for the i-th token at the t-th checkpoint.
- $\mathcal{E}_i^{(T)}$: The set of k experts activated for the i-th token at the $(t+1)$-th checkpoint.
- $\left| \mathcal{E}_i^{(t)} \backslash \mathcal{E}_i^{(T)} \right|$: The number of non-intersecting experts activated for the i-th token between the $(t+1)$-th and the t-th checkpoint

Router Change Rate is a quantitative metric to measure the stability of routing mechanism in Mixture-of-Experts (MoE) during training. Unlike router saturation, which assesses convergence towards a final routing decision, the router change rate evaluates fluctuations between consecutive checkpoints. A low router change rate indicates stable routing decisions across training intervals, implying that the gating mechanism has achieved consistent expert assignments, minimizing disruptions and promoting steady specialization of experts. Conversely, a high router change rate suggests volatility in routing decisions, reflecting ongoing exploration or adjustment, potentially introducing training inefficiencies and hindering expert specialization. Thus, monitoring the router

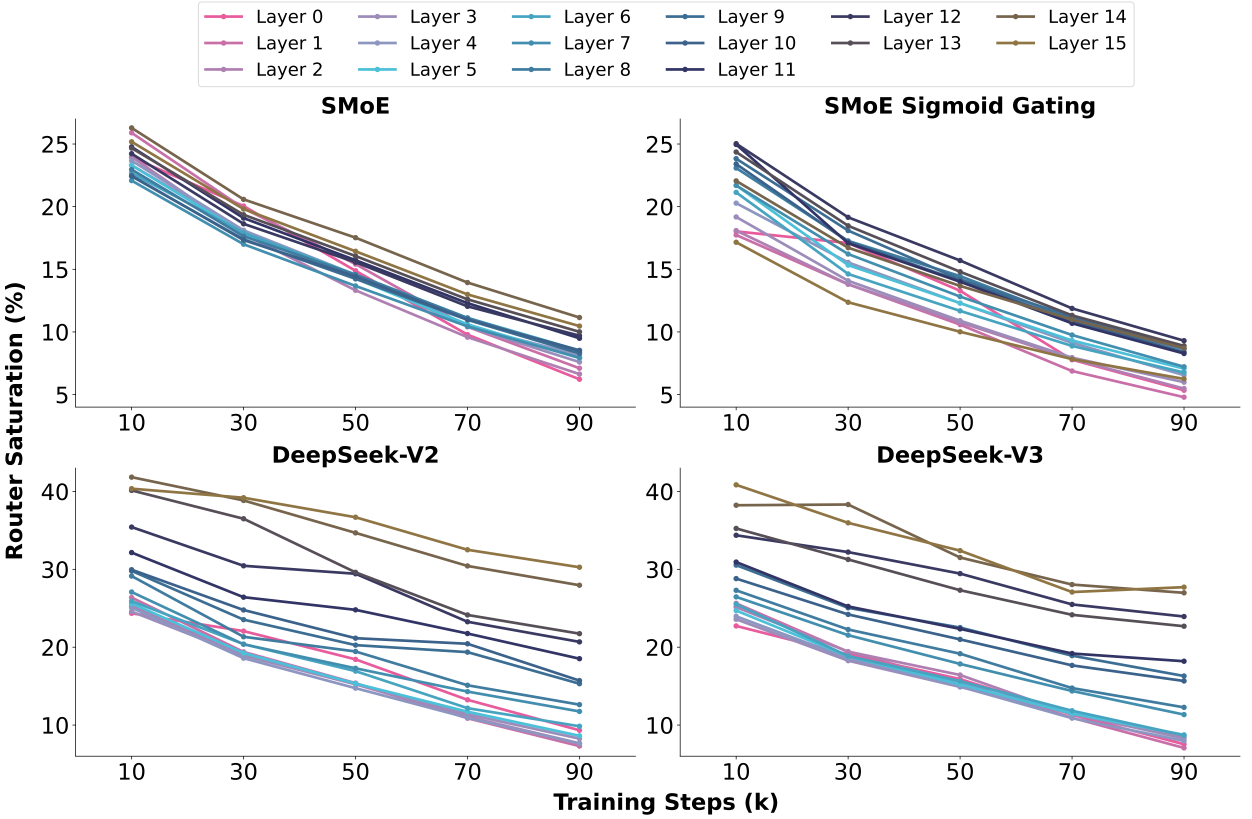


Figure 10: Router change rate across layers for 158M-parameter models in language modeling tasks. We compute router change rate by comparing the routing to the top-8 experts with SMoE and SMoE Sigmoid Gating, and the top-6 experts with DeepSeek variants.

change rate provides valuable insights into the dynamics of expert allocation stability, enabling deeper understanding and optimization of the routing strategy in MoE architectures.

Figure 9 and Figure 11 show the detailed router change rate for each layer with 158M and 679M parameters, respectively. Similar to router saturation, later layers show more stability with lower router change rate. However, the router change rate between layers show more consistency compared to router saturation. Layer 0 still has some differences in router change rate, the difference with other layers is still not too large, which show that although layer 0 saturates significantly slower, it still keep the stability of during training. When comparing between different model settings, the model with normalized sigmoid gating (SMoE Sigmoid Gating and DeepSeek-V3) shows lower and more consistent router change rate compared with the model with traditional softmax gating (SMoE, DeepSeek-V2).

H.3 Expert Utilization

To quantify the fairness of expert utilization in the Mixture-of-Experts (MoE) model, we apply Jain's Fairness Index to the router's resource allocation across n experts. Let $R = (r_1, r_2, \ldots, r_n)$ denote the utilization vector, where $r_i \geq 0$ represents the proportion of input tokens (or total routing

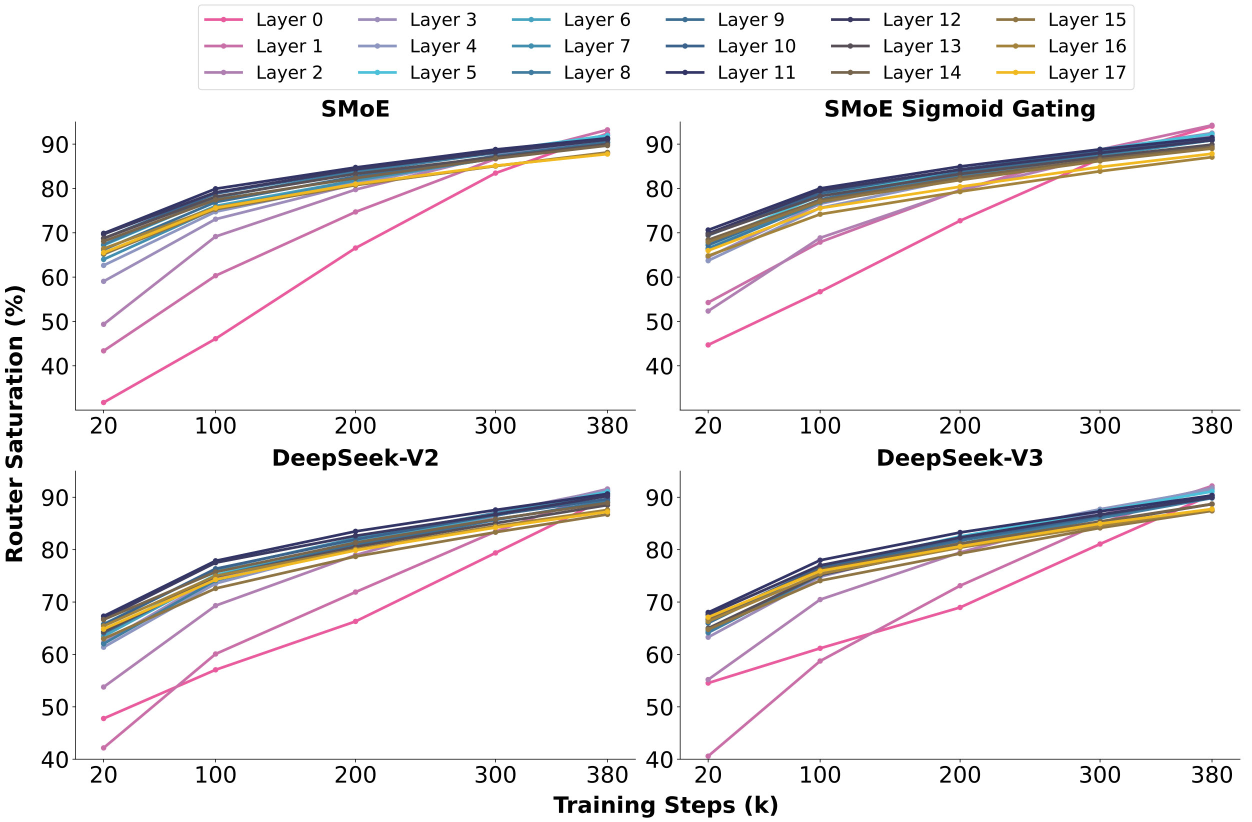


Figure 11: Router saturation across layers for 679M-parameter models in language modeling tasks. We compute saturation by comparing the routing to the top-8 experts with SMoE and SMoE Sigmoid Gating, and the top-6 experts with DeepSeek variants.

weight) assigned to expert i over a given evaluation window. The Jain's Fairness Index $J(R)$ is computed as:

$$J(R) = J(r_1, r_2, ..., r_n) = \frac{(\sum_{i=1}^{n} r_i)^2}{n \sum_{i=1}^{n} r_i^2},$$

This index ranges from $[1/n, 1]$, where $J(R) = 1$ indicates perfectly uniform expert usage, (i.e., all experts are used equally), where $J(R) = 1/n$ signifies complete imbalance, with only one expert active. Thus, higher values of $J(R)$ correspond to fairer and more evenly distributed expert selection.

Figure 13 presents a comparison of Jain's Fairness Index [29] across different Mixture-of-Experts (MoE) model configurations and scales. Across both 158M and 679M parameter models, all configurations exhibit a consistent pattern: fairness in expert utilization is highest in the initial layers and declines in subsequent layers, suggesting that earlier layers facilitate broader expert utilization. Notably, models employing normalized sigmoid gating (SMoE Sigmoid Gating and DeepSeek-V3) maintain a higher fairness index, especially in the later layers, indicating better expert utilization. These results highlight the efficacy of normalized sigmoid gating in promoting more balanced expert utilization throughout the network.

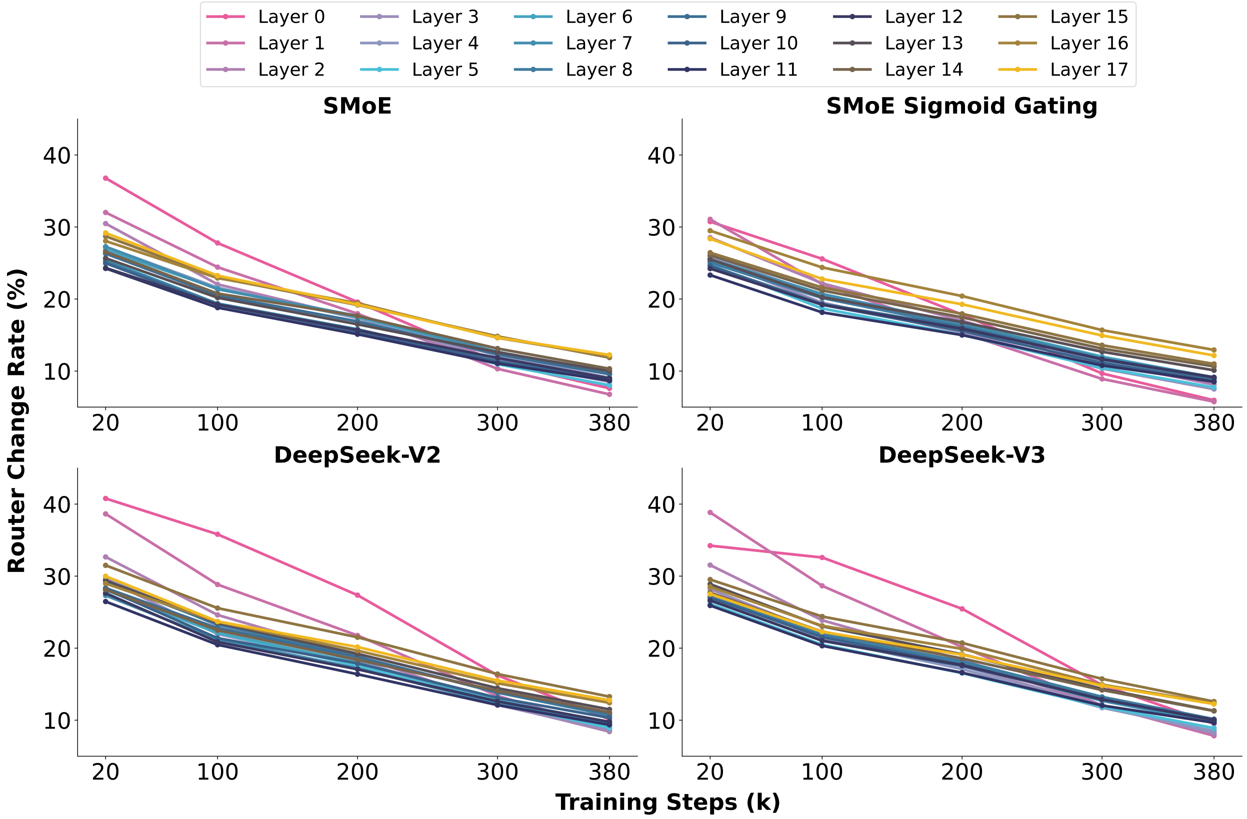


Figure 12: Router change rate across layers for 679M-parameter models in language modeling tasks. We compute router change rate by comparing the routing to the top-8 experts with SMoE and SMoE Sigmoid Gating, and the top-6 experts with DeepSeek variants.

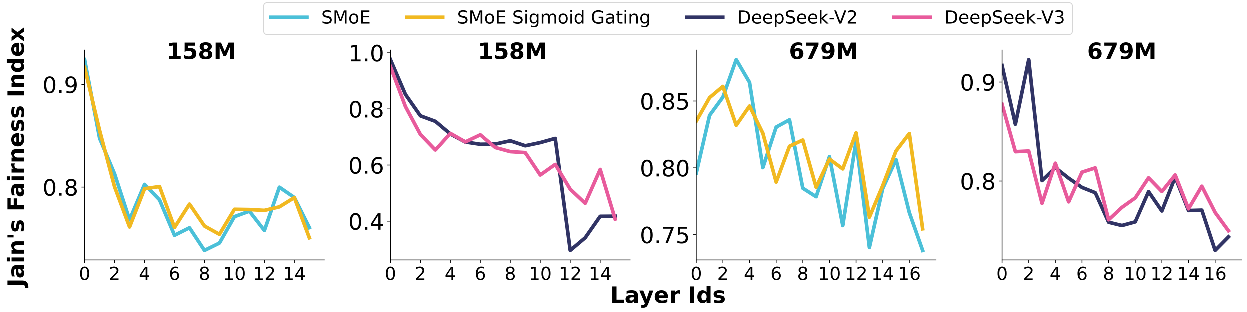


Figure 13: Jain's Fairness Index across MoE layers for language-modeling tasks with 158 M (left) and 679 M (right) parameter models.

I Experimental Details

I.1 Language Modeling

I.1.1 Datasets

SlimPajama. The SlimPajama [66] dataset is a filtered and deduplicated corpus of the 1.2T token RedPajama dataset [75] designed for language model pretraining. It contains around 627B tokens across diverse sources.

LAMBADA. The LAMBADA [58] dataset evaluates a model's ability to predict the final word of a passage, requiring understanding of broad discourse context. Each instance comprises a narrative where the target word is only predictable when considering the entire passage, challenging models to perform deep contextual comprehension beyond sentence-level cues

BLiMP. The Benchmark of Linguistic Minimal Pairs (BLiMP) [74] assesses language models' grasp of English grammar through 67 sub-datasets, each containing 1,000 minimal pairs. These pairs differ subtly to test specific syntactic, morphological, or semantic phenomena, enabling fine-grained evaluation of linguistic competence

Children's Book Test (CBT). CBT [24] measures a model's ability to utilize wider linguistic context by providing passages from children's books with a missing word to predict. The dataset distinguishes between predicting syntactic function words and semantically rich content words, emphasizing the importance of context in language understanding

HellaSwag. HellaSwag [81] challenges models with sentence completion tasks that require commonsense reasoning. Each instance presents a context and multiple plausible continuations, with only one being correct. The dataset is adversarially filtered to be trivial for humans but difficult for models, highlighting gaps in machine commonsense understanding.

PIQA. The Physical Interaction Question Answering (PIQA) [3] dataset tests models on physical commonsense reasoning. It comprises questions about everyday tasks, requiring knowledge of physical properties and affordances, challenging models to reason about the physical world without direct sensory experience.

ARC-Challenge. The AI2 Reasoning Challenge (ARC) [11] presents grade-school level multiple-choice science questions that necessitate reasoning and external knowledge. The Challenge set includes questions that are particularly difficult for models, serving as a benchmark for advanced question-answering capabilities .

OpenBookQA. OpenBookQA [35] consists of multiple-choice questions derived from a curated set of science facts, resembling open-book exams. Answering requires combining the provided facts with external commonsense knowledge, testing a model's ability to integrate information from multiple sources.

RACE. The Reading Comprehension Dataset from Examinations (RACE) [63] contains passages and questions from English exams for Chinese middle and high school students. With nearly 100,000 questions, it evaluates a model's reading comprehension and reasoning skills across diverse topics.

SIQA. Social IQa (SIQA) [63] focuses on social commonsense reasoning, presenting questions about everyday social interactions. Models must infer motivations, reactions, and social dynamics, challenging their understanding of human social behavior.

CommonSenseQA. CommonSenseQA [68] is a multiple-choice question-answering dataset that requires models to apply commonsense knowledge. Each question is designed to probe a specific aspect of commonsense reasoning, with distractor answers carefully crafted to be plausible yet incorrect.

I.1.2 Model Settings, Training Settings and Evaluation

Table 4: Comprehensive Model Configurations for Experimental Evaluation. SMoE refers to settings applied for both Vanilla SMoE and SMoE Sigmoid Gating, whereas DeepSeek corresponds to configurations used for DeepSeek-V2 and DeepSeek-V3 models.

Scale	Model	# params	# act. params	# trained tokens	d_{model}	H	d_{head}	N_E	K_r	N_s	Expert dim	N_{warmup}	κ
Small	**SMoE**	158M	36M	6.5B	512	4	82	66	8	0	128	0	0.1
	DeepSeek	158M	36M	6.5B	512	4	82	64	6	2	128	0	0.1
Large	**SMoE**	679M	131M	26.2B	1024	4	128	66	8	0	256	4000	0.25
	DeepSeek	679M	131M	26.2B	1024	4	128	64	6	2	256	4000	0.25

Training datasets. We conduct the experiments on language modeling using the popular SLimPajama [66] dataset. Due to the limited computational resource, we utilize only subsets of the SlimPajama [66] dataset containing 6.5B and 26.2B tokens to train our 158M and 679M parameter models, respectively.

Model Settings. Table 4 summarizes the comprehensive set of hyperparameters and configurations for both scales and the two model variants evaluated in our experiments. All models employ a total of $N_r = 66$ experts. For routing schemes, the baseline SMoE utilizes a top-8 expert routing strategy ($K_r = 8$), while the DeepSeek variants adopt a mixed routing approach comprising top-6 expert selection ($K_r = 6$) plus $N_s = 2$ shared experts. To align with the fine-grained expert segmentation proposed in DeepSeekMoE [12], we set the expert dimensionality to $1/4\ d_{model}$ and increase the expert count to 66 instead of the common settings with 16 experts. Additionally, the number of attention heads is uniformly set to $H = 4$ across both model scales. All models leverage Rotary Positional Embedding (RoPE) [67], PyTorch's optimized attention implementation, and employ pre-layernorm Transformers. To ensure balanced expert utilization, we use the standard load balancing loss defined in Switch Transformers [21].

Training Settings. All models are trained in PyTorch using a batch size of 64, context length of 1024, and a learning rate of $2.5e - 4$. We apply 4000 linear warm-up steps specifically for the larger-scale models and utilize the AdamW optimizer [46] with its default hyperparameters and a weight decay of 0.01. Gradient clipping is performed with threshold κ, and the precise number of linear warm-up steps (N_{warmup}) per model variant is provided in Table 4. We tokenize the input using SentencePiece [33], configured with a vocabulary size of 8000 tokens, which is trained on a representative subset of the SlimPajama dataset [66].

Evaluation. We evaluate our model with the Perplexity score (PPL) and zero-shot performance with nine different downstream tasks: LAMBADA [58], BLiMP [74], Children's Book Test [24], HellaSwag [81], PIQA[3], ARC-Challenge [11], RACE [35], SIQA [63] and CommonSenseQA [68]. For LAMBADA, we use the detokenized version from OpenAI, and we evaluate the top-1 accuracy of the last word (it can span multiple tokens; here we use greedy decoding). For CBT, BLiMP, and

RACE, we measure the accuracy of each task and report the average accuracy of the tasks.

Compute Resource. All models are trained and evaluated on a single node equipped with 4 NVIDIA A100 80GB CoWoS HBM2e PCIe 4.0 employing data-parallelism.

I.2 Vision Language Modeling

I.2.1 Datasets

LLaVA-558K. The LLaVA 558K [44] dataset is a curated subset of 558,000 image-text pairs derived from the LAION/CC/SBU dataset. It is designed for the pretraining stage of visual instruction tuning, facilitating the alignment between visual and language modalities. This dataset includes BLIP-generated captions and synthetic multimodal conversations, serving as a foundational resource for training models like LLaVA towards enhanced vision-language capabilities.
ALLaVA. ALLaVA [6] is a large-scale synthetic dataset comprising approximately 1.3 million samples, generated using GPT-4V. It includes fine-grained image annotations and complex reasoning visual question-answering pairs. The dataset aims to bridge the performance gap between traditional large vision-language models and more resource-efficient lite versions by providing high-quality training data for visual instruction tuning.
LLaVA-665K. The LLaVA-665K [42] dataset is an expanded and refined version of the original LLaVA instruction tuning dataset, containing 665,000 multimodal instruction-following samples. It integrates diverse sources such as VQAv2, GQA, OCR-VQA, and RefCOCO, among others, to enhance the model's performance across various vision-language tasks. This comprehensive dataset supports improved visual instruction tuning for models like LLaVA-1.5 [42].
AI2D. The AI2D (AI2 Diagrams) [31] dataset comprises over 5000 grade school science diagrams, annotated with more than 150,000 rich annotations and over 15000 corresponding multiple-choice questions. It serves as a resource for evaluating models' abilities in diagram understanding and visual reasoning within educational contexts.
MMStar. MMStar [7] is a meticulously curated benchmark designed to evaluate large vision-language models (LVLMs) on vision-indispensable tasks. It includes 1,500 samples across six core capabilities and 18 detailed axes, ensuring each sample necessitates visual understanding and minimizes data leakage.
POPE. The POPE (Polling-based Object Probing Evaluation) [40] dataset is developed to assess object hallucination in LVLMs. It provides a systematic approach to evaluate the consistency of object descriptions generated by models, highlighting tendencies to generate objects not present in the input images.
ScienceQA. ScienceQA [47] is a large-scale multimodal dataset featuring science questions enriched with lectures and explanations. It spans diverse subjects, including natural science, language science, and social science, aiming to evaluate models' abilities in multimodal reasoning and explanatory question answering.
TextVQA. The TextVQA [65] dataset focuses on visual question answering tasks that require reading and reasoning about text within images. It contains 45,336 questions over 28,408 images, challenging models to integrate textual and visual information effectively.
GQA. GQA (Graph Question Answering) [27] is a large-scale dataset designed for real-world visual reasoning and compositional question answering. It includes 22 million questions based on 113,000 images, each accompanied by scene graphs detailing objects, attributes, and relationships, facilitating

structured reasoning evaluations.

MME-RealWorld-Lite. MME-RealWorld-Lite [83] is a streamlined version of the MME-RealWorld benchmark, offering a subset of 50 samples per task to accelerate inference. It maintains the benchmark's focus on evaluating multimodal models in real-world scenarios with high-resolution images and complex tasks.

MMMU Pro. MMMU Pro [78] is an enhanced benchmark for assessing multimodal models' understanding and reasoning across multiple disciplines. It filters out questions answerable by text-only models, augments candidate options, and introduces vision-only input settings, thereby rigorously evaluating models' true multimodal capabilities.

OCRBench. OCRBench [45] is a comprehensive evaluation benchmark for optical character recognition (OCR) capabilities in large multimodal models. It encompasses 29 datasets covering tasks like text recognition, scene text-centric VQA, document-oriented VQA, key information extraction, and handwritten mathematical expression recognition, providing a thorough assessment of OCR performance.

I.2.2 Model Settings, Training Settings and Evaluation

Model Settings. We embrace the vision-language pre-training task [42], a challenging problem setting that enables effective model training with relatively limited data. We adopt the experiment setting in LIBMoE [57] with LLaVA architecture [44], which includes three modules: pre-trained Large Language Model, pre-trained visual encoder, and randomly initialized MLP connector. We employ the pre-trained SigLIP (Patch14-224) [82] as the vision encoder, pre-trained Phi-3.5-mini-instruct [1] as the LLM, and a randomly initialized MLP connector. In the Visual Instruction Tuning (VIT) stage, we adopt a sparse upcycling approach [32] and upcycle only the MLP Connector into 8 experts, employing a top-4 expert routing strategy, while the DeepSeek variants adopt a top-3 expert routing scheme with an additional shared expert. Thus, our model has approximately 4.4B parameters.

Training Settings. We follow LIBMoE [57] for the training settings. Specifically, our training recipe with three stages of training: pre-training, pre-finetuning, and Visual Instruction Tuning (VIT). In the first stage, we only pretrain the MLP connector for better alignment using LLaVA 558K dataset [44]. During the second pre-finetuning stage, we train all parameters using high-quality caption data with the ALLaVA [6] dataset with 708K samples, aiming to warm up the entire model. In the third stage, we upcycle the MLP Connector to MoE block and trained on visual instruction tuning data (a subset of LLaVA-665K [42] with 332K samples). The learning rate is set to $1e-3$ for pre-training the MLP connector and reduced to $2e-6$ for pre-finetuning and $4e-6$ for the final stage. All models are trained in PyTorch using a batch size of 4 and AdamW optimizer [46] with its default hyperparameters. We use Zero Redundancy Optimizer (ZeRO) [61] for memory optimization with Zero2 for the first stage and Zero3 for both pre-finetuning and VIT stages.

Evaluation. Our model is evaluated under the zero-shot setting across a diverse set of benchmarks encompassing various vision-language capabilities, such as perception, reasoning, OCR, instruction following, etc. The benchmarks considered include AI2D [31], MMStar [7], POPE [40], ScienceQA [47], TextVQA [65], GQA [27], MME-RealWorld-Lite [83], MMMU Pro [78], OCRBench [45].

Compute Resource. All models are trained and evaluated on a single node equipped with 4 NVIDIA A100 80GB CoWoS HBM2e PCIe 4.0 employing data-parallelism.

I.3 Training Time and Resource Allocation

Table 5 summarizes the training time and resource utilization across all experimental settings.

Table 5: Training Time and GPU Resource Allocation Across All Experimental Settings.

Model			**Training Time** (hours)	**Resourse**
Vision Language Modeling	Pre-Training		5.5	4xA100
	Pre-FineTuning		18	4xA100
	Visual Instruction Tuning	SMoE	10	4xA100
		SMoE Sigmoid Gating	10	4xA100
		DeepSeek-V2	10.5	4xA100
		DeepSeek-V3	10.5	4xA100
Language Modeling	158M parametes	SMoE	9.5	4xA100
		SMoE Sigmoid Gating	10	4xA100
		DeepSeek-V2	10.5	4xA100
		DeepSeek-V3	10.5	4xA100
	679M parametes	SMoE	65	4xA100
		SMoE Sigmoid Gating	65	4xA100
		DeepSeek-V2	71	4xA100
		DeepSeek-V3	71.5	4xA100

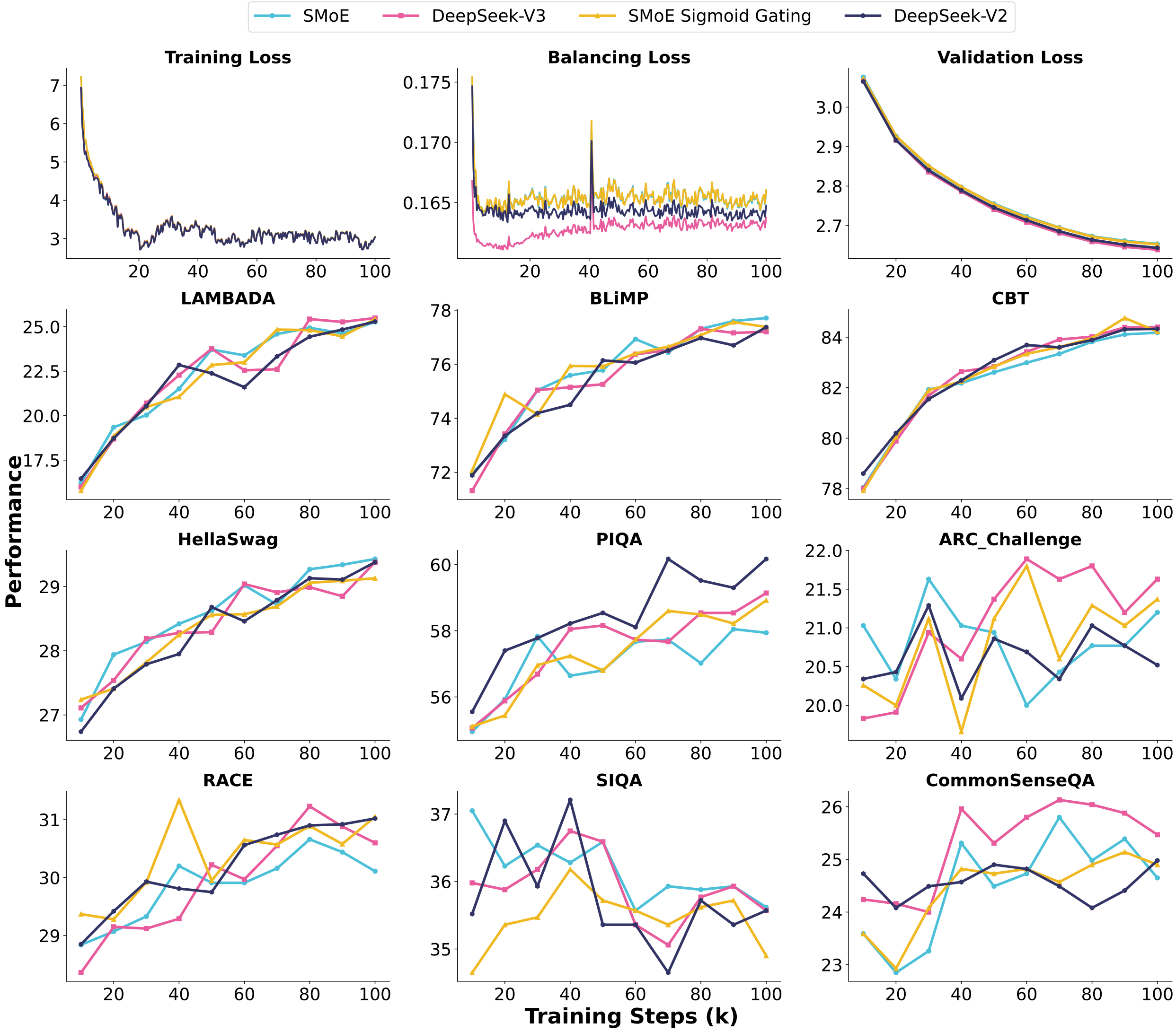


Figure 14: Benchmark curves during training in language modeling tasks for models with 158M parameters.

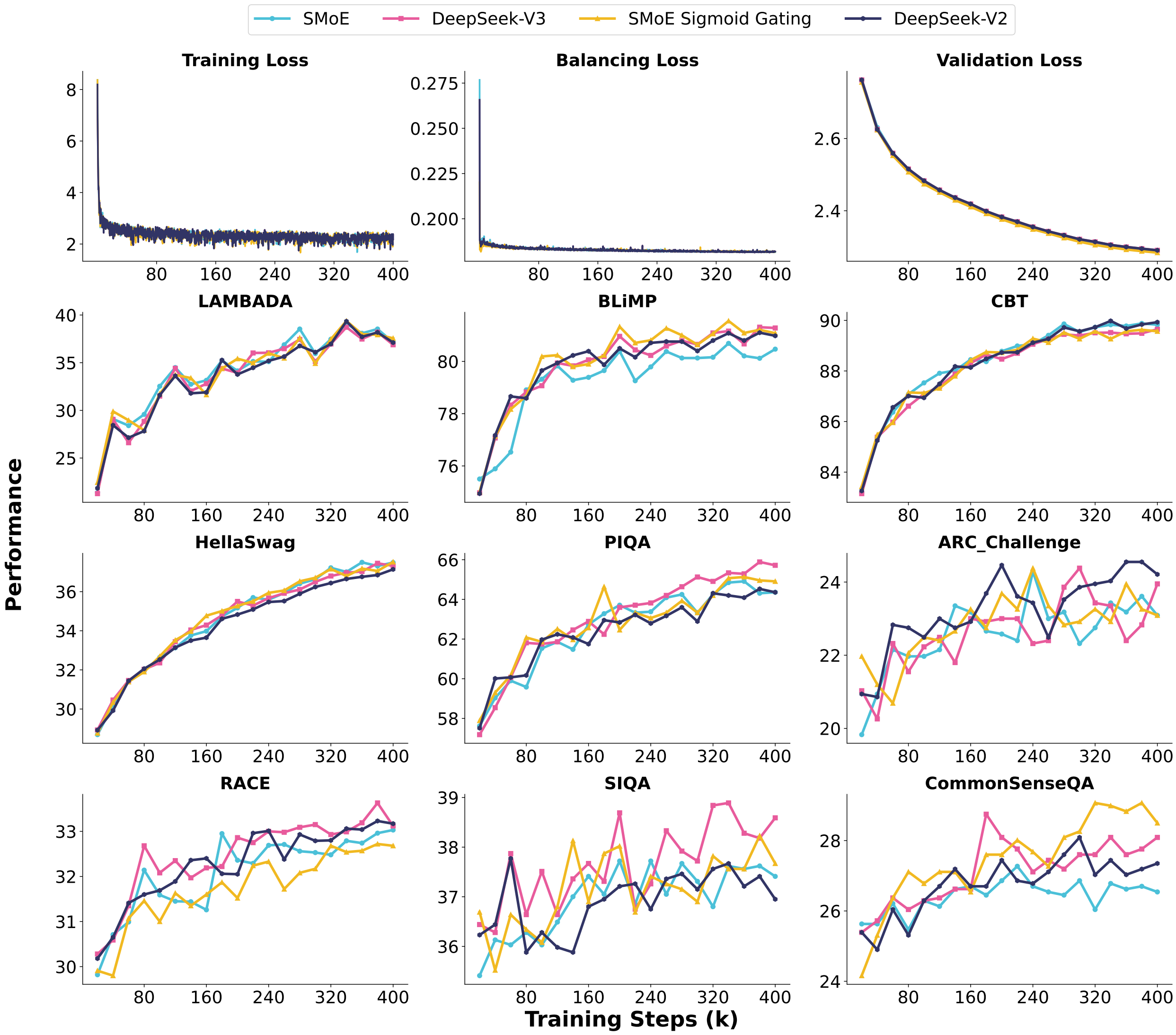


Figure 15: Benchmark curves during training in language modeling tasks for models with 679M parameters.

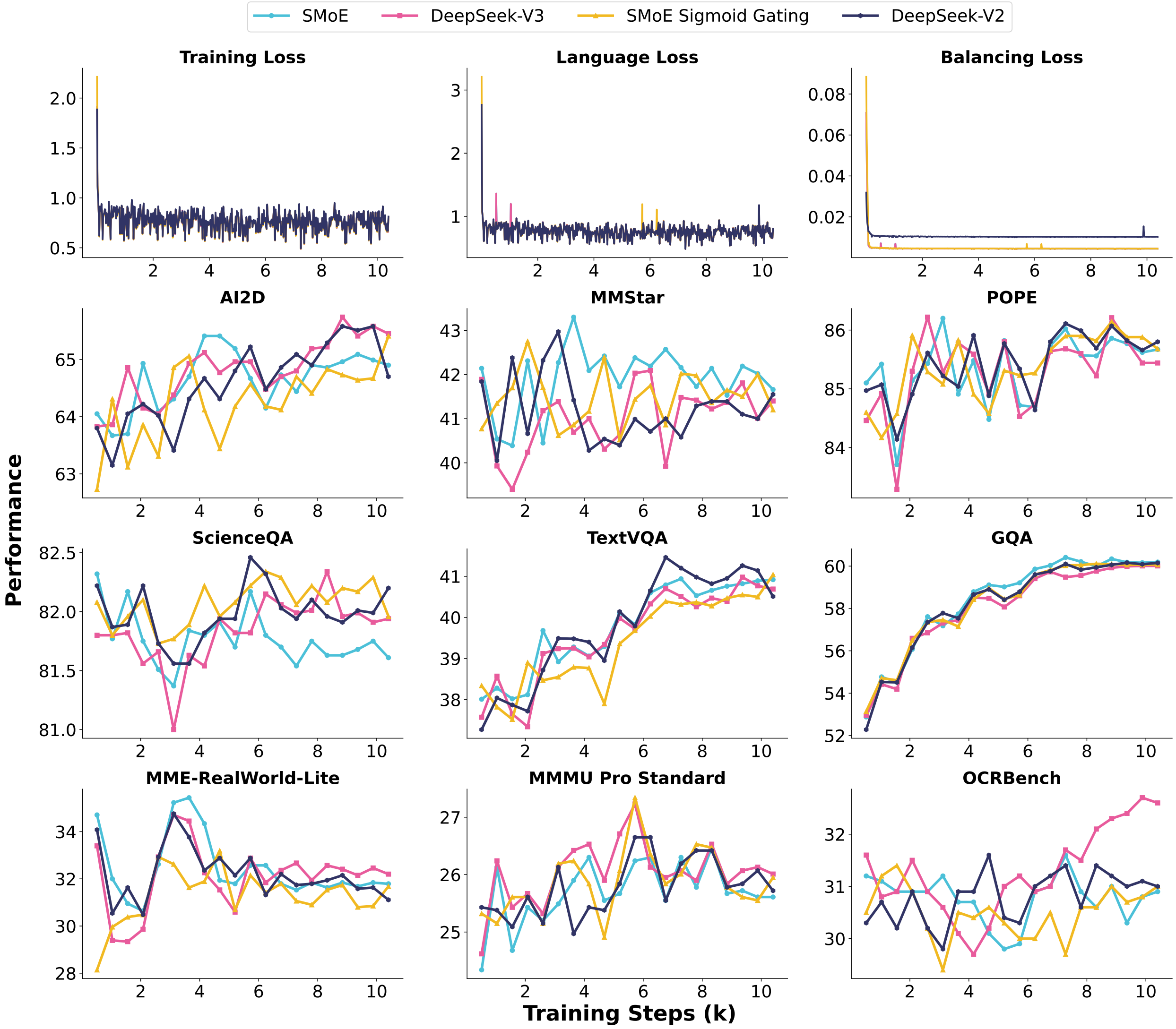


Figure 16: Benchmark curves during training in vision-language modeling tasks.

References

[1] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. (Cited on pages 9 and 89.)

[2] P. Akbarian, H. Nguyen, X. Han, and N. Ho. Quadratic gating functions in mixture of experts: A statistical insight. *arXiv preprint arXiv:2410.11222*, 2024. (Cited on page 19.)

[3] Y. Bisk, R. Zellers, J. Gao, Y. Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, 2020. (Cited on pages 8, 86, and 87.)

[4] J. S. O. Ceron, G. Sokar, T. Willi, C. Lyle, J. Farebrother, J. N. Foerster, G. K. Dziugaite, D. Precup, and P. S. Castro. Mixtures of experts unlock parameter scaling for deep RL. In *Forty-first International Conference on Machine Learning*, 2024. (Cited on page 1.)

[5] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Time series modeling by a regression approach based on a latent process. *Neural Networks*, 22(5-6):593–602, 2009. (Cited on page 76.)

[6] G. H. Chen, S. Chen, R. Zhang, J. Chen, X. Wu, Z. Zhang, Z. Chen, J. Li, X. Wan, and B. Wang. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*, 2024. (Cited on pages 88 and 89.)

[7] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. (Cited on pages 9, 88, and 89.)

[8] Z. Chen, Y. Deng, Y. Wu, Q. Gu, and Y. Li. Towards understanding the mixture-of-experts layer in deep learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23049–23062. Curran Associates, Inc., 2022. (Cited on pages 1 and 19.)

[9] Z. Chi, L. Dong, S. Huang, D. Dai, S. Ma, B. Patra, S. Singhal, P. Bajaj, X. Song, X.-L. Mao, H. Huang, and F. Wei. On the representation collapse of sparse mixture of experts. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. (Cited on page 19.)

[10] Y. Chow, A. Tulepbergenov, O. Nachum, D. Gupta, M. Ryu, M. Ghavamzadeh, and C. Boutilier. A Mixture-of-Expert Approach to RL-based Dialogue Management. In *The Eleventh International Conference on Learning Representations*, 2023. (Cited on page 1.)

[11] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. (Cited on pages 8, 86, and 87.)

[12] D. Dai, C. Deng, C. Zhao, R. X. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu, Z. Xie, Y. K. Li, P. Huang, F. Luo, C. Ruan, Z. Sui, and W. Liang. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.04088*, 2024. (Cited on pages 1, 2, and 87.)

[13] D. Dai, L. Dong, S. Ma, B. Zheng, Z. Sui, B. Chang, and F. Wei. Stablemoe: Stable routing strategy for mixture of experts. *arXiv preprint arXiv:2204.08396*, 2022. (Cited on page 11.)

[14] DeepSeek-AI et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024. (Cited on page 2.)

[15] DeepSeek-AI et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. (Cited on page 2.)

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977. (Cited on page 76.)

[17] N. T. Diep, H. Nguyen, C. Nguyen, M. Le, D. M. H. Nguyen, D. Sonntag, M. Niepert, and N. Ho. On zero-initialized attention: Optimal prompt and gating factor estimation. *arXiv preprint arXiv:2502.03029*, 2025. (Cited on page 19.)

[18] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. Yu, O. Firat, B. Zoph, L. Fedus, M. Bosma, Z. Zhou, T. Wang, E. Wang, K. Webster, M. Pellat, K. Robinson, K. Meier-Hellstern, T. Duke, L. Dixon, K. Zhang, Q. Le, Y. Wu, Z. Chen, and C. Cui. Glam: Efficient scaling of language models with mixture-of-experts. In *ICML*, 2022. (Cited on page 1.)

[19] S. Faria and G. Soromenho. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, 2010. (Cited on page 1.)

[20] W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. (Cited on pages 1 and 8.)

[21] W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23:1–39, 2022. (Cited on page 87.)

[22] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. (Cited on page 2.)

[23] X. Han, H. Nguyen, C. Harris, N. Ho, and S. Saria. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. In *Advances in Neural Information Processing Systems*, 2024. (Cited on page 1.)

[24] F. Hill, A. Bordes, S. Chopra, and J. Weston. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*, 2015. (Cited on pages 8, 86, and 87.)

[25] N. Ho and X. Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6):2726 – 2755, 2016. Publisher: Institute of Mathematical Statistics and Bernoulli Society. (Cited on page 18.)

[26] N. Ho, C.-Y. Yang, and M. I. Jordan. Convergence rates for Gaussian mixtures of experts. *Journal of Machine Learning Research*, 23(323):1–81, 2022. (Cited on page 19.)

[27] D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. (Cited on pages 9, 88, and 89.)

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

[29] R. K. Jain, D.-M. W. Chiu, W. R. Hawe, et al. A quantitative measure of fairness and discrimination. *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, 21(1), 1984. (Cited on page 84.)

[30] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mixtral of experts. *arxiv preprint arxiv 2401.04088*, 2024. (Cited on page 1.)

[31] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016. (Cited on pages 9, 88, and 89.)

[32] A. Komatsuzaki, J. Puigcerver, J. Lee-Thorp, C. R. Ruiz, B. Mustafa, J. Ainslie, Y. Tay, M. Dehghani, and N. Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints. *arXiv preprint arXiv:2212.05055*, 2022. (Cited on pages 9 and 89.)

[33] T. Kudo and J. Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018. (Cited on page 87.)

[34] J. Kwon and C. Caramanis. EM Converges for a Mixture of Many Linear Regressions. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1727–1736. PMLR, Aug. 2020. (Cited on page 1.)

[35] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. (Cited on pages 8, 86, and 87.)

[36] M. Le, C. Nguyen, H. Nguyen, Q. Tran, T. Le, and N. Ho. Revisiting prefix-tuning: Statistical benefits of reparameterization among prompts. In *The Thirteenth International Conference on Learning Representations*, 2025. (Cited on page 19.)

[37] M. Le, A. N. The, H. Nguyen, T. T. N. Vu, H. T. Pham, L. N. Van, and N. Ho. Mixture of experts meets prompt-based continual learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. (Cited on page 19.)

[38] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In *International Conference on Learning Representations*, 2021. (Cited on page 1.)

[39] H. Li, S. Lin, L. Duan, Y. Liang, and N. Shroff. Theory on mixture-of-experts in continual learning. In *The Thirteenth International Conference on Learning Representations*, 2025. (Cited on page 19.)

[40] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. (Cited on pages 9, 88, and 89.)

[41] H. Liang, Z. Fan, R. Sarkar, Z. Jiang, T. Chen, K. Zou, Y. Cheng, C. Hao, and Z. Wang. M^{3}ViT: Mixture-of-Experts Vision Transformer for Efficient Multi-task Learning with Model-Accelerator Co-design. In *NeurIPS*, 2022. (Cited on page 1.)

[42] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. (Cited on pages 9, 88, and 89.)

[43] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *NeurIPS*, 2023. (Cited on page 8.)

[44] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. (Cited on pages 9, 88, and 89.)

[45] Y. Liu, Z. Li, M. Huang, B. Yang, W. Yu, C. Li, X.-C. Yin, C.-L. Liu, L. Jin, and X. Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024. (Cited on pages 9 and 89.)

[46] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. (Cited on pages 87 and 89.)

[47] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. (Cited on pages 9, 88, and 89.)

[48] J. Ludziejewski, J. Krajewski, K. Adamczewski, M. Pióro, M. Krutul, S. Antoniak, K. Ciebiera, K. Król, T. Odrzygóźdź, P. Sankowski, M. Cygan, and S. Jaszczur. Scaling laws for fine-grained mixture of experts. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024. (Cited on page 12.)

[49] T. Manole and N. Ho. Refined convergence rates for maximum likelihood estimation under finite mixture models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14979–15006. PMLR, 17–23 Jul 2022. (Cited on page 5.)

[50] E. F. Mendes and W. Jiang. Convergence rates for mixture-of-experts. *arXiv preprint arxiv 1110.2058*, 2011. (Cited on page 19.)

[51] N. Muennighoff, L. Soldaini, D. Groeneveld, K. Lo, J. Morrison, S. Min, W. Shi, P. Walsh, O. Tafjord, N. Lambert, et al. Olmoe: Open mixture-of-experts language models. *arXiv preprint arXiv:2409.02060*, 2024. (Cited on page 10.)

[52] H. Nguyen, P. Akbarian, T. Nguyen, and N. Ho. A general theory for softmax gating multinomial logistic mixture of experts. In *Proceedings of the ICML*, 2024. (Cited on page 1.)

[53] H. Nguyen, P. Akbarian, T. Pham, T. Nguyen, S. Zhang, and N. Ho. Statistical advantages of perturbing cosine router in mixture of experts. In *International Conference on Learning Representations*, 2025. (Cited on page 19.)

[54] H. Nguyen, P. Akbarian, F. Yan, and N. Ho. Statistical perspective of top-k sparse softmax gating mixture of experts. In *International Conference on Learning Representations*, 2024. (Cited on pages 19 and 69.)

[55] H. Nguyen, N. Ho, and A. Rinaldo. Convergence rates for softmax gating mixture of experts. *arXiv preprint arXiv:2503.03213*, 2025. (Cited on page 19.)

[56] H. Nguyen, T. Nguyen, and N. Ho. Demystifying softmax gating function in Gaussian mixture of experts. In *Advances in Neural Information Processing Systems*, 2023. (Cited on pages 5, 8, 18, and 19.)

[57] N. V. Nguyen, T. T. Doan, L. Tran, V. Nguyen, and Q. Pham. Libmoe: A library for comprehensive benchmarking mixture of experts in large language models. *arXiv preprint arXiv:2411.00918*, 2024. (Cited on pages 8, 9, and 89.)

[58] D. Paperno, G. Kruszewski, A. Lazaridou, Q. N. Pham, R. Bernardi, S. Pezzelle, M. Baroni, G. Boleda, and R. Fernández. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*, 2016. (Cited on pages 8, 86, and 87.)

[59] Q. Pham, G. Do, H. Nguyen, T. Nguyen, C. Liu, M. Sartipi, B. T. Nguyen, S. Ramasamy, X. Li, S. Hoi, and N. Ho. Competesmoe – effective training of sparse mixture of experts via competition. *arXiv preprint arXiv:2402.02526*, 2024. (Cited on page 19.)

[60] Qwen et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025. (Cited on page 2.)

[61] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020. (Cited on page 89.)

[62] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. S. Pint, D. Keysers, and N. Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 34, pages 8583–8595. Curran Associates, Inc., 2021. (Cited on page 1.)

[63] M. Sap, H. Rashkin, D. Chen, R. LeBras, and Y. Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019. (Cited on pages 8, 86, and 87.)

[64] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *In International Conference on Learning Representations*, 2017. (Cited on pages 1 and 2.)

[65] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. (Cited on pages 9, 88, and 89.)

[66] D. Soboleva, F. Al-Khateeb, R. Myers, J. R. Steeves, J. Hestness, and N. Dey. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama, 2023. (Cited on pages 8, 9, 86, and 87.)

[67] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. (Cited on page 87.)

[68] A. Talmor, J. Herzig, N. Lourie, and J. Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018. (Cited on pages 8, 86, and 87.)

[69] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. (Cited on page 2.)

[70] H. Teicher. Identifiability of finite mixtures. *Ann. Math. Statist.*, 32:1265–1269, 1963. (Cited on page 68.)

[71] T. Truong, C. Nguyen, H. Nguyen, M. Le, T. Le, and N. Ho. Replora: Reparameterizing low-rank adaptation via the perspective of mixture of experts. *arXiv preprint arXiv:2502.03044*, 2025. (Cited on page 19.)

[72] S. van de Geer. *Empirical processes in M-estimation.* Cambridge University Press, 2000. (Cited on pages 62 and 63.)

[73] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. (Cited on page 19.)

[74] A. Warstadt, A. Parrish, H. Liu, A. Mohananey, W. Peng, S.-F. Wang, and S. R. Bowman. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. (Cited on pages 8, 86, and 87.)

[75] M. Weber, D. Fu, Q. Anthony, Y. Oren, S. Adams, A. Alexandrov, X. Lyu, H. Nguyen, X. Yao, V. Adams, et al. Redpajama: an open dataset for training large language models. *Advances in neural information processing systems*, 37:116462–116492, 2024. (Cited on page 86.)

[76] F. Xue, Z. Zheng, Y. Fu, J. Ni, Z. Zheng, W. Zhou, and Y. You. Openmoe: An early effort on open mixture-of-experts language models. *arXiv preprint arXiv:2402.01739*, 2024. (Cited on page 10.)

[77] F. Yan, H. Nguyen, P. Akbarian, N. Ho, and A. Rinaldo. Sigmoid self-attention is better than softmax self-attention: A mixture-of-experts perspective. *arXiv preprint arXiv:2502.00281*, 2025. (Cited on page 19.)

[78] X. Yue, T. Zheng, Y. Ni, Y. Wang, K. Zhang, S. Tong, Y. Sun, B. Yu, G. Zhang, H. Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. (Cited on pages 9 and 89.)

[79] S. Yun, I. Choi, J. Peng, Y. Wu, J. Bao, Q. Zhang, J. Xin, Q. Long, and T. Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. (Cited on page 1.)

[80] A. Zeevi, R. Meir, and V. Maiorov. Error bounds for functional approximation and estimation using mixtures of experts. *IEEE Transactions on Information Theory*, 44(3):1010–1025, 1998. (Cited on page 19.)

[81] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019. (Cited on pages 8, 86, and 87.)

[82] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. (Cited on pages 9 and 89.)

[83] Y.-F. Zhang, H. Zhang, H. Tian, C. Fu, S. Zhang, J. Wu, F. Li, K. Wang, Q. Wen, Z. Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024. (Cited on pages 9 and 89.)