Convergence Rates for Softmax Gating Mixture of Experts

Huy Nguyen   Nhat Ho[⋆]   Alessandro Rinaldo[⋆]

The University of Texas at Austin[†]

March 6, 2025

Abstract

Mixture of experts (MoE) has recently emerged as an effective framework to advance the efficiency and scalability of machine learning models by softly dividing complex tasks among multiple specialized sub-models termed experts. Central to the success of MoE is an adaptive softmax gating mechanism which takes responsibility for determining the relevance of each expert to a given input and then dynamically assigning experts their respective weights. Despite its widespread use in practice, a comprehensive study on the effects of the softmax gating on the MoE has been lacking in the literature. To bridge this gap in this paper, we perform a convergence analysis of parameter estimation and expert estimation under the MoE equipped with the standard softmax gating or its variants, including a dense-to-sparse gating and a hierarchical softmax gating, respectively. Furthermore, our theories also provide useful insights into the design of sample-efficient expert structures. In particular, we demonstrate that it requires polynomially many data points to estimate experts satisfying our proposed *strong identifiability* condition, namely a commonly used two-layer feed-forward network. In stark contrast, estimating linear experts, which violate the strong identifiability condition, necessitates exponentially many data points as a result of intrinsic parameter interactions expressed in the language of partial differential equations. All the theoretical results are substantiated with a rigorous guarantee.

1   Introduction

Introduced by Jacob et al. [18], mixture of experts (MoE) has been known as a powerful statistical machine learning framework that generalizes the concept of conventional mixture models [25] based on the principle of divide and conquer. More specifically, it incorporates the power of multiple sub-models referred to as experts through an adaptive gating mechanism. Here, each expert can be a classifier [4, 31], a regression function [10] or a feed-forward network (FFN) that specializes in some specific tasks [42, 9]. Meanwhile, a gating function is formulated as an FFN followed by a softmax normalization, which we refer to as the softmax gating function throughout the paper. Its responsibility is to dynamically assign a corresponding weight for each individual expert in a way that experts which are more relevant to the input will have larger weights than others. Therefore, the weight set of the MoE model varies with the input value rather than remains constant as that of the traditional mixture model (see Figure 1a). This feature accounts for the flexibility and adaptivity of the MoE, leading to a series of MoE applications in classification [46], multi-task learning [14, 12], speech recognition [39, 47], and bioinformatics [40], etc.

In order to enhance the computational efficiency of the MoE, Shazeer et al. [42] has recently developed a sparse variant of the softmax gating function known as the Top-K sparse softmax gating

---

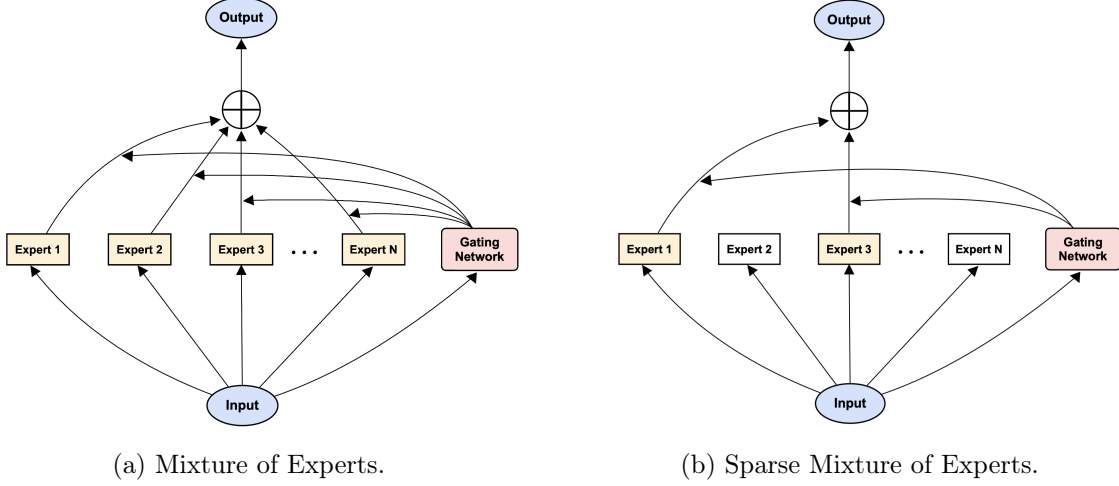(a) Mixture of Experts.   (b) Sparse Mixture of Experts.

Figure 1: Illustration of mixture of experts.

function. In particular, the gating network is trained to route each input to only the K most relevant expert networks rather than all of them (see Figure 1b). This mechanism can be viewed as a form of conditional computation [1, 6], and it allows the sparse MoE to reduce the computational overhead significantly while maintaining or even improving the model performance. A typical application of the sparse MoE is in the Transformer model [45], which has been the state-of-the-art architecture for many natural language processing tasks. More concretely, the FFN in a transformer layer is often replaced with a sparse mixture of smaller FFNs whose total number of parameters matches that of the original FFN layer, which might contain up to billions of parameters [20, 9]. Since only a few FFNs are activated per input, it is obvious that the computational cost decreases substantially. Moreover, the usage of multiple FFNs might help learn domain-specific or task-specific tasks better than using a single FFN. For example, when the dataset consists of diverse data modalities such as texts, images, time series, etc., then each FFN can be trained to specialize in processing a specific data modality to make the most out of the dataset [13, 49], thereby enhancing the model performance. As a result, the sparse MoE has been widely utilized in several large-scale models, namely large language models [26, 20, 11, 43], computer vision [41, 24, 3], domain generalization [23, 32], and reinforcement learning [2, 7].

Despite its widespread use, there are two main disadvantages of the sparse gating function proposed in [42]. First, fixing the number of activated experts per input not only hinders the expert exploration but also makes the gating function discontinuous, probably causing some challenges in terms of optimization. Second, since the gating network lacks the experience of expert selection in early training, it may also exhibit undesirable instability. For those reasons, Nie et al. [37] propose a dense-to-sparse gating function which involves a temperature parameter in the dense softmax gating to control the weight distribution. In particular, when the temperature parameter tends to infinity, the mixture weights are uniformly distributed, meaning that all the experts are activated and assigned identical weights. On the other hand, when the temperature parameter approaches zero, the weight distribution converges to a one-hot vector, indicating that only one expert is activated per input. This strategy helps smooth the expert selection process as well as dynamically adjust the

sparsity level of the MoE model. Therefore, the dense-to-sparse gating has been leveraged in several works on MoE, namely [5, 30, 8].

In addition to the standard MoE and its sparse versions that we have discussed so far, we would also like to bring attention to a hierarchical MoE (HMoE) model introduced by Jordan et al. [22] as a tree-structured architecture for supervised learning. The HMoE encompasses multiple layers of gating networks to organize the experts in a multi-level hierarchy as illustrated in Figure 2. Such hierarchical design has been demonstrated to facilitate the expert specialization and enhance the model generalization, particularly in the scenarios involving diverse data distributions [33, 17] or complex decision-making tasks [19, 29].
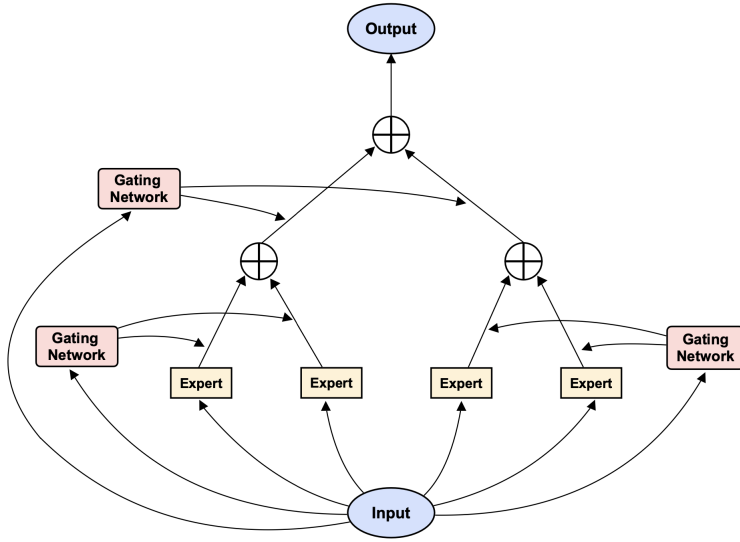

Figure 2: Illustration of two-level hierarchical mixture of experts.

Related works. Given the success of applying the MoE and HMoE models in practice, it is natural to ask about the theoretical attempts to understand those models. First, Zeevi et al. [50] established the error of approximating a target function in the Sobolev class using a mixture of generalized linear experts under the L^p norm. After that, Jiang et al. [21] proceeded to studied the approximation power of the HMoE models where exponential family regression models with generalized linear mean functions were aggregated. They demonstrated that such models were able to approximate one-parameter exponential family densities with arbitrary smooth mean functions in a transformed Sobolev space when the number of experts increased. Next, Mendes et al. [28] investigated the convergence rate of the maximum likelihood estimator (MLE) under the MoE model where each expert was formulated as a polynomial regression model. From the theoretical results, they provided some implications for the optimal number of experts and the complexity of the expert models. The convergence analysis of the MLE was then continued in [16] and [35] but under the Gaussian MoE models. Those works pointed out that the parameter estimation rates were negatively affected by some intrinsic interaction among the gating and expert parameters expressed in the language of partial differential equations (PDEs). Unlike previous works where the MoE was associated with a probability distribution, Nguyen et al. [34] considered a regression framework where the regression

3

function took the form of a deterministic softmax gating MoE, and studied the problem of expert estimation using the least square method. In particular, they derived a so-called strong identifiability condition to characterize the types of experts that admitted faster convergence rates than others, namely those formulated as two-layer FFNs with GELU or sigmoid activation function. In this work, we will first revisit those results and then extend them to the settings where the regression function admits the form of a dense-to-sparse gating MoE and a hierarchical MoE, respectively, which have remained elusive in the literature. For the sake of presentation, let us introduce the formal problem statement for the setting of the softmax gating MoE below and then defer those for their variants to Section 3 and Section 4.

Problem setting. We assume that an i.i.d. sample of size n: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ in $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$ is generated according to the model

$$Y_i = f_{G_*}(X_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{1}$$

Above, we assume that $X_1, X_2, \ldots, X_n$ are i.i.d. samples from some probability distribution μ. Meanwhile, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent noise random variables such that their conditional probability distributions given the input are Gaussian, that is, $\varepsilon_i | X_i \sim \mathcal{N}(0, \nu)$, for all $1 \leq i \leq n$. We note in passing that the Gaussian assumption is simply for the ease of proof arguments. Next, the regression function $f_{G_*}(\cdot)$ is assumed to take the form of the softmax gating mixture of k_* experts, namely

$$f_{G_*}(x) := \sum_{i=1}^{k_*} \text{Softmax}((\omega_i^*)^\top x + \beta_i^*) \cdot \mathcal{E}(x, \eta_i^*), \tag{2}$$

where $(\beta_i^*, \omega_i^*, \eta_i^*)_{i=1}^{k_*}$ are true yet unknown parameters belonging to the parameter space $\Theta \subseteq \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^q$ and $G_* := \sum_{i=1}^{k_*} \exp(\beta_i^*) \delta_{(\omega_i^*, \eta_i^*)}$ denotes the associated *mixing measure*, that is, a weighted sum of Dirac delta measures. The corresponding regression functions to the settings of the dense-to-sparse gating MoE and the HMoE are given in equations (12) and (20), respectively. Additionally, the terms $\mathcal{E}(x, \eta_i^*)$, for $1 \leq i \leq k_*$, are referred to as experts.

It is worth noting that expert specialization is an essential problem in the literature of MoE models [9, 38]. Therefore, our main objective is to learn how fast we can estimate the parametric ground-truth experts $\mathcal{E}(x, \eta_i^*)$, which can be obtained by determining the convergence rate of parameter estimation. Since the number of experts k_* is unknown in practice, we over-specify the model in equation (2) by a mixture of k experts, where $k > k_*$ is a given threshold. Then, we can estimate the unknown parameters $(\beta_i^*, \omega_i^*, \eta_i^*)_{i=1}^{k_*}$ via estimating the mixing measure G_* using the least square method as follows:

$$\widehat{G}_n := \underset{G \in \mathcal{G}_k(\Theta)}{\arg\min} \sum_{i=1}^n \left(Y_i - f_G(X_i)\right)^2, \tag{3}$$

where $\mathcal{G}_k(\Theta) := \{G = \sum_{i=1}^{k'} \exp(\beta_i) \delta_{(\omega_i, \eta_i)} : 1 \leq k' \leq k, \ (\beta_i, \omega_i, \eta_i) \in \Theta\}$ stands for the set of all mixing measures with no more than k atoms.

Contributions. In this paper, we analyze the convergence behavior of parameter estimation and expert estimation under the (hierarchical) MoE with the softmax gating and the dense-to-sparse

4

Table 1: Summary of estimation rates for strongly identifiable experts and linear experts.

	Strongly Identifiable Experts	Linear Experts	Theorems
Softmax gating MoE	$\mathcal{O}_P([\log(n)/n]^{1/4})$	$\mathcal{O}_P(1/\log^\lambda(n))$	Thm.1, Thm.2
Dense-to-sparse gating MoE with linear router	$\mathcal{O}_P(1/\log^\lambda(n))$	$\mathcal{O}_P(1/\log^\lambda(n))$	Thm.3
Dense-to-sparse gating MoE with general router	$\mathcal{O}_P([\log(n)/n]^{1/4})$	$\mathcal{O}_P(1/\log^\lambda(n))$	Thm.4
Hierarchical MoE	$\mathcal{O}_P([\log(n)/n]^{1/4})$	$\mathcal{O}_P(1/\log^\lambda(n))$	Thm.5, Thm.6

gating. Our ultimate goal is to find the optimal expert structure for each MoE model to provide a useful guide on the model design for its practical applications. Our contributions are three-fold and can be summarized as follows (see also Table 1):

1. Softmax gating MoE: We re-state the strong identifiability condition in [34] for characterizing the structure of experts admitting faster convergence rates than others. The main intuition behind that condition is to eliminate interactions among parameters through some PDEs. Then, we demonstrate that the rates for estimating strongly identifiable experts, including those formulated as FFNs with GELU or sigmoid activation, are parametric on the sample size. On the other hand, linear experts, which fail to satisfy the strong identifiability condition, are shown to have slower estimation rates than any polynomial rates.

2. Dense-to-sparse gating MoE: Recall that we involve the temperature parameter in the dense-to-sparse gating function to control the sparsity of the MoE model. However, our theory reveals that the temperature has an undesirable interaction with the gating parameters expressed via a PDE, leading to a substantial deceleration in the expert convergence regardless of the expert structure. In response to this issue, we generalize the linear router inside the dense-to-sparse gating function to a general router and then establish a condition to survey which combinations of the router and the expert structure will accelerate the expert convergence.

3. Hierarchical MoE: For the softmax gating hierarchical MoE, we discover that experts satisfying the aforementioned strong identifiability condition still enjoy a faster estimation rate than others. At the same time, we show that the rates for estimating linear experts are slower than any polynomial rates due to the interaction between gating parameters and expert parameters.

Organization. The paper proceeds as follows. In Section 2, we recall the convergence analysis of the softmax gating MoE equipped with strongly identifiable experts and linear experts. Next, we study the effects of involving the temperature parameter in the softmax gating function on the convergence of expert estimation in Section 3. Subsequently, we generalize the results to the setting of the HMoE model in Section 4 before providing an in-depth discussion on the theoretical results in Section 5. Full proofs and additional results can be found in the Appendices.

Notations. For any $n \in \mathbb{N}$, we denote $[n]$ as the set $= \{1, 2, \ldots, n\}$. Additionally, for any set S, we refer to $|S|$ as its cardinality. Next, for any vectors $v := (v_1, v_2, \ldots, v_d) \in \mathbb{R}^d$ and $\alpha := (\alpha_1, \alpha_2, \ldots, \alpha_d) \in \mathbb{N}^d$, we let $v^\alpha = v_1^{\alpha_1} v_2^{\alpha_2} \ldots v_d^{\alpha_d}$, $|v| := v_1 + v_2 + \ldots + v_d$ and $\alpha! := \alpha_1! \alpha_2! \ldots \alpha_d!$,

while $\|v\|$ stands for its 2-norm value. Lastly, 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 $n \in \mathbb{N}$, where $C > 0$ is some universal constant. The notation $a_n = \mathcal{O}_P(b_n)$ indicates that a_n/b_n is stochastically bounded.

2   Softmax Gating MoE

In this section, we first establish the convergence rate of estimation for the regression function under some mild assumptions. This lays the foundation for deriving a condition to capture the optimal expert structure in terms of sample efficiency in Section 2.1. Next, we determine the convergence behavior of estimation for linear experts, which fail to satisfy that condition, in Section 2.2.

Assumptions. To begin with, let us introduce the essential yet mild assumptions for our subsequent convergence analysis.

(A.1) The parameter space Θ is a compact subset of $\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^q$, and its dimension is fixed unless stated otherwise. Meanwhile, the input space $\mathcal{X}$ is a bounded subset of $\mathbb{R}$.

(A.2) The ground-truth expert parameters $\eta_1^*, \eta_2^*, \ldots, \eta_{k_*}^*$ are distinct. Furthermore, the expert functions $\mathcal{E}(x, \eta)$ is bounded and Lipschitz continuous with respect to η for almost every x.

(A.3) The last pair of gating parameters are zero, that is, $\beta_{k_*}^* = 0$ and $\omega_{k_*}^* = 0_d$.

(A.4) At least one of the ground-truth gating parameters $\omega_1^*, \omega_2^*, \ldots, \omega_{k_*}^*$ is different from zero.

Above, the first assumption (A.1) is to ensure the convergence of least square estimators, whereas the second (A.2) is necessary for the distinction among experts. Next, the third assumption (A.3) helps prevent the invariance to translation of the softmax gating function, which negatively affects the identifiability of the softmax gating MoE model. The last assumption (A.4) is to guarantee that the gating value is input-dependent as in practice.

Given the above assumptions, we are ready to study the convergence behavior of regression estimation in the following proposition.

Proposition 1. *For a least square estimator $\widehat{G}_n$ in equation* (3)*, the convergence rate of regression function $f_{\widehat{G}_n}$ is given by*

$$\|f_{\widehat{G}_n} - f_{G_*}\|_{L^2(\mu)} = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}}). \tag{4}$$

The proof of Proposition 1 can be found in Appendix C.1. The bound in equation (4) indicates that the least square estimator $f_{\widehat{G}_n}$ converges to its true regression function f_{G_*} under the $L^2(\mu)$ norm at the rate of order $\mathcal{O}_P([\log(n)/n]^{\frac{1}{2}})$, which is parametric on the sample size n. As a consequence, in order to capture the convergence behavior of parameter estimation and expert estimation, it is sufficient to build a loss function among parameters $\mathcal{L}(\widehat{G}_n, G_*)$ such that $\|f_{\widehat{G}_n} - f_{G_*}\|_{L^2(\mu)} \gtrsim \mathcal{L}(\widehat{G}_n, G_*)$, which implies that $\mathcal{L}(\widehat{G}_n, G_*) = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}})$. Given these results, we attempt to construct a condition to characterize *strongly identifiable* expert functions that require fewer data points to estimate than others.

2.1 Strongly identifiable experts

Before presenting the condition for the strongly identifiable expert structure, let us highlight the challenges of establishing the essential L^2-lower bound $\|f_{\widehat{G}_n} - f_{G_*}\|_{L^2(\mu)} \gtrsim \mathcal{L}(G, G_*)$ to determine the convergence rates of parameter estimation and expert estimation.

Challenges. In our proof, a key step to derive the aforementioned L^2-lower bound is to decompose the discrepancy between the estimation of the regression function and its true counterpart, that is, $f_{\widehat{G}_n}(x) - f_{G_*}(x)$, using Taylor expansions to the function $x \mapsto F(x; \omega, \eta) := \exp(\omega^\top x)\mathcal{E}(x, \eta)$ staying implicit in the representation of regression function. Furthermore, it is necessary to guarantee that the function F and its partial derivatives resulting from the Taylor expansions are linearly independent so that when the regression discrepancy goes to zero as $n \to \infty$, the coefficients of those terms, which encompass the parameter discrepancies, also converge to zero. Therefore, we need to establish a non-trivial algebraic independence condition on the expert function $\mathcal{E}(x, \eta)$ called *strong identifiability* in Definition 1 to ensure such a linear independent property. This requires us to adopt new techniques as previous works on the softmax gating MoE [35, 31] employ only linear experts.

Definition 1 (Strong Identifiability). *An expert function $x \mapsto \mathcal{E}(x, \eta)$ is said to be strongly identifiable if it is twice differentiable with respect to its parameter η for almost every x, and the set of functions in x*

$$\left\{ x^\nu \cdot \frac{\partial^{|\rho|}\mathcal{E}}{\partial \eta^\rho}(x, \eta_j) : j \in [k], \nu \in \mathbb{N}^d, \ \rho \in \mathbb{N}^q, 0 \leq |\nu| + |\rho| \leq 2 \right\}$$

is linearly independent for almost every x for any $k \geq 1$ and distinct parameters $\eta_1, \eta_2 \ldots, \eta_k$.

Example. We can justify that several experts employed in practice, namely FFNs with activation functions like GELU [15], sigmoid, tanh, and non-linear transformed input, satisfy the strong identifiability condition. For instance, let us consider a two-layer FFN with normalized input, i.e.

$$\mathcal{E}(x, (a, b, c)) = c\sigma\left(a\frac{x}{\|x\|} + b\right),$$

where σ is one among the activation functions GELU, sigmoid and tanh, and $x, a \in \mathbb{R}^d, b, c, \in \mathbb{R}$. On the other hand, the strong identifiability will be violated if the activation function σ is of polynomial form, e.g., $\sigma(z) = z^p$ for all $z \in \mathbb{R}$ for some positive integer $p \in \mathbb{N}$.

Intuitively, the linear independence of functions in Definition 3.1 helps eliminate potential interactions among parameters expressed in the language of partial differential equations (see e.g., equation (10) and equation (16) where gating parameters β_1 interact with expert parameters a). Such interactions are demonstrated to result in significantly slow expert estimation rates (see Theorem 4.4 and Theorem 4.6).

Now, it is necessary to construct a loss function among parameters to capture the convergence rate of parameter estimation. In previous works, Nguyen et al. [36] utilized the generalized Wasserstein divergence between mixing measures to determine parameter estimation rates under the setting of classical mixture models. Then, this divergence was adopted again in [16] to establish the convergence rates of parameter estimation in Gaussian mixture of experts. However, there is an

inherent drawback of the generalized Wasserstein. In particular, this divergence implies estimation rates for all individual parameters while these rates should vary with the number of fitted parameters. In response to this issue, we propose using a loss function built based on the concept of Voronoi cells [27] to accurately characterize parameter estimation rates.
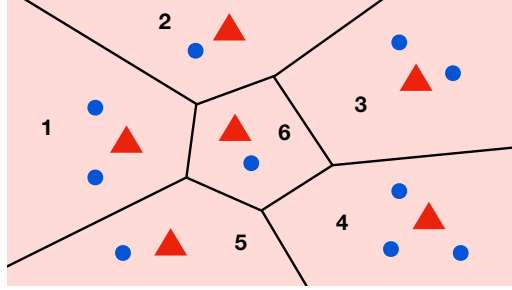

Figure 3: Illustration of Voronoi cells generated by $k_* = 6$ atoms of the ground-truth G_* (red triangles) and $k = 10$ fitted atoms of the estimator $\widehat{G}_n$ (blue rounds). By definition, each Voronoi cell is generated by one ground-truth atom, and its cardinality equals the number of corresponding fitted atoms. For instance, the red triangle in cell 4 is fitted by three blue rounds, implying that the cardinality of Voronoi cell 4 is three.

Voronoi loss function. For a mixing measure G with $1 \leq k' \leq k$ atoms, we partition its atoms into the set of Voronoi cells $\{\mathcal{A}_j \equiv \mathcal{A}_j(G) : j \in [k_*]\}$ generated by the atoms of the ground-truth mixing measure G_* defined as

$$\mathcal{A}_j := \left\{ i \in [k'] : \|\theta_i - \theta_j^*\| \leq \|\theta_i - \theta_\ell^*\|, \forall \ell \neq j \right\}, \tag{5}$$

with $\theta_i := (\omega_i, \eta_i)$ and $\theta_j^* := (\omega_j^*, \eta_j^*)$ for all $j \in [k_*]$ (see Figure 3). Then, the Voronoi loss of interest is given by

$$\mathcal{L}_1(G, G_*) := \sum_{j=1}^{k_*} \left| \sum_{i \in \mathcal{A}_j} \exp(\beta_i) - \exp(\beta_j^*) \right| + \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\beta_i) \left[\|\Delta\omega_{ij}\| + \|\Delta\eta_{ij}\| \right]$$

$$+ \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp(\beta_i) \left[\|\Delta\omega_{ij}\|^2 + \|\Delta\eta_{ij}\|^2 \right], \tag{6}$$

where we denote $\Delta\omega_{ij} := \omega_i - \omega_j$ and $\Delta\eta_{ij} := \eta_i - \eta_j^*$. It is worth noting that the cardinality of a Voronoi cell $\mathcal{A}_j$ is exactly the number of parameters approaching the ground-truth atom θ_j^* for all $j \in [k_*]$. By convention, if a Voronoi cell $\mathcal{A}_j$ is an empty set, then we set the respective summation to zero. Regarding the Voronoi loss function, we notice that $\mathcal{L}_1(G, G_*) = 0$ holds if and only if $G \equiv G_*$. This property indicates that when the Voronoi loss $\mathcal{L}_1(G, G_*)$ becomes small enough, the parameter discrepancies $\Delta\omega_{ij}$ and $\Delta\eta_{ij}$ are also small accordingly. Therefore, it is a suitable loss function for the sake of capturing parameter estimation rates. Furthermore, the Voronoi loss $\mathcal{L}_1(G, G_*)$ induces more accurate convergence rates of individual parameter estimation than the generalized Wasserstein divergence. In particular, suppose that there exist a sequence of mixing measures (G_n) such that $\mathcal{L}_1(G_n, G_*)$ converges to zero at a rate $\gamma_n = o(1)$ as $n \to \infty$, we deduce that the convergence rates for estimating *exactly-specified parameters* θ_j^*, whose Voronoi cells $\mathcal{A}_j$ have only one element, are also γ_n. Meanwhile, *over-specified parameters* θ_j^*, whose Voronoi cells have more than one element,

admit the estimation rates of order $\gamma_n^{\frac{1}{2}}$. On the other hand, if we used the generalized Wasserstein divergence, then the parameter estimation rates would be the same irrespective of the Voronoi cell cardinality. However, as the Voronoi loss $\mathcal{L}_1(G, G_*)$ is not symmetric, it is not a proper metric.

Given the above loss function, we provide the convergence rate of parameter estimation in Theorem 1.

Theorem 1. *Suppose that the expert function $x \mapsto \mathcal{E}(x, \eta)$ satisfies the strong identifiability condition presented in Definition 1, then the lower bound $\|f_G - f_{G_*}\|_{L^2(\mu)} \gtrsim \mathcal{L}_1(G, G_*)$ holds for all $G \in \mathcal{G}_k(\Theta)$, indicating that*

$$\mathcal{L}_1(\widehat{G}_n, G_*) = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}}). \tag{7}$$

The proof of Theorem 1 can be found in Appendix A. A few remarks regarding the result of this theorem are in order.

(i) Parameter estimation rates: Putting the bound (7) and the definition of the Voronoi loss $\mathcal{L}_1$ in equation (6) together, it follows that the rates for estimating exactly-specified parameters ω_j^*, η_j^* are of the order $\mathcal{O}_P([\log(n)/n]^{\frac{1}{2}})$, while over-specified parameters ω_j^*, η_j^* admit slower estimation rates of the order $\mathcal{O}_P([\log(n)/n]^{\frac{1}{4}})$. This highlights the benefit of the Voronoi loss function over the generalized Wasserstein divergence as the former can distinguish the estimation rates of exactly-specified and over-specified parameters while the latter induces identical convergence rates for all parameter estimations.

(ii) Expert estimation rates: By definition, since $\mathcal{E}(x, \eta)$ is a strongly identifiable expert function, it is twice differentiable over a bounded domain, implying that it is also a Lipschitz function. Assume that the least square estimator $\widehat{G}_n$ is represented as $\widehat{G}_n = \sum_{i=1}^{\widehat{k}_n} \exp(\widehat{\beta}_i)\delta_{(\widehat{\omega}_i^n, \widehat{\eta}_i^n)}$, then the following inequality holds for all $i \in \mathcal{A}_j(\widehat{G}_n)$ and $j \in [k_*]$:

$$\sup_x |\mathcal{E}(x, \widehat{\eta}_i^n) - \mathcal{E}(x, \eta_j^*)| \lesssim \|\widehat{\eta}_i^n - \eta_j^*\|. \tag{8}$$

From this result, we deduce that the convergence rates of estimating experts $\mathcal{E}(x, \eta_j^*)$ are identical to the rates for estimating parameters η_j^*, which are of the order $\mathcal{O}_P([\log(n)/n]^{\frac{1}{2}})$ when $|\mathcal{A}_j| = 1$ and become slower at the order $\mathcal{O}_P([\log(n)/n]^{\frac{1}{4}})$ when $|\mathcal{A}_j| > 1$. These rates imply that to achieve an approximation of the experts $\mathcal{E}(x, \eta_j^*)$ with a given error $\epsilon > 0$, we need polynomially many data points of order $\mathcal{O}(\epsilon^{-2})$ or $\mathcal{O}(\epsilon^{-4})$. To see the effects of the strong identifiability condition on the parameter and expert estimation problem more clearly, we will demonstrate in the next section that it requires a substantially larger sample size to estimate non-strongly identifiable experts.

2.2 Linear experts

Moving to this section, we complement our convergence analysis of parameter and expert estimations under the softmax gating MoE by taking into account a class of linear experts of the form $\mathcal{E}(x, (a, b)) = a^\top x + b$ for $(a, b) \in \mathbb{R}^d \times \mathbb{R}$, which violate the strong identifiability condition in Definition 1.

Parameter interaction. Let us recall that a key step in establishing the L^2-lower bound in Theorem 1 is to decompose the regression discrepancy $f_{\widehat{G}_n}(x) - f_{G_*}(x)$ into a combination of linearly

independent terms by applying Taylor expansions to the function $x \mapsto F(x; \omega, \eta) = \exp(\omega^\top x)\mathcal{E}(x, \eta)$. since the strong identifiability condition is not satisified under the scenario of linear experts $\mathcal{E}(x, (a, b)) = a^\top x + b$, there exist undesirable linear dependence among the function F and its derivative with respect to parameters a, b, η expressed via the following partial differential equation (PDE):

$$\frac{\partial^2 F}{\partial \omega \partial b}(x; \omega_i^*, a_i^*, b_i^*) = \frac{\partial F}{\partial a}(x; \omega_i^*, a_i^*, b_i^*). \tag{9}$$

We refer to the above PDE as an interaction between the gating parameters ω and the expert parameters a, b. This interaction has been encountered in the setting of softmax gating Gaussian mixture of linear experts [35]. More specifically, Nguyen et al. [35] showed that although parameter estimation rates were still of polynomial orders given the aforementioned parameter interaction, these rates were significantly slow, as they hinged upon the solvability of some intrinsic system of polynomial equations. Subsequently, we will demonstrate in Theorem 2 that the parameter estimation rates become significantly slower than polynomial orders under the deterministic softmax gating MoE due to the parameter interaction in equation (9). For that purpose, let us define a Voronoi loss function tailored to the setting of linear experts as follows:

$$\mathcal{L}_{2,r}(G, G_*) := \sum_{j=1}^{k_*} \Big| \sum_{i \in \mathcal{A}_j} \exp(\beta_i) - \exp(\beta_j^*) \Big| + \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp(\beta_i) \Big[\|\Delta\omega_{ij}\|^r + \|\Delta a_{ij}\|^r + \|\Delta b_{ij}\|^r \Big], \tag{10}$$

for any $r \geq 1$, where we denote $\Delta a_{ij} := a_i - a_j^*$ and $\Delta b_{ij} := b_i - b_j^*$.

Theorem 2. *Assume that the experts take the linear form $a^\top x + b$, then the following minimax lower bound of estimating G_* holds true for any $r \geq 1$:*

$$\inf_{\overline{G}_n \in \mathcal{G}_k(\Theta)} \sup_{G \in \mathcal{G}_k(\Theta) \backslash \mathcal{G}_{k_*-1}(\Theta)} \mathbb{E}_{f_G}[\mathcal{L}_{2,r}(\overline{G}_n, G)] \gtrsim n^{-1/2}, \tag{11}$$

where $\mathbb{E}_{f_G}$ indicates the expectation taken with respect to the product measure with f_G^n and the infimum is over all estimators taking values in $\mathcal{G}_k(\Theta)$.

The proof of Theorem 2 is in Appendix B.1. The bound in equation (11) reveals that the convergence rates of estimating ground-truth parameters ω_j^*, a_j^*, and b_j^* are slower than polynomial rates $\mathcal{O}_P(n^{-1/2r})$ for any $r \geq 1$. Therefore, these rates can become as slow as $\mathcal{O}_P(1/\log^\lambda(n))$ for some positive constant λ. Regarding the expert estimation, it is worth noting that

$$\sup_x \Big| (\widehat{a}_i^n)^\top x + \widehat{b}_i^n - ((a_j^*)^\top x + b_j^*) \Big| \leq \sup_x \|\widehat{a}_i^n - a_j^*\| \cdot \|x\| + |\widehat{b}_i^n - b_j^*|.$$

Recall from Assumption (A.1) that the input space $\mathcal{X}$ is bounded. Thus, the above inequality implies that the rates for estimating experts $(a_j^*)^\top x + b_j^*$ are also slower than any polynomial rates and could be of the order $\mathcal{O}_P(1/\log^\lambda(n))$. In that case, it requires exponentially many data points of order $\mathcal{O}(\exp(\epsilon^{-1/\lambda}))$ to obtain an expert approximation with a predetermined error ϵ. Compared to the results in Theorem 1, we see that using strongly identifiable experts is more sample efficient as we need only a polynomial number of data points to estimate them.

3 Dense-to-sparse Gating MoE

In this section, we investigate the convergence analysis of parameter estimation and expert estimation under the dense-to-sparse gating MoE. We begin our analysis with a linear router inside the dense-to-sparse gating in Section 3.1, and then extend to the setting of a class of general routers, which are optimal for the parameter and expert convergence, in Section 3.2.

3.1 Linear Router

Firstly, we will exhibit the problem setup for analyzing the convergence of dense-to-sparse gating MoE with a linear router.

Problem setting. We assume that an i.i.d. sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ in $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$ is also generated according to a regression framework as in equation (1) but with another MoE-based regression function as $Y_i = g_{G_*}(X_i) + \varepsilon_i$ for all $i \in [k_*]$, where the regression function $g_{G_*}(\cdot)$ is defined as

$$g_{G_*}(x) := \sum_{i=1}^{k_*} \text{Softmax}\Big(\frac{(\omega_i^*)^\top x + \beta_i^*}{\tau^*}\Big) \cdot \mathcal{E}(x, \eta_i^*). \tag{12}$$

Here, we keep the definitions and assumptions of the input X_i and the noise variable ε_i unchanged for all $i \in [k_*]$. Recall that the temperature τ^* is involved to control the sparsity level of the MoE model, making the expert selection process smooth and stable. In particular, when the temperature parameter goes to infinity, the weight distribution becomes roughly uniform, implying that all the experts are activated. In contrast, when the temperature parameter tends to zero, the mixture weights behave as a one-hot vector, meaning that only one expert is activated. Additionally, by abuse of notation, we still denote $G_* := \sum_{i=1}^{k_*} \exp(\beta_i^*) \delta_{(\omega_i^*, \tau^*, \eta_i^*)}$ as a mixing measure associated with unknown parameters $(\beta_i^*, \omega_i^*, \tau^*, \eta_i^*)_{i=1}^{k_*} \in \Theta \subseteq \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}_+ \times \mathbb{R}^q$ although this formulation is different from that in the setting of softmax gating MoE due to the appearance of the temperature parameter τ^*. Then, the least square estimator of the ground-truth mixing measure G_* is now given by:

$$\widetilde{G}_n := \underset{G \in \mathcal{G}_k(\Theta)}{\arg \min} \sum_{i=1}^{n} \Big(Y_i - g_G(X_i)\Big)^2, \tag{13}$$

where $\mathcal{G}_k(\Theta) := \{G = \sum_{i=1}^{k'} \exp(\beta_i) \delta_{(\omega_i, \tau, \eta_i)} : 1 \leq k' \leq k, \ (\beta_i, \omega_i, \tau, \eta_i) \in \Theta\}$ denotes the set of all mixing measures with no more than k atoms. Analogously to Section 2, we provide in the following proposition the convergence rate of estimating the ground-truth regression function $g_{G_*}(x)$.

Proposition 2. *For a least square estimator $\widetilde{G}_n$ in equation (13), the convergence rate of regression function $g_{\widetilde{G}_n}$ is given by*

$$\|g_{\widetilde{G}_n} - g_{G_*}\|_{L^2(\mu)} = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}}). \tag{14}$$

Since this proposition can be proved in a similar fashion to Proposition 1, its proof is omitted. The above result reveals that the regression function estimation $g_{\widetilde{G}_n}$ converges to its ground-truth

version g_{G_*} at the parametric rate on the sample size, that is, $\mathcal{O}_P([\log(n)/n]^{\frac{1}{2}})$. Next, we proceed to determine the convergence rates of parameter and expert estimations based on this result. It comes to our attention that although the temperature helps stabilize and smooth the expert selection process, it induces an intrinsic interaction with gating parameters which might harm the parameter and expert convergence.

Interaction between the temperature and gating parameters. Given the bound in equation (14), we continue to leverage the same strategy to determine parameter and expert estimation rates as in Section 2, that is, to derive the L^2 lower bound $\|g_{\widetilde{G}_n} - g_{G_*}\|_{L^2(\mu)} \gtrsim \mathcal{L}(\widetilde{G}_n, G_*)$ for some loss function $\mathcal{L}$ that will be defined later. However, an obstacle arises when decomposing the discrepancy $\|g_{\widetilde{G}_n} - g_{G_*}\|_{L^2(\mu)}$ into a combination of linearly independent terms, which is a key step in the strategy. In particular, we observe that there is an interaction between the temperature parameter τ and the gating parameter ω via the following PDE:

$$\frac{\partial F}{\partial \tau}(x; \omega_i^*, \tau^*, \eta_i^*) = -\frac{1}{\tau^*}(\omega_i^*)^\top \frac{\partial F}{\partial \omega}(x; \omega_i^*, \tau^*, \eta_i^*), \tag{15}$$

where we define $F(x; \omega, \tau, \eta) := \exp\left(\frac{\omega^\top x}{\tau}\right)\mathcal{E}(x, \eta)$ by abuse of notation. Unlike the interaction in equation (9) which occurs when the expert function takes a linear form, the above interaction holds true irrespective of the expert structure. As a result, we will illustrate in Theorem 3 that the parameter interaction (15) negatively affects the convergence of parameter and expert estimations under the dense-to-sparse gating MoE by involving the following Voronoi loss function:

$$\mathcal{L}_{3,r}(G, G_*) := \sum_{j=1}^{k_*} \Big| \sum_{i\in\mathcal{A}_j} \exp\left(\frac{\beta_{0i}}{\tau}\right) - \exp\left(\frac{\beta_{0j}^*}{\tau^*}\right) \Big| + \sum_{j=1}^{k_*} \sum_{i\in\mathcal{A}_j} \exp\left(\frac{\beta_{0i}}{\tau}\right)[\|\Delta\omega_{ij}\|^r + \|\Delta\tau\|^r + \|\Delta\eta_{ij}\|^r], \tag{16}$$

for any $r \geq 1$, where we denote $\Delta\tau := \tau - \tau^*$.

Theorem 3. *The following minimax lower bound of estimating G_* holds true for any $r \geq 1$:*

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

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

The proof of this theorem is deferred to Appendix B.2. Similarly to Theorem 2, the result of Theorem 3 indicates that the rates for estimating parameters ω_j^*, τ^*, and η_j^* are slower than polynomial rates of order $\mathcal{O}_P(n^{-1/2r})$ for any $r \geq 1$. Furthermore, it follows from the inequality (8) that the expert estimation shares the same convergence behavior. Thus, both the parameter and expert estimation rates could be as slow as $\mathcal{O}_P(1/\log^\lambda(n))$ for some constant $\lambda > 0$. It should be noted that this slow estimation rate applies for any expert function employed in the dense-to-sparse gating MoE model with the linear router (12). This situation necessitates a new router formulation to avoid the parameter interaction (15), thereby improving the parameter and expert estimation rates. We will address this issue in the next section.

3.2 General Router

In this section, we aim to explore a class of new router structures such that the interaction between the temperature and gating parameters via the PDE (15), which induces slow expert estimation rates, no longer exists. For that purpose, we first generalize the ground-truth regression function $g_{G_*}(x)$ in equation (12) by replacing a linear router $(\omega_j^*)^\top x$ with a general router $\pi(x, \omega_j^*)$ as follows:

$$g_{G_*}(x) := \sum_{i=1}^{k_*} \text{Softmax}\Big(\frac{\pi(x, \omega_i^*) + \beta_i^*}{\tau^*}\Big) \cdot \mathcal{E}(x, \eta_i^*). \tag{17}$$

Here, we will keep all the assumptions imposed on the setting of linear router unchanged throughout this section unless stating otherwise. Next, we introduce a so-called *algebraic independence* condition on the router function $x \mapsto \pi(x, \omega)$ and the expert function $\mathcal{E}(x, \eta)$ in Definition 2 to prevent any interaction among the temperature, gating and expert parameters expressed in the language of PDEs such as those in equation (15) from happening.

Definition 2 (Algebraic Independence). *We say that a router function $x \mapsto \pi(x, \omega)$ and an expert function $x \mapsto \mathcal{E}(x, \eta)$ are algebraically independent if they are twice differentiable w.r.t their parameters ω and η, and the set of functions in x*

$$\Big\{ \mathcal{E}(x, \eta_j), \pi(x, \omega_j)\mathcal{E}(x, \eta_j), \frac{\partial \pi}{\partial \omega^{(u)}}(x, \omega_j)\mathcal{E}(x, \eta_j), \frac{\partial \pi}{\partial \omega^{(u)}}(x, \omega_j)\frac{\partial \pi}{\partial \omega^{(v)}}(x, \omega_j)\mathcal{E}(x, \eta_j),$$

$$\frac{\partial^2 \pi}{\partial \omega^{(u)}\partial \omega^{(v)}}(x, \omega_j)\mathcal{E}(x, \eta_j), \pi(x, \omega_j)\frac{\partial \pi}{\partial \omega^{(u)}}(x, \omega_j)\mathcal{E}(x, \eta_j), \frac{\partial \mathcal{E}}{\partial \eta^{(u)}}(x, \eta_j), \pi(x, \omega_j)\frac{\partial \mathcal{E}}{\partial \eta^{(u')}}(x, \eta_j),$$

$$\frac{\partial \pi}{\partial \omega^{(u)}}(x, \omega_j)\frac{\partial \mathcal{E}}{\partial \eta^{(u')}}(x, \eta_j), \frac{\partial^2 \mathcal{E}}{\partial \eta^{(u')}\partial \eta^{(v')}}(x, \eta_j) : j \in [k], 1 \le u, v \le d, 1 \le u', v' \le q \Big\}.$$

is linearly independent for almost every x for any $k \ge 1$, distinct gating parameters $\omega_1, \omega_2, \ldots, \omega_k$ and distinct expert parameters $\eta_1, \eta_2, \ldots, \eta_k$.

Example. It can be validated that the router function $\pi(x, \omega) = \sigma_1(\omega^\top x)$ and the expert function $\mathcal{E}(x, \eta) = c\sigma_2\Big(a\frac{x}{\|x\|} + b\Big)$ are algebraically independent, where σ_1, σ_2 are two different activation functions among the functions GELU, sigmoid and tanh, and $\omega \in \mathbb{R}^q$, $x, a \in \mathbb{R}^d$, $b, c, \in \mathbb{R}$. However, if the activation functions σ_1 and σ_2 take polynomial forms $z \in \mathbb{R} \mapsto z^p$, for some $p \in \mathbb{N}$, then the previous pair of router function and expert function violates the algebraic independence condition.

Given the algebraic independence condition on the router function and the expert function, we are now ready to analyze the convergence of parameter and expert estimations under the dense-to-sparse gating MoE in Theorem 4, which involves the following Voronoi loss function:

$$\mathcal{L}_4(G, G_*) := \sum_{j=1}^{k_*} \Big| \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_{0i}}{\tau}\Big) - \exp\Big(\frac{\beta_{0j}^*}{\tau^*}\Big)\Big| + \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_{0i}}{\tau}\Big)[\|\Delta\omega_{ij}\| + \|\Delta\tau\| + \|\Delta\eta_{ij}\|]$$

$$+ \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_{0i}}{\tau}\Big)[\|\Delta\omega_{ij}\|^2 + \|\Delta\tau\|^2 + \|\Delta\eta_{ij}\|^2]. \tag{18}$$

Theorem 4. *Suppose that the router function $x \mapsto \pi(x, \omega)$ and the expert function $x \mapsto \mathcal{E}(x, \eta)$ are algebraically independent, then the lower bound $\|g_G - g_{G_*}\|_{L^2(\mu)} \gtrsim \mathcal{L}_4(G, G_*)$ holds for all $G \in \mathcal{G}_k(\Theta)$, indicating that*

$$\mathcal{L}_4(\widetilde{G}_n, G_*) = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}}). \tag{19}$$

The proof of Theorem 4 is in Appendix B.3.

(i) Parameter estimation rates: From the bound (19) and the formulation of the Voronoi loss $\mathcal{L}_4$, we deduce that exactly-specified parameters $\omega_j^*, \tau^*, \eta_j^*$ enjoy estimation rates of the order $\mathcal{O}_P([\log(n)/n]^{\frac{1}{2}})$, whereas those for their over-specified counterparts are of the order $\mathcal{O}_P([\log(n)/n]^{\frac{1}{4}})$. Compared to the linear router scenario in Theorem 3, we see that the convergence rates of parameter estimation are significantly improved when the router function and the expert function are algebraically independent.

(ii) Expert estimation rates: By employing the inequality (8), we notice that the convergence rates of expert estimation also benefit from the enhancement of parameter estimation rates. In particular, ground-truth experts $\mathcal{E}(x, \eta_j^*)$ admit estimation rates of the order $\mathcal{O}_P([\log(n)/n]^{\frac{1}{2}})$ when they are exactly-specified, and of the order $\mathcal{O}_P([\log(n)/n]^{\frac{1}{4}})$ when they are over-specified. As a consequence, only polynomially many data points of order $\mathcal{O}(\epsilon^{-2})$ or $\mathcal{O}(\epsilon^{-4})$ are required to obtain the expert approximation with the error ϵ, which are substantially improved in comparison with exponentially many data points required in the linear router scenario. This result indicates that the algebraic independence between the router function and the expert function helps increase the model sample efficiency.

4 Hierarchical Softmax Gating MoE

In this section, we extend our previous convergence analysis of parameter estimation and expert estimation to the setting of hierarchical MoE (HMoE) model. For simplicity, we will consider only the two-level structure of the HMoE with a note that higher-level HMoE models can be analyzed in a similar fashion.

Problem setting. Here, we continue to assume that $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ in $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$ is drawn from a regression framework $Y_i = h_{G_*}(X_i) + \varepsilon_i$ for all $i \in [k_*]$, with the HMoE-based regression function $h_{G_*}(\cdot)$ given by

$$h_{G_*}(x) := \sum_{i_1=1}^{k_1^*} \mathrm{Softmax}((\omega_{i_1}^*)^\top x + \beta_{i_1}^*) \sum_{i_2=1}^{k_2^*} \mathrm{Softmax}((\kappa_{i_2|i_1}^*)^\top x + \nu_{i_2|i_1}^*) \cdot \mathcal{E}(x, \eta_{i_1 i_2}^*). \tag{20}$$

Note that all the definitions and assumptions of the input X_i and the noise variable ε_i from previous sections will still be applied in this setting. Meanwhile, the ground-truth mixing measure G_* inherits the hierarchical structure of the HMoE and is formulated as $G_* := \sum_{i_1=1}^{k_1^*} \exp(\beta_{i_1}^*) \sum_{i_2=1}^{k_2^*} \exp(\nu_{i_2|i_1}^*) \delta_{(\omega_{i_1}^*, \kappa_{i_2|i_1}^*, \eta_{i_1 i_2}^*)}$, where $(\beta_{i_1}^*, \omega_{i_1}^*, \nu_{i_2|i_1}^*, \kappa_{i_2|i_1}^*, \eta_{i_1 i_2}^*)_{i_1 \in [k_1^*], i_2 \in [k_2^*]}$ denotes true yet unknown parameters in the compact space $\Theta \subseteq \mathbb{R} \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^q$. Throughout this work, we refer to k_1^* as the number of expert groups, while k_2^* represents the number of experts in each group. As the convergence analysis would

become needlessly complicated if the number of expert groups k_1^* was unknown, we assume that its value is known for ease of presentation. Nevertheless, the number of experts in each group k_2^* still remains unknown. Under these assumptions, we over-specify the ground-truth model (20) by taking into account a least square estimator within a class of HMoE models with k_1^* expert groups, each of which has at most $k_2 > k_2^*$ experts, as follows:

$$\check{G}_n := \underset{G \in \mathcal{G}_{k_1^* k_2}(\Theta)}{\arg\min} \sum_{i=1}^{n} \Big(Y_i - h_G(X_i) \Big)^2, \tag{21}$$

with

$$\mathcal{G}_{k_1^* k_2}(\Theta) := \{ G = \sum_{i_1=1}^{k_1^*} \exp(\beta_{i_1}) \sum_{i_2=1}^{k_2'} \exp(\nu_{i_2|i_1}) \delta_{(\omega_{i_1}, \kappa_{i_2|i_1}, \eta_{i_2|i_1})} : 1 \le k_2' \le k_2,$$

$$(\beta_{i_1}, \omega_{i_1}, \nu_{i_2|i_1}, \kappa_{i_2|i_1}, \eta_{i_1 i_2}) \in \Theta \}$$

standing for the set of all feasible mixing measures. Given the above estimator, we will demonstrate that the corresponding regression function estimation $h_{\check{G}_n}(x)$ converges to the ground-truth regression function $h_{G_*}(x)$ at the parametric rate on the sample size in Proposition 3 whose proof is left in Appendix C.2.

Proposition 3. *For a least square estimator $\check{G}_n$ in equation (21), the convergence rate of regression function $h_{\check{G}_n}$ is given by*

$$\| h_{\check{G}_n} - h_{G_*} \|_{L^2(\mu)} = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}}). \tag{22}$$

Toward the goal of analyzing the convergence behavior of parameter and expert estimations in the HMoE, we also need to decompose the difference of regression functions $h_{\check{G}_n}(x) - h_{G_*}(x)$ into a combination of linearly independent terms. However, due to the hierarchical structure of the HMoE, this decomposition will be achieved by applying Taylor expansions to the function $x \mapsto \check{F}(x; \omega, \kappa, \eta) := \exp(\omega^\top x) \exp(\kappa^\top x) \mathcal{E}(x, \eta)$ rather than the function $x \mapsto F(x; \omega, \eta) := \exp(\omega^\top x) \mathcal{E}(x, \eta)$ as in the MoE setting. Recall that, the function $\check{F}$ and its partial derivatives from the Taylor expansions are forced to be linearly independent so that when the previous regression difference approaches zero as $n \to \infty$, the coefficients of those terms, which contain parameter discrepancies, also go to zero, ensuring the convergence of parameter estimation. Notably, we will show in Section 4.1 that if the expert function $\mathcal{E}(x, \eta)$ meets the strong identifiability condition in Definition 1, then the aforementioned linear independence condition is secured and, therefore, ground-truth parameters and experts admit estimation rates of polynomial orders. On the other hand, if the expert function takes a linear form $\mathcal{E}(x, (a, b)) = a^\top x + b$, then it is not strongly identifiable, leading to slow parameter and expert estimation rates exhibited in Section 4.2.

4.1 Strongly identifiable experts

Before presenting the result statement, it is necessary to construct a Voronoi loss function tailored to the HMoE setting.

Voronoi loss function. Due to the two-level hierarchical structure of the HMoE, we need to employ distinct sets of Voronoi cells to capture the convergence of parameter estimations in each

level. In particular, given an arbitrary mixing measure $G \in \mathcal{G}_{k_1^* k_2}(\Theta)$, we partition its atoms into the Voronoi cells $\{\mathcal{A}_{j_1}(G) : j_1 \in [k_1^*]\}$ and $\{\mathcal{A}_{j_2|j_1}(G) : j_1 \in [k_1^*], j_2 \in [k_2^*]\}$ generated by the atoms of the ground-truth mixing measure G_* (see Figure 4), where

$$\mathcal{A}_{j_1} \equiv \mathcal{A}_{j_1}(G) := \{i_1 \in [k_1^*] : \|\omega_{i_1} - \omega_{j_1}^*\| \le \|\omega_{i_1} - \omega_{\ell_1}^*\|, \forall \ell_1 \ne j_1\}, \tag{23}$$

$$\mathcal{A}_{j_2|j_1} \equiv \mathcal{A}_{j_2|j_1}(G) := \{i_2 \in [k_2] : \|\theta_{i_2|j_1} - \theta_{j_2|j_1}^*\| \le \|\theta_{i_2|j_1} - \theta_{\ell_2|j_1}^*\|, \forall \ell_2 \ne j_2\}, \tag{24}$$

with $\theta_{i_2|j_1} := (\kappa_{i_2|j_1}, \eta_{j_1 i_2})$ and $\theta_{j_2|j_1}^* := (\kappa_{j_2|j_1}^*, \eta_{j_1 j_2}^*)$. Then, the Voronoi loss function of interest is defined as

$$\mathcal{L}_5(G, G_*) := \sum_{j_1=1}^{k_1^*} \Big| \sum_{i_1 \in \mathcal{A}_{j_1}} \exp(\beta_{i_1}) - \exp(\beta_{j_1}^*) \Big| + \sum_{j_1=1}^{k_1^*} \sum_{i_1 \in \mathcal{A}_{j_1}} \exp(\beta_{i_1}) \|\omega_{i_1} - \omega_{j_1}^*\|$$

$$+ \sum_{j_1=1}^{k_1^*} \sum_{i_1 \in \mathcal{A}_{j_1}} \exp(\beta_{i_1}) \Bigg[\sum_{j_2:|\mathcal{A}_{j_2|j_1}|=1} \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|i_1})(\|\kappa_{i_2|i_1} - \kappa_{j_2|j_1}^*\| + \|\eta_{i_1 i_2} - \eta_{j_1 j_2}^*\|)$$

$$+ \sum_{j_2:|\mathcal{A}_{j_2|j_1}|>1} \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|i_1})(\|\kappa_{i_2|i_1} - \kappa_{j_2|j_1}^*\|^2 + \|\eta_{i_1 i_2} - \eta_{j_1 j_2}^*\|^2) \Bigg]$$

$$+ \sum_{j_1=1}^{k_1^*} \sum_{i_1 \in \mathcal{A}_{j_1}} \exp(\beta_{i_1}) \sum_{j_2=1}^{k_2^*} \Big| \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|i_1}) - \exp(\nu_{j_2|j_1}^*) \Big|. \tag{25}$$

Equipped with the Voronoi loss $\mathcal{L}_5$, we will illustrate that strongly identifiable experts in the HMoE model admit estimation rates of polynomial orders in Theorem 5 whose proof is left in Appendix B.4.

Theorem 5. *Suppose that the expert function $x \mapsto \mathcal{E}(x, \eta)$ satisfies the strong identifiability condition presented in Definition 1, then the lower bound $\|h_G - h_{G_*}\|_{L^2(\mu)} \gtrsim \mathcal{L}_5(G, G_*)$ holds for all $G \in \mathcal{G}_{k_1^* k_2}(\Theta)$, indicating that*

$$\mathcal{L}_5(\check{G}_n, G_*) = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}}). \tag{26}$$

We provide below some implications on the parameter and expert estimation rates from the above results.

(i) Parameter estimation rates: It follows from the formulation of the Voronoi loss $\mathcal{L}_5$ and the bound (26) that all the first-level parameters $\omega_{j_1}^*$ share the same estimation rate of the order $\mathcal{O}_P([\log(n)/n]^{\frac{1}{2}})$. Regarding the second-level parameters $\kappa_{j_2|j_1}^*$ and $\eta_{j_1 j_1}^*$, the rates for estimating them are of the order $\mathcal{O}_P([\log(n)/n]^{\frac{1}{2}})$ if they are exactly-specified, that is, $|\mathcal{A}_{j_2|j_1}| = 1$. However, if they are over-specified, that is, $|\mathcal{A}_{j_2|j_1}| > 1$, their estimation rates become slower of the order $\mathcal{O}_P([\log(n)/n]^{\frac{1}{4}})$.

(ii) Expert estimation rates: Recall that a strongly identifiable expert function is twice differentiable over a bounded domain, so it is also a Lipschitz function, which implies that

$$\sup_x |\mathcal{E}(x, \check{\eta}_{i_1 i_2}^n) - \mathcal{E}(x, \eta_{j_1 j_2}^*)| \lesssim \|\check{\eta}_{i_1 i_2}^n - \eta_{j_1 j_2}^*\|. \tag{27}$$
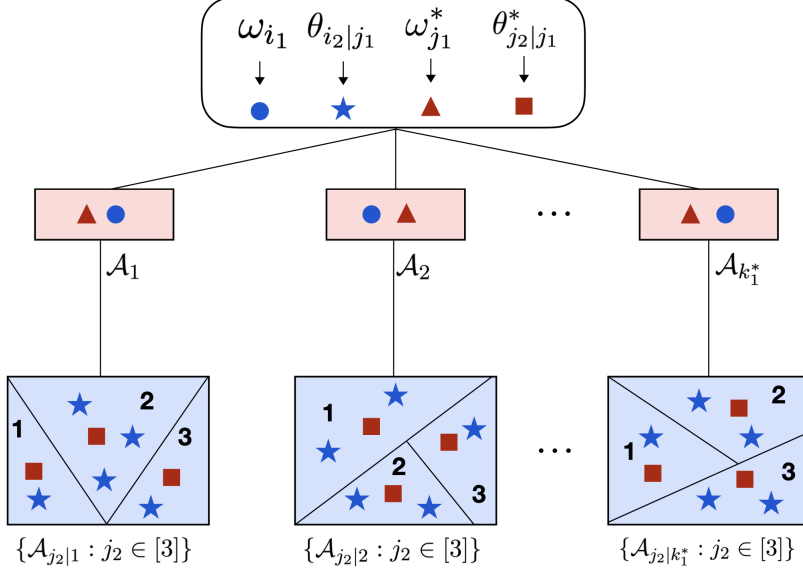
Figure 4: Illustration of Voronoi cells given in equations (23) and (24). Above, Voronoi cells $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_{k_1^*}$ in the first level are generated by ground-truth parameters $\omega_1^*, \omega_2^*, \ldots, \omega_{k_1^*}^*$ (red triangles), respectively. As the number of expert groups k_1^* is known, each Voronoi cell $\mathcal{A}_{j_1}$ contains one fitted parameter ω_{i_1} (blue round). In the second level, each rectangle represent a set of $k_2^* = 3$ Voronoi cells $\{\mathcal{A}_{j_2|j_1} : j_2 \in [k_2^*]\}$ generated by ground-truth parameters $\theta_{j_2|j_1}^* := (\kappa_{j_2|j_1}^*, \eta_{j_1 j_2}^*)$ (red squares), for $j_1 \in [k_1^*]$, each of which contains a total of $k_2 = 5$ fitted parameters $\theta_{i_2|j_1} := (\kappa_{i_2|j_1}, \eta_{i_2 j_1})$ (blue stars).

Together with the above parameter estimation rates, this inequality indicates that exactly-specified expert $\mathcal{E}(x, \eta_{j_1 j_2}^*)$ admit the estimation rates of the order $\mathcal{O}_P([\log(n)/n]^{\frac{1}{2}})$, while those for their over-specified counterparts are slower of the order $\mathcal{O}_P([\log(n)/n]^{\frac{1}{4}})$. As a consequence, we need polynomially many data points of the order $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-4})$ to approximate these experts with a given error ϵ. It can be seen that the expert convergence behavior under the two-level HMoE is similar to that under the MoE considered in Section 2.

4.2   Linear experts

In this section, we draw our attention to the convergence analysis for the hierarchical mixture of linear experts taking the form $\mathcal{E}(x, (a, b)) = a^\top x + b$ for $(a, b) \in \mathbb{R}^d \times \mathbb{R}$, which fail to meet the strong identifiability condition in Definition 1.

Parameter interaction. In order to obtain an L^2-lower bound as in Theorem 5, a crucial step is to apply Taylor expansions to the function $x \mapsto \check{F}(x; \omega, \kappa, a, b) = \exp(\omega^\top x) \exp(\kappa^\top x)(a^\top x + b)$ to decompose the regression discrepancy $h_{\check{G}_n}(x) - h_{G_*}(x)$ into a combination of linearly independent terms. Nevertheless, we discover that this step is hindered by interactions among parameters ω, κ, a, b

from both levels of the HMoE model via the following PDEs:

$$\frac{\partial^2 \breve{F}}{\partial \omega \partial b}(x; \omega_{j_1}^*, \kappa_{j_2|j_1}^*, a_{j_1 j_2}^*, b_{j_1 j_2}^*) = \frac{\partial^2 \breve{F}}{\partial \kappa \partial b}(x; \omega_{j_1}^*, \kappa_{j_2|j_1}^*, a_{j_1 j_2}^*, b_{j_1 j_2}^*) = \frac{\partial \breve{F}}{\partial a}(x; \omega_{j_1}^*, \kappa_{j_2|j_1}^*, a_{j_1 j_2}^*, b_{j_1 j_2}^*). \tag{28}$$

As a result of the above parameter interactions, we will exhibit in Theorem 6 that the convergence rates of parameter and expert estimations in the HMoE are negatively affected to be slower than polynomial orders. It is worth noting that since the number of expert groups k_1^* is known, the first-level parameters $\omega_{j_1}^*$ are exactly-specified and, thus, should have estimation rates of polynomial orders as in Theorem 5. Nevertheless, due to their interactions with second-level parameters $\kappa_{j_2|j_1}^*$ and $a_{j_1 j_2}^*, b_{j_1 j_2}^*$ in equation (28), the rates for estimating them will also be decelerated. To capture such convergence behavior, we will employ the following Voronoi loss function for our analysis:

$$\mathcal{L}_{6,r}(G, G_*) := \sum_{j_1=1}^{k_1^*} \Big| \sum_{i_1 \in \mathcal{A}_{j_1}} \exp(\beta_{i_1}) - \exp(\beta_{j_1}^*) \Big| + \sum_{j_1=1}^{k_1^*} \sum_{i_1 \in \mathcal{A}_{j_1}} \exp(\beta_{i_1}) \|\omega_{i_1} - \omega_{j_1}^*\|^r$$

$$+ \sum_{j_1=1}^{k_1^*} \sum_{i_1 \in \mathcal{A}_{j_1}} \exp(\beta_{i_1}) \sum_{j_2=1}^{k_2^*} \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|i_1})(\|\kappa_{i_2|i_1} - \kappa_{j_2|j_1}^*\|^r + \|a_{i_1 i_2} - a_{j_1 j_2}^*\|^r + |b_{i_1 i_2} - b_{j_1 j_2}^*|^r)$$

$$+ \sum_{j_1=1}^{k_1^*} \sum_{i_1 \in \mathcal{A}_{j_1}} \exp(\beta_{i_1}) \sum_{j_2=1}^{k_2^*} \Big| \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|i_1}) - \exp(\nu_{j_2|j_1}^*) \Big|, \tag{29}$$

for any $r \geq 1$. Now, we are ready to present the main result of this section.

Theorem 6. *Assume that the experts take the linear form $a^\top x + b$, then the following minimax lower bound of estimating G_* holds true for any $r \geq 1$:*

$$\inf_{\overline{G}_n \in \mathcal{G}_{k_1^* k_2}(\Theta)} \sup_{G \in \mathcal{G}_{k_1^* k_2}(\Theta) \setminus \mathcal{G}_{k_1^*(k_2^*-1)}(\Theta)} \mathbb{E}_{h_G}[\mathcal{L}_{6,r}(\overline{G}_n, G)] \gtrsim n^{-1/2}, \tag{30}$$

where $\mathbb{E}_{h_G}$ indicates the expectation taken with respect to the product measure with h_G^n and the infimum is over all estimators taking values in $\mathcal{G}_{k_1^* k_2}(\Theta)$.

The proof of Theorem 6 is in Appendix B.5. Putting the above result and the formulation of the Voronoi loss $\mathcal{L}_{6,r}$ together, it follows that the estimation rates for the parameters $\omega_{j_1}^*, \kappa_{j_2|j_1}^*, a_{j_1 j_2}^*, b_{j_1 j_2}^*$, which are involved in the interactions (28), are slower than polynomial orders $\mathcal{O}_P(n^{-1/2r})$ for any $r \geq 1$. Thus, these parameter estimation rates can become as slow as $\mathcal{O}_P(1/\log^\lambda(n))$ for some constant $\lambda > 0$. Furthermore, expert estimation rates are also affected by slow estimation rates of the expert parameters $a_{j_1 j_2}^*, b_{j_1 j_2}^*$ via the following inequality:

$$\sup_x \Big| ((\breve{a}_{i_1 i_2}^n)^\top x + \breve{b}_{i_1 i_2}^n) - ((a_{j_1 j_2}^*)^\top x + b_{j_1 j_2}^*) \Big| \leq \sup_x \|\breve{a}_{i_1 i_2}^n - a_{j_1 j_2}^*\| \cdot \|x\| + |\breve{b}_{i_1 i_2}^n - b_{j_1 j_2}^*|.$$

In particular, since the input space $\mathcal{X}$ is assumed to be bounded, then the estimation rates for the experts $(a_{j_1 j_2}^*)^\top x + b_{j_1 j_2}^*$ are slower than any polynomial order and, therefore, could be as slow as $\mathcal{O}_P(1/\log^\lambda(n))$. Consequently, we would need exponentially many data points of order $\mathcal{O}(\exp(\epsilon^{-1/\lambda}))$ to approximate them with a predetermined error ϵ, which is significantly larger than a polynomial number of data points needed for strongly identifiable experts. Hence, the claim that using strongly identifiable experts is more sample efficient than employing linear experts still holds under the HMoE setting.

5 Discussion

In this paper, we aim to investigate the impacts of the softmax gating and its variants, namely the dense-to-sparse gating and the hierarchical softmax gating, on the convergence behavior of parameter estimation and expert estimation in the mixture of experts (MoE). We show that the convergence rates of expert estimation are of polynomial orders when the expert function meets the strong identifiability condition. However, when using the dense-to-sparse gating with a linear router, the expert estimation rates become slower than any polynomial rates regardless of the expert structure. In response to this issue, we establish the algebraic independence condition to characterize the pairs of a router function and an expert function that lead to improved parameter and expert estimation rates of polynomial orders. On the other hand, our findings reveal that the convergence rates of estimating linear experts are always slower than polynomial rates irrespective of the gating choice. In conclusion, our convergence analysis provide two following important practical implications for the design of expert and gating structures:

(i) Expert design: Using strongly identifiable experts, namely experts formulated as two-layer feed-forward networks, is more sample-efficient than employing linear experts.

(ii) Gating design: When incorporating the temperature parameter into the MoE to smoothly adjust the model sparsity, it is necessary to replace a linear router with a general router that is algebraically independent of the expert function. For example, when experts are two-layer feed-forward networks, the router should take the form of a multi-layer perceptron following by a non-linear activation function.

At the same time, there are two inherent limitations in the analysis. Firstly, we assume that the data are generated from an MoE-based regression framework, which could not be satisfied in the real-world scenario where the regression function $s(\cdot)$ takes an arbitrary form. In that case, the least square estimator $\widehat{G}_n$ converges to the set of mixing measures $\overline{G} \in \mathcal{G}_k(\Theta)$ which minimizes the distance $\|f_G - s\|_{L^2(\mu)}$. Note that the current techniques of analyzing the convergence of least square estimation are tailored to the scenario when the function space is convex [44]. Since the space $\mathcal{G}_k(\Theta)$ is non-convex, it is necessary to develop further techniques to handle this real-world setting. Secondly, for the sake of theory, we consider only a single MoE layer, while practitioners often employ multiple MoE layers in practice [9, 20]. As these two potential directions stay beyond the scope of our work, we leave them for future development.

<div align="center">

Supplement to

"Convergence Rates for Softmax Gating Mixture of Experts"

</div>

A Proof of Theorem 1

Proof overview. In this proof, we focus on establishing the following inequality:

$$\inf_{G \in \mathcal{G}_k(\Theta)} \|f_G - f_{G_*}\|_{L^2(\mu)} / \mathcal{L}_1(G, G_*) > 0 \tag{31}$$

For this sake, we first derive the local part of the inequality (31), which is given by

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{G}_k(\Theta): \mathcal{L}_1(G, G_*) \leq \varepsilon} \|f_G - f_{G_*}\|_{L^2(\mu)} / \mathcal{L}_1(G, G_*) > 0. \tag{32}$$

We will prove the local part by contradiction. In particular, assume that it does not hold, then there exists a mixing measure sequence (G_n) such that $\mathcal{L}_1(G_n, G_*) \to 0$ and $\|f_{G_n} - f_{G_*}\|_{L^2(\mu)}/\mathcal{L}_1(G_n, G_*) \to 0$ as $n \to \infty$. Then, we divide the proof into three main stages.

Step 1 - Decompose the difference between regression functions. Firstly, we aim to decompose the regression discrepancy $f_{G_n}(x) - f_{G_*}(x)$ using Taylor expansions.

Step 2 - Non-vanishing coefficients. Secondly, we show that not all the coefficients in the decomposition of $[f_{G_n}(x) - f_{G_*}(x)]/\mathcal{L}_1(G_n, G_*)$ in Step 1 converge to zero.

Step 3 - Applications of the Fatou's lemma. Finally, we show a contradiction to the result in Step 2. Indeed, by means of the Fatou's lemma, since the term $\|f_{G_n} - f_{G_*}\|_{L^2(\mu)}/\mathcal{L}_1(G_n, G_*) \to 0$ as $n \to \infty$, we deduce $|f_{G_n} - f_{G_*}|/\mathcal{L}_1(G_n, G_*) \to 0$. Furthermore, as the expert function satisfies the strong identifiability condition in Definition 1, that is, the expert function and its derivatives are linearly independent, then all the coefficients in the representation of $[f_{G_n}(x) - f_{G_*}(x)]/\mathcal{L}_1(G_n, G_*)$ go to zero, contradicting to the result of Step 2. Hence, we obtain the local part (32).

As a consequence, we can find a positive constant ε' such that

$$\inf_{G \in \mathcal{G}_k(\Theta) : \mathcal{L}_1(G, G_*) \leq \varepsilon'} \|f_G - f_{G_*}\|_{L^2(\mu)}/\mathcal{L}_1(G, G_*) > 0.$$

Therefore, it suffices to prove the inequality

$$\inf_{G \in \mathcal{G}_k(\Theta) : \mathcal{L}_1(G, G_*) > \varepsilon'} \|f_G - f_{G_*}\|_{L^2(\mu)}/\mathcal{L}_1(G, G_*) > 0, \tag{33}$$

which we refer to as the global part of the inequality (31). Putting the local part (32) and the global part (33) together, we achieve our desired inequality in equation (31).

In the sequel, we will streamline the detailed proofs of the local part and the global part, respectively.

Proof of the local part (32). Suppose that the local part (32) does not hold, then we can find a sequence of mixing measures $G_n = \sum_{i=1}^{k_*} \exp(\beta_i^n) \delta_{(\omega_i^n, \eta_i^n)}$ in $\mathcal{G}_k(\Theta)$ satisfying $\mathcal{L}_{1n} := \mathcal{L}_1(G_n, G_*) \to 0$ and

$$\|f_{G_n} - f_{G_*}\|_{L^2(\mu)}/\mathcal{L}_{1n} \to 0, \tag{34}$$

as $n \to \infty$. For ease of presentation, we denote $\mathcal{A}_j^n := \mathcal{A}_j(G_n)$ as a Voronoi cell of G_n generated by the j-th atom of G_*, for all $j \in [k_*]$. As we use asymptotic arguments throughout this proof, these Voronoi cells can be assumed to be independent of the sample size n, that is, $\mathcal{A}_j = \mathcal{A}_j^n$. Then, we can rewrite the Voronoi loss $\mathcal{L}_{1n}$ as

$$\mathcal{L}_{1n} := \sum_{j=1}^{k_*} \left| \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) - \exp(\beta_j^*) \right| + \sum_{j : |\mathcal{A}_j| > 1} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) \left[\|\Delta\omega_{ij}^n\|^2 + \|\Delta\eta_{ij}^n\|^2 \right]$$
$$+ \sum_{j : |\mathcal{A}_j| = 1} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) \left[\|\Delta\omega_{ij}^n\| + \|\Delta\eta_{ij}^n\| \right], \tag{35}$$

where we denote $\Delta\omega_{ij}^n := \omega_i^n - \omega_j^*$ and $\Delta\eta_{ij}^n := \eta_i^n - \eta_j^*$.

Recall that $\mathcal{L}_{1n} \to 0$ as $n \to \infty$, then we have that $(\omega_i^n, \eta_i^n) \to (\omega_j^*, \eta_j^*)$ and $\sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) \to \exp(\beta_j^*)$ for any $i \in \mathcal{A}_j$ and $j \in [k_*]$. Subsequently, we divide the proof of local part into three main steps. In

the first step, we attempt to decompose the difference $f_{G_n}(x) - f_{G_*}(x)$ into a combination of linearly independent terms using Taylor expansions. Then, we demonstrate that not all the coefficients in the representation of $[f_{G_n}(x) - f_{G_*}(x)]/\mathcal{L}_{1n}$ converge to zero as $n \to \infty$ in the second step. In the final step, we employ the Fatou's lemma to show a contradiction that all the previous coefficients necessarily go to zero and, thus, complete the proof of the local part.

Step 1 - Decompose the difference between regression functions. To begin with, we use Taylor expansions to decompose the quantity $Q_n(x) := [\sum_{j=1}^{k_*} \exp((\omega_j^*)^\top x + \beta_j^*)] \cdot [f_{G_n}(x) - f_{G_*}(x)]$ into a combination of linearly independent terms. For this sake, we denote $F(x; \omega, \eta) := \exp(\omega^\top x)\mathcal{E}(x, \eta)$ and $H(x; \omega) = \exp(\omega^\top x)f_{G_n}(x)$. Then, $Q_n(x)$ can be decomposed as

$$
\begin{aligned}
Q_n(x) = &\sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n)\Big[F(x; \omega_i^n, \eta_i^n) - F(x; \omega_j^*, \eta_j^*)\Big] \\
&- \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n)\Big[H(x; \omega_i^n) - H(x; \omega_j^*)\Big] \\
&+ \sum_{j=1}^{k_*} \Big(\sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) - \exp(\beta_j^*)\Big)\Big[F(x; \omega_j^*, \eta_j^*) - H(x; \omega_j^*)\Big] \\
&:= A_n(x) - B_n(x) + C_n(x).
\end{aligned}
\tag{36}
$$

Next, we will decompose the terms $A_n(x)$ and $B_n(x)$, respectively.

Step 1A - Decompose $A_n(x)$. We can represent $A_n(x)$ as

$$
\begin{aligned}
A_n(x) := &\sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n)\Big[F(x; \omega_i^n, \eta_i^n) - F(x; \omega_j^*, \eta_j^*)\Big] \\
&+ \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n)\Big[F(x; \omega_i^n, \eta_i^n) - F(x; \omega_j^*, \eta_j^*)\Big] \\
&:= A_{n,1}(x) + A_{n,2}(x).
\end{aligned}
$$

Regarding the term $A_{n,1}(x)$, by applying the first-order Taylor expansion to the function $F(x; \omega_i^n, \eta_i^n)$ around the point (ω_j^*, η_j^*), we have

$$
A_{n,1}(x) = \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) \sum_{|\alpha|=1} (\Delta\omega_{ij}^n)^{\alpha_1}(\Delta\eta_{ij}^n)^{\alpha_2} \cdot \frac{\partial F}{\partial\omega^{\alpha_1}\partial\eta^{\alpha_2}}(x; \omega_j^*, \eta_j^*) + R_1(x),
$$

where $R_1(x)$ is a Taylor remainder such that $R_1(x)/\mathcal{L}_{1n} \to 0$ as $n \to \infty$. Note that the first derivatives of F w.r.t its parameters ω and η are given by

$$
\begin{aligned}
\frac{\partial F}{\partial\omega}(x; \omega_j^*, \eta_j^*) &= x \exp((\omega_j^*)^\top x)\mathcal{E}(x, \eta_j^*) = x \cdot F(x; \omega_j^*, \eta_j^*), \\
\frac{\partial F}{\partial\eta}(x; \omega_j^*, \eta_j^*) &= \exp((\omega_j^*)^\top x) \cdot \frac{\partial\mathcal{E}}{\partial\eta}(x, \eta_j^*) := F_1(x; \omega_j^*, \eta_j^*).
\end{aligned}
$$

Then, $A_{n,1}(x)$ can be rewritten as

$$
A_{n,1}(x) = \sum_{j:|\mathcal{A}_j|=1} E_{n,1,j}(x) + R_1(x),
\tag{37}
$$

21

in which $E_{n,1,j}(x) = \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) \left[(\Delta \omega_{ij}^n)^\top x \cdot F(x; \omega_j^*, \eta_j^*) + (\Delta \eta_{ij}^n)^\top F_1(x; \omega_j^*, \eta_j^*) \right]$.

Analogously, by means of the second-order Taylor expansion, we can decompose $A_{n,2}(x)$ as

$$A_{n,2}(x) = \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) \sum_{|\alpha|=1}^{2} \frac{1}{\alpha!} (\Delta \omega_{ij}^n)^{\alpha_1} (\Delta \eta_{ij}^n)^{\alpha_2} \cdot \frac{\partial^{|\alpha_1|+|\alpha_2|} F}{\partial \omega^{\alpha_1} \partial \eta^{\alpha_2}} (x; \omega_j^*, \eta_j^*) + R_2(x),$$

where $R_2(x)$ is a Taylor remainder such that $R_2(x)/\mathcal{L}_{1n} \to 0$ as $n \to \infty$. By taking the second derivatives of F w.r.t its parameters, we have

$$\frac{\partial^2 F}{\partial \omega \partial \omega^\top} (x; \omega_j^*, \eta_j^*) = xx^\top \cdot F(x; \omega_j^*, \eta_j^*), \quad \frac{\partial^2 F}{\partial \omega \partial \eta^\top} (x; \omega_j^*, \eta_j^*) = x \cdot [F_1(x; \omega_j^*, \eta_j^*)]^\top,$$

$$\frac{\partial^2 F}{\partial \eta \partial \eta^\top} (x; \omega_j^*, \eta_j^*) = \exp((\omega_j^*)^\top x) \cdot \frac{\partial^2 \mathcal{E}}{\partial \eta \partial \eta^\top} := F_2(x; \omega_j^*, \eta_j^*)$$

Then, we can represent $A_{n,2}(x)$ as

$$A_{n,2}(x) = \sum_{j:|\mathcal{A}_j|>1} [E_{n,1,j}(x) + E_{n,2,j}(x)] + R_2(x), \tag{38}$$

where

$$E_{n,2,j}(x) := \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) \left\{ \left[x^\top \left(M_d \odot (\Delta \omega_{ij}^n)(\Delta \omega_{ij}^n)^\top \right) x \right] \cdot F(x; \omega_j^*, a_j^*, b_j^*) \right.$$

$$\left. + \left[x^\top (\Delta \omega_{ij}^n)(\Delta \eta_{ij}^n)^\top F_1(x; \omega_j^*, \eta_j^*) \right] + \left[(\Delta \eta_{ij}^n)^\top \left(M_d \odot F_2(x; \omega_j^*, \eta_j^*) \right)(\Delta \eta_{ij}^n) \right] \right\},$$

with M_d being an $d \times d$ matrix whose diagonal entries are $\frac{1}{2}$ while other entries are 1.

Step 1B - Decompose $B_n(x)$. Next, we apply the same strategy of decomposing $A_n(x)$ for partitioning the term $B_n(x)$. In particular, we first have

$$B_n(x) = \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) \left[H(x; \omega_i^n) - H(x; \omega_j^*) \right]$$

$$+ \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) \left[H(x; \omega_i^n) - H(x; \omega_j^*) \right]$$

$$:= B_{n,1}(x) + B_{n,2}(x).$$

Regarding the term $B_{n,1}(x)$, by employing the first-order Taylor expansion to the function $H(x; \omega_i^n)$ around the point ω_j^*, we have

$$B_{n,1}(x) = \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n)(\Delta \omega_{ij}^n)^\top x \cdot H(x; \omega_j^*) + R_3(x), \tag{39}$$

where $R_3(x)$ is a Taylor remainder such that $R_3(x)/\mathcal{L}_{1n} \to 0$ as $n \to \infty$. Similarly, by means of the second-order Taylor expansion, we get

$$B_{n,2}(x) = \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) \left[(\Delta \omega_{ij}^n)^\top x + (\Delta \omega_{ij}^n)^\top \left(M_d \odot xx^\top \right)(\Delta \omega_{ij}^n) \right] \cdot H(x; \omega_j^*) + R_4(x), \tag{40}$$

where $R_4(x)$ is a Taylor remainder such that $R_4(x)/\mathcal{L}_{1n} \to 0$ as $n \to \infty$.

Putting the above results together, we see that $[A_n(x) - R_1(x) - R_2(x)]/\mathcal{L}_{1n}$, $[B_n(x) - R_3(x) - R_4(x)]/\mathcal{L}_{1n}$ and $C_n(x)/\mathcal{L}_{1n}$ can be written as a combination of elements from the following set

$$\left\{ F(x; \omega_j^*, \eta_j^*),\ x^{(u)} F(x; \omega_j^*, \eta_j^*),\ x^{(u)} x^{(v)} F(x; \omega_j^*, \eta_j^*) : u, v \in [d],\ j \in [k_*] \right\},$$

$$\cup \left\{ [F_1(x; \omega_j^*, \eta_j^*)]^{(u)},\ x^{(u)} [F_1(x; \omega_j^*, \eta_j^*)]^{(v)} : u, v \in [d],\ j \in [k_*] \right\},$$

$$\cup \left\{ [F_2(x; \omega_j^*, \eta_j^*)]^{(uv)} : u, v \in [d],\ j \in [k_*] \right\},$$

$$\cup \left\{ H(x; \omega_j^*),\ x^{(u)} H(x; \omega_j^*),\ x^{(u)} x^{(v)} H(x; \omega_j^*) : u, v \in [d],\ j \in [k_*] \right\}.$$

Step 2 - Non-vanishing coefficients. Moving to this step, we utilize the proof by contradiction method to show that not all coefficients in the representations of $[A_n - R_1(x) - R_2(x)]/\mathcal{L}_{1n}$, $[B_n - R_3(x) - R_4(x)]/\mathcal{L}_{1n}$ and $C_n(x)/\mathcal{L}_{1n}$ converge to zero as $n \to \infty$. Indeed, assume by contrary that all of them go to zero and consider the coefficients of the term

- $F(x; \omega_j^*, \eta_j^*)$ for $j \in [k_*]$, we have

$$\frac{1}{\mathcal{L}_{1n}} \cdot \sum_{j=1}^{k_*} \left| \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) - \exp(\beta_j^*) \right| \to 0;$$

- $x^{(u)} F(x; \omega_j^*, \eta_j^*)$ for $u \in [d]$ and $j : |\mathcal{A}_j| = 1$, we have

$$\frac{1}{\mathcal{L}_{1n}} \cdot \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) \|\Delta \omega_{ij}^n\|_1 \to 0.$$

  Since the ℓ_1-norm is equivalent to the ℓ_2-norm, we deduce

$$\frac{1}{\mathcal{L}_{1n}} \cdot \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) \|\Delta \omega_{ij}^n\| \to 0;$$

- $[F_1(x; \omega_j^*, \eta_j^*)]^{(u)}$ for $u \in [d]$ and $j : |\mathcal{A}_j| = 1$, we have

$$\frac{1}{\mathcal{L}_{1n}} \cdot \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) \|\Delta \eta_{ij}^n\| \to 0;$$

- $[x^{(u)}]^2 F(x; \omega_j^*, \eta_j^*)$ for $u \in [d]$ and $j : |\mathcal{A}_j| > 1$, we have

$$\frac{1}{\mathcal{L}_{1n}} \cdot \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) \|\Delta \omega_{ij}^n\|^2 \to 0 :$$

- $[F_2(x; \omega_j^*, \eta_j^*)]^{(uu)}$ for $u \in [d]$ and $j : |\mathcal{A}_j| > 1$, we have

$$\frac{1}{\mathcal{L}_{1n}} \cdot \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) \|\Delta \eta_{ij}^n\|^2 \to 0;$$

23

Combine all the above limits, we arrive at $1 = \mathcal{L}_{1n}/\mathcal{L}_{1n} \to 0$ as $n \to \infty$, which is a contradiction. Thus, it follows that at least one among the coefficients in the representations of $[A_n(x) - R_1(x) - R_2(x)]/\mathcal{L}_{1n}$, $[B_n(x) - R_3(x) - R_4(x)]/\mathcal{L}_{1n}$ and $C_n(x)/\mathcal{L}_{1n}$ does not converge to zero.

Step 3 - Application of the Fatou's lemma. Now, we attempt to exhibit a contradiction to the conclusion of Step 2. Let m_n be the maximum of the absolute values of the coefficients in the representations of $[A_n(x) - R_1(x) - R_2(x)]/\mathcal{L}_{1n}$, $[B_n(x) - R_3(x) - R_4(x)]/\mathcal{L}_{1n}$ and $C_n(x)/\mathcal{L}_{1n}$. As at least one among those coefficients does not go zero, we have $1/m_n \not\to \infty$.

Recall from the hypothesis in equation (34) that we have $\|f_{G_n} - f_{G_*}\|_{L^2(\mu)}/\mathcal{L}_{1n} \to 0$ as $n \to \infty$. This result also implies that $\|f_{G_n} - f_{G_*}\|_{L^1(\mu)}/\mathcal{L}_{1n} \to 0$. Then, by applying the Fatou's lemma, we get

$$0 = \lim_{n\to\infty} \frac{\|f_{G_n} - f_{G_*}\|_{L^1(\mu)}}{m_n \mathcal{L}_{1n}} \geq \int \liminf_{n\to\infty} \frac{|f_{G_n}(x) - f_{G_*}(x)|}{m_n \mathcal{L}_{1n}} \mathrm{d}\mu(x) \geq 0.$$

As a result, we have $[f_{G_n}(x) - f_{G_*}(x)]/[m_n \mathcal{L}_{1n}] \to 0$ for almost every x. Since the parameter Θ is compact, the term $\sum_{j=1}^{k_*} \exp((\omega_j^*)^\top x + \beta_j^*)$ is bounded. Thus, it follows that $Q_n(x)/[m_n \mathcal{L}_{1n}] \to 0$ as $n \to \infty$, or equivalently,

$$\frac{1}{m_n \mathcal{L}_{1n}} \cdot \Big[(A_{n,1}(x) - R_1(x) + A_{n,2}(x) - R_2(x)) - (B_{n,1}(x) - R_3(x) + B_{n,2}(x) - R_4(x)) + C_n(x) \Big] \to 0. \tag{41}$$

For ease of presentation, we denote

$$\frac{1}{m_n \mathcal{L}_{1n}} \cdot \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n)(\Delta\omega_{ij}^n) \to \phi_{1,j}, \quad \frac{1}{m_n \mathcal{L}_{1n}} \cdot \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n)(\Delta\omega_{ij}^n)(\Delta\omega_{ij}^n)^\top \to \phi_{2,j},$$

$$\frac{1}{m_n \mathcal{L}_{1n}} \cdot \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n)(\Delta\eta_{ij}^n) \to \varphi_{1,j}, \quad \frac{1}{m_n \mathcal{L}_{1n}} \cdot \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n)(\Delta\eta_{ij}^n)(\Delta\eta_{ij}^n)^\top \to \varphi_{2,j},$$

$$\frac{1}{m_n \mathcal{L}_{1n}} \cdot \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n)(\Delta\omega_{ij}^n)(\Delta\eta_{ij}^n)^\top \to \zeta_j, \quad \frac{1}{m_n \mathcal{L}_{1n}} \cdot \Big(\sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) - \exp(\beta_j^*) \Big) \to \xi_j,$$

$$F_{\rho j} := F_\rho(x; \omega_j^*, \eta_j^*), \quad H_j = H(x; \omega_j^*).$$

Above, at least one among the limits $\phi_{1,j}^{(u)}$, $\phi_{2,j}^{(uu)}$, $\varphi_{1,j}^{(u)}$, $\varphi_{2,j}^{(uu)}$ and ξ_j, for $j \in [k_*]$, is non-zero as a result of the conclusion in Step 2. From the formulation of

- $A_{n,1}(x)$ in equation (37), we have

$$\frac{A_{n,1}(x) - R_1(x)}{m_n \mathcal{L}_{1n}} \to \sum_{j:|\mathcal{A}_j|=1} \Big[\phi_{1,j}^\top x \cdot F_j + \varphi_{1,j}^\top F_{1j} \Big]. \tag{42}$$

- $A_{n,2}(x)$ in equation (38), we have

$$\frac{A_{n,2}(x) - R_2(x)}{m_n \mathcal{L}_{2n}} \to \sum_{j:|\mathcal{A}_j|>1} \Bigg\{ \Big[\phi_{1,j}^\top x + x^\top \Big(M_d \odot \phi_{2,j} \Big) x \Big] \cdot F_j + [\varphi_{1,j}^\top + x^\top \zeta_j] \cdot F_{1j}$$

$$+ \Big[M_d \odot \varphi_{2,j} \Big] \odot F_{2j} \Bigg\}. \tag{43}$$

24

- $B_{n,1}(x)$ in equation (39), we have

$$\frac{B_{n,1}(x) - R_3(x)}{m_n \mathcal{L}_{2n}} \to \sum_{j:|\mathcal{A}_j|=1} [\phi_{1,j}^\top x \cdot H_j]. \tag{44}$$

- $B_{n,2}(x)$ in equation (40), we have

$$\frac{B_{n,2}(x) - R_4(x)}{m_n \mathcal{L}_{2n}} \to \sum_{j:|\mathcal{A}_j|>1} \left[\phi_{1,j}^\top x + x^\top \left(M_d \odot \phi_{2,j}\right)x\right] \cdot H_j. \tag{45}$$

- $C_n(x)$ in equation (36), we have

$$\frac{C_n(x)}{m_n \mathcal{L}_{2n}} \to \sum_{j=1}^{k_*} \xi_j [F_j - H_j]. \tag{46}$$

Note that from equation (41), it follows that the sum of the limits in equations (42), (43), (44), (45) and (46) is equal to zero, that is,

$$\sum_{j:|\mathcal{A}_j|=1} \left[\phi_{1,j}^\top x \cdot F_j + \varphi_{1,j}^\top F_{1j}\right] + \sum_{j:|\mathcal{A}_j|>1} \left\{\left[\phi_{1,j}^\top x + x^\top \left(M_d \odot \phi_{2,j}\right)x\right] \cdot F_j + [\varphi_{1,j}^\top + x^\top \zeta_j] \cdot F_{1j}\right.$$

$$\left. + \left[M_d \odot \varphi_{2,j}\right] \odot F_{2j}\right\} - \sum_{j:|\mathcal{A}_j|=1} [\phi_{1,j}^\top x \cdot H_j] - \sum_{j:|\mathcal{A}_j|>1} \left[\phi_{1,j}^\top x + x^\top \left(M_d \odot \phi_{2,j}\right)x\right] \cdot H_j$$

$$+ \sum_{j=1}^{k_*} \xi_j [F_j - H_j] = 0. \tag{47}$$

Subsequently, we aim to demonstrate that the values of $\phi_{1,j}^{(u)}$, $\phi_{2,j}^{(uu)}$, $\varphi_{1,j}^{(u)}$, $\varphi_{2,j}^{(uu)}$ and ξ_j are all zero for all $j \in [k_*]$. Indeed, let $P_1, P_2, \ldots, P_\ell$ be the partition of the set $\{\exp((\omega_j^*)^\top x) : j \in [k_*]\}$, for some $\ell \in [k_*]$, such that

(i) $\omega_j^* = \omega_{j'}^*$ for any $j, j' \in P_i$ and $i \in [\ell]$;

(ii) $\omega_j^* \neq \omega_{j'}^*$ when j and j' do not belong to the same set P_i for any $i \in [\ell]$.

Then, the set $\{\exp((\omega_{j_1}^*)^\top x), \ldots, \exp((\omega_{j_\ell}^*)^\top x)\}$, where $j_i \in P_i$, is linearly independent. This result together with equation (47) implies that for any $i \in [\ell]$, we have

$$\sum_{j \in P_i:|\mathcal{A}_j|=1} \left[(\xi_j + \phi_{1,j}^\top x) \cdot \mathcal{E}_j + \varphi_{1,j}^\top \mathcal{E}_{1j}\right] + \sum_{j \in P_i:|\mathcal{A}_j|>1} \left\{\left[\phi_{1,j}^\top x + x^\top \left(M_d \odot \phi_{2,j}\right)x\right] \cdot \mathcal{E}_j\right.$$

$$\left. + [\varphi_{1,j}^\top + x^\top \zeta_j] \cdot \mathcal{E}_{1j} + \left[M_d \odot \varphi_{2,j}\right] \odot \mathcal{E}_{2j}\right\} - \sum_{j \in P_i:|\mathcal{A}_j|=1} [(\phi_{1,j}^\top x + \xi_j) \cdot f_{G_*}(x)]$$

$$- \sum_{j \in P_i:|\mathcal{A}_j|>1} \left[\xi_j + \phi_{1,j}^\top x + x^\top \left(M_d \odot \phi_{2,j}\right)x\right] \cdot f_{G_*}(x) = 0,$$

25

where we denote $\mathcal{E}_j := \mathcal{E}(x, \eta_j^*)$, $\mathcal{E}_{1j} := \frac{\partial \mathcal{E}}{\partial \eta}(x, \eta_j^*)$ and $\mathcal{E}_{2j} := \frac{\partial^2 \mathcal{E}}{\partial \eta \partial \eta^\top}(x, \eta_j^*)$. Furthermore, since the expert function $x \mapsto \mathcal{E}(x, \eta)$ satisfies the strong identifiability condition in Definition 1, then the set

$$\left\{ x^\nu \cdot \frac{\partial^{|\rho|} \mathcal{E}}{\partial \eta^\rho}(x, \eta_j^*) : j \in [k_*], \nu \in \mathbb{N}^d, \ \rho \in \mathbb{N}^q, 0 \le |\nu| + |\rho| \le 2 \right\}$$

is linearly independent for almost every x. Thus, it follows that $\xi_j = 0$, $\phi_{1,j} = \varphi_{1,j} = 0_d$ and $\phi_{2,j} = \varphi_{2,j} = \zeta_j = 0_{d \times d}$ for any $j \in P_i$ and $i \in [\ell]$. This contradicts the fact that not all the values of $\phi_{1,j}^{(u)}$, $\phi_{2,j}^{(uu)}$, $\varphi_{1,j}^{(u)}$, $\varphi_{2,j}^{(uu)}$ and ξ_j, for $j \in [k_*]$, are zero. Therefore, we obtain the local part (32), that is,

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{G}_k(\Theta): \mathcal{L}_1(G, G_*) \le \varepsilon} \|f_G - f_{G_*}\|_{L^2(\mu)} / \mathcal{L}_1(G, G_*) > 0.$$

Consequently, we can find a positive constant ε' such that

$$\inf_{G \in \mathcal{G}_k(\Theta): \mathcal{L}_1(G, G_*) \le \varepsilon'} \|f_G - f_{G_*}\|_{L^2(\mu)} / \mathcal{L}_1(G, G_*) > 0.$$

Proof of the global part (33)**:** Given the above result, it is sufficient to prove that

$$\inf_{G \in \mathcal{G}_k(\Theta): \mathcal{L}_1(G, G_*) > \varepsilon'} \|f_G - f_{G_*}\|_{L^2(\mu)} / \mathcal{L}_1(G, G_*) > 0.$$

Suppose that the global part does not hold, then there exists a sequence of mixing measures $G_n' \in \mathcal{G}_k(\Theta)$ satisfying $\mathcal{L}_1(G_n', G_*) > \varepsilon'$ and

$$\lim_{n \to \infty} \frac{\|f_{G_n'} - f_{G_*}\|_{L^2(\mu)}}{\mathcal{L}_1(G_n', G_*)} = 0,$$

indicating that $\|f_{G_n'} - f_{G_*}\|_{L^2(\mu)} \to 0$ as $n \to \infty$. Since the parameter space Θ is compact, we can substitute the sequence (G_n') with one of its subsequences that converges to $G' \in \mathcal{G}_k(\Theta)$. In addition, as we have $\mathcal{L}_1(G_n', G_*) > \varepsilon'$, it follows that $\mathcal{L}_1(G', G_*) > \varepsilon'$. Now, by applying the Fatou's lemma, we get

$$0 = \lim_{n \to \infty} \|f_{G_n'} - f_{G_*}\|_{L^2(\mu)}^2 \ge \int \liminf_{n \to \infty} \left| f_{G_n'}(x) - f_{G_*}(x) \right|^2 \, \mathrm{d}\mu(x).$$

The above result implies that $f_{G'}(x) = f_{G_*}(x)$ for almost every x. From Proposition 4 in Appendix C.3, we obtain $G' \equiv G_*$. As a consequence, we have $\mathcal{L}_1(G', G_*) = 0$, which contradicts the fact that $\mathcal{L}_1(G', G_*) > \varepsilon' > 0$. Hence, we achieve the global part and complete the proof of Theorem 1.

B Proofs of other Theorems

B.1 Proof of Theorem 2

Proof overview. In this proof, we first demonstrate that the limit

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{G}_k(\Theta): \mathcal{L}_{2,r}(G, G_*) \le \varepsilon} \|f_G - f_{G_*}\|_{L^2(\mu)} / \mathcal{L}_{2,r}(G, G_*) = 0 \tag{48}$$

holds for any $r \geq 1$. Given the above result, we proceed to derive the minimax lower bound in Theorem 2:

$$\inf_{\overline{G}_n \in \mathcal{G}_k(\Theta)} \sup_{G \in \mathcal{G}_k(\Theta) \setminus \mathcal{G}_{k_*-1}(\Theta)} \mathbb{E}_{f_G}[\mathcal{L}_{2,r}(\overline{G}_n, G)] \gtrsim n^{-1/2}. \tag{49}$$

Proof of equation (48): It suffices to construct a mixing measure sequence (G_n) satisfying $\mathcal{L}_{2,r}(G_n, G_*) \to 0$ and

$$\|f_{G_n} - f_{G_*}\|_{L^2(\mu)} / \mathcal{L}_{2,r}(G_n, G_*) \to 0, \tag{50}$$

as $n \to \infty$. Now, let us take into account the sequence (G_n) defined as $G_n := \sum_{i=1}^{k_*+1} \exp(\beta_i^n) \delta_{(\omega_i^n, a_i^n, b_i^n)}$, where we set

$\exp(\beta_1^n) = \exp(\beta_2^n) = \frac{1}{2}\exp(\beta_1^*) + \frac{1}{2n^{r+1}}$ and $\exp(\beta_i^n) = \exp(\beta_{(i-1)}^n)$ for any $3 \leq i \leq k_* + 1$;

$\omega_1^n = \omega_2^n = \omega_1^*$ and $\omega_i^n = \omega_{(i-1)}^n$ for any $3 \leq i \leq k_* + 1$;

$a_1^n = a_2^n = a_1^*$ and $a_i^n = a_{i-1}^n$ for any $3 \leq i \leq k_* + 1$;

$b_1^n = b_1^* + \frac{1}{n}$, $b_2^n = b_1^* - \frac{1}{n}$ and $b_i^n = b_{i-1}^n$ for any $3 \leq i \leq k_* + 1$.

Then, the Voronoi loss $\mathcal{L}_{2,r}(G_n, G_*)$ can be rewritten as

$$\mathcal{L}_{2,r}(G_n, G_*) = \frac{1}{n^{r+1}} + \left[\exp(\beta_1^*) + \frac{1}{n^{r+1}}\right] \cdot \frac{1}{n^r} = \mathcal{O}(n^{-r}). \tag{51}$$

The above formulation implies that $\mathcal{L}_{2,r}(G_n, G_*) \to 0$ as $n \to \infty$. Thus, it suffices to derive equation (50). To this end, we decompose the quantity $Q_n(x) := [\sum_{j=1}^{k_*} \exp((\omega_j^*)^\top x + \beta_j^*)] \cdot [f_{G_n}(x) - f_{G_*}(x)]$ as

$$Q_n(x) = \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n)\Big[\exp((\omega_i^n)^\top x)((a_i^n)^\top x + b_i^n) - \exp((\omega_j^*)^\top x)((a_j^*)^\top x + b_j^*)\Big]$$

$$- \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp(\beta_i^n)\Big[\exp((\omega_i^n)^\top x) - \exp((\omega_j^*)^\top x)\Big] f_{G_n}(x)$$

$$+ \sum_{j=1}^{k_*} \Big(\sum_{i \in \mathcal{A}_j} \exp(\beta_i^n) - \exp(\beta_j^*)\Big)\Big[\exp((\omega_j^*)^\top x)((a_j^*)^\top x + b_j^*) - \exp((\omega_j^*)^\top x)f_{G_n}(x)\Big]$$

$$:= A_n(x) - B_n(x) + C_n(x).$$

By plugging the values of $\exp(\beta_i^n), \omega_i^n, a_i^n$ and b_i^n into the terms $A_n(x), B_n(x)$ and $C_n(x)$, we have

$$A_n(x) = \sum_{i=1}^{2} \exp(\beta_i^n)\exp((\omega_1^*)^\top x)(b_i^n - b_1^*) = \frac{1}{2}\left[\exp(\beta_1^*) + \frac{1}{n^{r+1}}\right]\exp((\omega_1^*)^\top x)[(b_1^n - b_1^*) + (b_2^n - b_1^*)] = 0,$$

$$B_n(x) = \sum_{i=1}^{2} \exp(\beta_i^n)[\exp((\omega_i^n)^\top x) - \exp((\omega_1^*)^\top x)]f_{G_n}(x) = 0,$$

$$C_n(x) = \Big(\sum_{i=1}^{2} \exp(\beta_i^n) - \exp(\beta_1^*)\Big)\Big[\exp((\beta_1^*)^\top x)((a_1^*)^\top x + b_1^*) - \exp((\beta_1^*)^\top x)f_{G_n}(x)\Big] = \mathcal{O}(n^{-(r+1)})$$

Since $\mathcal{L}_{2,r}(G_n, G_*) = \mathcal{O}(n^{-r})$, we can justify that $C_n(x)/\mathcal{L}_{2,r}(G_n, G_*) \to 0$, thereby leading to $Q_n(x)/\mathcal{L}_{2,r}(G_n, G_*) \to 0$ as $n \to \infty$ for almost every x. Note that the term $\sum_{j=1}^{k_*} \exp((\omega_j^*)^\top x + \beta_j^*)$ is bounded, then we also have that $[f_{G_n}(x) - f_{G_*}(x)]/\mathcal{L}_{2,r} \to 0$ for almost every x. As a consequence, we obtain $\|f_{G_n} - f_{G_*}\|_{L^2(\mu)}/\mathcal{L}_{2,r} \to 0$ as $n \to \infty$. Hence, the proof of claim (48) is completed.

Proof of equation (49)**:** Since the noise variables ε_i given the input X_i follows Gaussian distributions, we have $Y_i|X_i \sim \mathcal{N}(f_{G_*}(x_i), \nu)$ for all $i \in [n]$. Next, it follows from equation (48) that for a small enough constant $\varepsilon > 0$ and a fixed constant $c > 0$ that will be determined later, there exists a mixing measure $G'_* \in \mathcal{G}_k(\Theta)$ satisfying $\mathcal{L}_{2,r}(G'_*, G_*) = 2\varepsilon$ and $\|f_{G'_*} - f_{G_*}\|_{L^2(\mu)} \le c \cdot \varepsilon$. According to Le Cam's lemma [48], since the Voronoi loss $\mathcal{L}_{2,r}$ satisfies the weak triangle inequality, we have

$$
\inf_{\overline{G}_n \in \mathcal{G}_k(\Theta)} \sup_{G \in \mathcal{G}_k(\Theta) \setminus \mathcal{G}_{k_*-1}(\Theta)} \mathbb{E}_{f_G}[\mathcal{L}_{2,r}(\overline{G}_n, G)]
$$
$$
\gtrsim \frac{\mathcal{L}_{2,r}(G'_*, G_*)}{8} \exp(-n\mathbb{E}_{X \sim \mu}[\mathrm{KL}(\mathcal{N}(f_{G'_*}(x), \nu), \mathcal{N}(f_{G_*}(x), \nu))])
$$
$$
\gtrsim \varepsilon \cdot \exp(-n\|f_{G'_*} - f_{G_*}\|_{L^2(\mu)}^2)
$$
$$
\gtrsim \varepsilon \cdot \exp(-cn\varepsilon^2), \tag{52}
$$

where the second inequality is due to the fact that $\mathrm{KL}(\mathcal{N}(f_{G'_*}(x), \nu), \mathcal{N}(f_{G_*}(x), \nu)) = \frac{[f_{G'_*}(x) - f_{G_*}(x)]^2}{2\nu}$. By setting $\varepsilon = n^{-1/2}$, we get $\varepsilon \cdot \exp(-cn\varepsilon^2) = n^{-1/2} \exp(-c)$. Hence, we obtain the desired minimax lower bound in equation (49) and complete the proof.

B.2   Proof of Theorem 3

In this proof, we will focus on showing that the limit

$$
\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{G}_k(\Theta): \mathcal{L}_{3,r}(G, G_*) \le \varepsilon} \frac{\|g_G - g_{G_*}\|_{L^2(\mu)}}{\mathcal{L}_{3,r}(G, G_*)} = 0, \tag{53}
$$

holds for any $r \ge 1$. Then, by employing the arguments for proving equation (49) in Appendix B.1, we achieve the minimax lower bound in Theorem 3, that is,

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

Proof of equation (53)**:** It is sufficient to build a mixing measure sequence (G_n) that satisfies $\mathcal{L}_{3,r}(G_n, G_*) \to 0$ and

$$
\|g_G - g_{G_*}\|_{L^2(\mu)}/\mathcal{L}_{3,r}(G_n, G_*) \to 0, \tag{54}
$$

as $n \to \infty$. For this sake, let us consider the sequence defined as $G_n := \sum_{i=1}^{k_*} \exp\left(\frac{\beta_i^n}{\tau^n}\right) \delta_{(\omega_i^n, \tau^n, \eta_i^n)}$, where we define for any $j \in [k_*]$ that

$\eta_i^n = \eta_j^*$, for any $i \in \mathcal{A}_j$;

$\omega_i^n = \omega_j^* + s_{n,j}$, for any $i \in \mathcal{A}_j$;

$\tau^n = \tau^* + t_n$;

$\beta_i^n = \tau^n \cdot \left[\frac{\beta_j^*}{\tau^*} - \log(|\mathcal{A}_j|)\right]$, which implies that $\sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right) = \exp\left(\frac{\beta_j^*}{\tau^*}\right)$, for any $i \in \mathcal{A}_j$,

where $s_{n,j} := (s_{n,j}^{(1)}, \ldots, s_{n,j}^{(d)}) \in \mathbb{R}^d$ and $t_n \in \mathbb{R}$ will be determined later such that $s_{n,j}^{(u)} \to 0$ and $t_n \to 0$ as $n \to \infty$ for any $u \in [d]$ and $j \in [k_*]$. Then, we can represent the Voronoi loss $\mathcal{L}_{3,r}(G_n, G_*)$ as

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

Due to the aforementioned properties of sequences $(s_{n,j})$ and (t_n), it follows that $\mathcal{L}_{3,r}(G_n, G_*) \to 0$ as $n \to \infty$. Thus, we can complete the proof by deriving equation (54).

To achieve this goal, we take into account the quantity $Q_n(x) := \left[\sum_{j=1}^{k_*} \exp\left(\frac{(\omega_i^*)^\top x + \beta_i^*}{\tau^*}\right)\right] \cdot \left[g_{G_n}(x) - g_{G_*}(x)\right]$, which can be decomposed as

$$Q_n(x) = \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)\left[\exp\left(\frac{(\omega_i^n)^\top x}{\tau^n}\right)\mathcal{E}(x, \eta_i^n) - \exp\left(\frac{(\omega_j^*)^\top x}{\tau^*}\right)\mathcal{E}(x, \eta_j^*)\right]$$

$$- \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)\left[\exp\left(\frac{(\omega_i^n)^\top x}{\tau^n}\right)g_{G_n}(x) - \exp\left(\frac{(\omega_j^*)^\top x}{\tau^*}\right)g_{G_n}(x)\right]$$

$$+ \sum_{j=1}^{k_*} \left[\sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right) - \exp\left(\frac{\beta_j^*}{\tau^*}\right)\right]\left[\exp\left(\frac{(\omega_j^*)^\top x}{\tau^*}\right)\mathcal{E}(x, \eta_j^*) - \exp\left(\frac{(\omega_j^*)^\top x}{\tau^*}\right)g_{G_n}(x)\right]$$

$$:= A_n(x) - B_n(x) + C_n(x).$$

By plugging in the values of η_i^n, ω_i^n, τ^n and β_i^n, the term $A_n(x)$ becomes

$$A_n(x) = \sum_{i=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)\left[\exp\left(\frac{(\omega_i^n)^\top x}{\tau^n}\right) - \exp\left(\frac{(\omega_j^*)^\top x}{\tau^*}\right)\right]\mathcal{E}(x, \eta_j^*).$$

By applying the first-order Taylor expansions to the function $\exp\left(\frac{(\omega_i^n)^\top x}{\tau^n}\right)$ around the point (ω_j^*, τ^*), we have

$$A_n(x) = \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \sum_{u=1}^{d} \exp\left(\frac{\beta_i^n}{\tau^n}\right)\left[\frac{s_{n,j}^{(u)}}{\tau^*} - \frac{t_n(\omega_j^*)^{(u)}}{(\tau^*)^2}\right] \cdot x^{(u)} \exp\left(\frac{(\omega_j^*)^\top x}{\tau^*}\right)\mathcal{E}(x, \eta_j^*) + R_1(x),$$

where $R_1(x)$ is a Taylor remainder such that $R_1(x)/\mathcal{L}_{3,r}(G_n, G_*) \to 0$ as $n \to \infty$. Then, by setting

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

we get $A_n(x)/\mathcal{L}_{3,r}(G_n, G_*) \to 0$ as $n \to \infty$.

Analogously, we also have $B_n(x)/\mathcal{L}_{3,r}(G_n, G_*) \to 0$ as $n \to \infty$. Furthermore, since it can be checked that $C_n(x) = 0$, we deduce $Q_n(x)/\mathcal{L}_{3,r}(G_n, G_*) \to 0$ as $n \to \infty$. Note that the term $\left[\sum_{j=1}^{k_*} \exp\left(\frac{(\omega_j^*)^\top x + \beta_j^*}{\tau^*}\right)\right]$ is bounded, then it follows that

$$|g_{G_n}(x) - g_{G_*}(x)|/\mathcal{L}_{3,r}(G_n, G_*) \to 0$$

as $n \to \infty$ for almost every x. The above result directly leads to equation (54). Hence, the proof is completed.

B.3 Proof of Theorem 4

In this proof, we aim to establish the following inequality:

$$\inf_{G \in \mathcal{G}_k(\Theta)} \|g_G - g_{G_*}\|_{L^2(\mu)} / \mathcal{L}_4(G, G_*) > 0. \tag{55}$$

By employing the proof framework of Theorem 1 in Appendix A, we will first derive the local part of equation (55), which is given by

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{G}_k(\Theta): \mathcal{L}_4(G, G_*) \leq \varepsilon} \|g_G - g_{G_*}\|_{L^2(\mu)} / \mathcal{L}_4(G, G_*) > 0. \tag{56}$$

As a consequence, there exists some $\varepsilon' > 0$ such that

$$\inf_{G \in \mathcal{G}_k(\Theta): \mathcal{L}_4(G, G_*) \leq \varepsilon'} \|g_G - g_{G_*}\|_{L^2(\mu)} / \mathcal{L}_4(G, G_*) > 0.$$

Thus, it is sufficient to prove the global part of equation (55), that is,

$$\inf_{G \in \mathcal{G}_k(\Theta): \mathcal{L}_4(G, G_*) > \varepsilon'} \|g_G - g_{G_*}\|_{L^2(\mu)} / \mathcal{L}_4(G, G_*) > 0.$$

Notably, since the global part can be argued in a similar fashion to that in Appendix A, it is omitted here. In other words, we will provide only the proof of local part here.

Proof of the local part (56): Assume by contrary that the above inequality does not hold true, then there exists a sequence of mixing measures $G_n = \sum_{i=1}^{k_*} \exp(\beta_i^n) \delta_{(\omega_i^n, \tau^n, \eta_i^n)}$ in $\mathcal{G}_k(\Theta)$ such that $\mathcal{L}_{4n} := \mathcal{L}_4(G_n, G_*) \to 0$ and

$$\|g_{G_n} - g_{G_*}\|_{L^2(\mu)} / \mathcal{L}_{4n} \to 0, \tag{57}$$

as $n \to \infty$. Let us denote by $\mathcal{A}_j^n := \mathcal{A}_j(G_n)$ a Voronoi cell of G_n generated by the j-th components of G_*. Since our arguments are asymptotic, we may assume that those Voronoi cells do not depend on the sample size, i.e. $\mathcal{A}_j = \mathcal{A}_j^n$. Thus, the Voronoi loss $\mathcal{L}_{4n}$ can be represented as

$$\mathcal{L}_{4n} := \sum_{j=1}^{k_*} \left| \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right) - \exp\left(\frac{\beta_j^*}{\tau^*}\right) \right| + \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right) \left[\|\Delta\omega_{ij}^n\| + |\Delta\tau^n| + \|\Delta\eta_{ij}^n\| \right]$$

$$+ \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right) \left[\|\Delta\omega_{ij}^n\|^2 + |\Delta\tau^n|^2 + \|\Delta\eta_{ij}^n\|^2 \right], \tag{58}$$

where we denote $\Delta\omega_{ij}^n := \omega_i^n - \omega_j^*$, $\Delta\tau^n := \tau^n - \tau^*$, and $\Delta\eta_{ij}^n := \eta_i^n - \eta_j^*$.

Since $\mathcal{L}_{4n} \to 0$, we get that $(\omega_i^n, \tau^n, \eta_i^n) \to (\omega_j^*, \tau^*, \eta_j^*)$ and $\sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right) \to \exp\left(\frac{\beta_j^*}{\tau^*}\right)$ as $n \to \infty$ for any $i \in \mathcal{A}_j$ and $j \in [k_*]$. Now, we divide the proof of local part into three steps as follows:

30

Step 1 - Decompose the difference between regression functions. To begin with, we use Taylor expansions to decompose the quantity $Q_n(x) := \left[\sum_{j=1}^{k_*} \exp \left(\frac{\pi(x,\omega_j^*) + \beta_j^*}{\tau^*} \right) \right] \cdot [g_{G_n}(x) - g_{G_*}(x)]$ into a combination of linearly independent terms. For this sake, we denote $F(x; \omega, \tau, \eta) := \exp \left(\frac{\pi(x,\omega)}{\tau} \right) \mathcal{E}(x, \eta)$ and $H(x; \omega, \tau) = \exp \left(\frac{\pi(x,\omega)}{\tau} \right) g_{G_n}(x)$. Then, $Q_n(x)$ can be decomposed as

$$
\begin{aligned}
Q_n(x) = & \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp \left(\frac{\beta_i^n}{\tau^n} \right) \left[F(x; \omega_i^n, \tau^n, \eta_i^n) - F(x; \omega_j^*, \tau^*, \eta_j^*) \right] \\
& - \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp \left(\frac{\beta_i^n}{\tau^n} \right) \left[H(x; \omega_i^n, \tau^n) - H(x; \omega_j^*, \tau^*) \right] \\
& + \sum_{j=1}^{k_*} \left(\sum_{i \in \mathcal{A}_j} \exp \left(\frac{\beta_i^n}{\tau^n} \right) - \exp \left(\frac{\beta_j^*}{\tau^*} \right) \right) \left[F(x; \omega_j^*, \tau^*, \eta_j^*) - H(x; \omega_j^*, \tau^*) \right] \\
& := A_n(x) - B_n(x) + C_n(x).
\end{aligned}
\tag{59}
$$

Next, we will decompose the terms $A_n(x)$ and $B_n(x)$, respectively.

Step 1A- Decompose $A_n(x)$. We can represent $A_n(x)$ as

$$
\begin{aligned}
A_n(x) := & \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp \left(\frac{\beta_i^n}{\tau^n} \right) \left[F(x; \omega_i^n, \tau^n, \eta_i^n) - F(x; \omega_j^*, \tau^*, \eta_j^*) \right] \\
& + \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp \left(\frac{\beta_i^n}{\tau^n} \right) \left[F(x; \omega_i^n, \tau^n, \eta_i^n) - F(x; \omega_j^*, \tau^*, \eta_j^*) \right] \\
& := A_{n,1}(x) + A_{n,2}(x).
\end{aligned}
$$

By applying the first-order Taylor expansion, we have

$$
A_{n,1}(x) = \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp \left(\frac{\beta_i^n}{\tau^n} \right) \sum_{|\alpha|=1} (\Delta \omega_{ij}^n)^{\alpha_1} (\Delta \tau^n)^{\alpha_2} (\Delta \eta_{ij}^n)^{\alpha_3} \cdot \frac{\partial F}{\partial \omega^{\alpha_1} \partial \tau^{\alpha_2} \partial \eta^{\alpha_3}} (x; \omega_j^*, \tau^*, \eta_j^*)
$$
$$
+ R_1(x),
$$

where $R_1(x)$ is a Taylor remainder such that $R_1(x)/\mathcal{L}_{4n} \to 0$ as $n \to \infty$. Note that the first derivatives of F w.r.t its parameters are given by

$$
\begin{aligned}
\frac{\partial F}{\partial \omega}(x; \omega_j^*, \tau^*, \eta_j^*) &= \frac{1}{\tau^*} \cdot \frac{\partial \pi}{\partial \omega}(x, \omega_j^*) \exp \left(\frac{\pi(x, \omega_j^*)}{\tau^*} \right) \mathcal{E}(x, \eta_j^*) = \frac{1}{\tau^*} \cdot \frac{\partial \pi}{\partial \omega}(x, \omega_j^*) F(x; \omega_j^*, \tau^*, \eta_j^*), \\
\frac{\partial F}{\partial \tau}(x; \omega_j^*, \tau^*, \eta_j^*) &= -\frac{\pi(x, \omega_j^*)}{(\tau^*)^2} \exp \left(\frac{\pi(x, \omega_j^*)}{\tau^*} \right) \mathcal{E}(x, \eta_j^*) = -\frac{\pi(x, \omega_j^*)}{(\tau^*)^2} F(x; \omega_j^*, \tau^*, \eta_j^*), \\
\frac{\partial F}{\partial \eta}(x; \omega_j^*, \tau^*, \eta_j^*) &= \exp \left(\frac{\pi(x, \omega_j^*)}{\tau^*} \right) \frac{\partial \mathcal{E}}{\partial \eta}(x, \eta_j^*) := F_1(x; \omega_j^*, \tau^*, \eta_j^*).
\end{aligned}
$$

Then, $A_{n,1}(x)$ can be rewritten as

$$
A_{n,1}(x) = \sum_{j:|\mathcal{A}_j|=1} A_{n,1,j}(x) + R_1(x),
\tag{60}
$$

31

where

$$A_{n,1,j}(x) := \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)\left[\left(\frac{1}{\tau^*}(\Delta\omega_{ij}^n)^\top \frac{\partial \pi}{\partial \omega}(x, \omega_j^*) - (\Delta\tau^n) \cdot \frac{\pi(x, \omega_j^*)}{(\tau^*)^2}\right) \cdot F(x; \omega_j^*, \tau^*, \eta_j^*) + \right.$$

$$\left. (\Delta\eta_{ij}^n)^\top F_1(x; \omega_j^*, \tau^*, \eta_j^*)\right].$$

Next, by means of the second-order Taylor expansion, we can decompose $A_{n,2}(x)$ as

$$A_{n,2}(x) = \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right) \sum_{|\alpha|=1}^{2} \frac{1}{\alpha!}(\Delta\omega_{ij}^n)^{\alpha_1}(\Delta\tau^n)^{\alpha_2}(\Delta\eta_{ij}^n)^{\alpha_3} \cdot \frac{\partial^{|\alpha_1|+\alpha_2+|\alpha_3|}F}{\partial\omega^{\alpha_1}\partial\tau^{\alpha_2}\partial\eta^{\alpha_3}}(x; \omega_j^*, \tau^*, \eta_j^*) + R_2(x),$$

where $R_2(x)$ is a Taylor remainder such that $R_2(x)/\mathcal{L}_{4n} \to 0$ as $n \to \infty$. By taking the second derivatives of F w.r.t its parameters, we have

$$\frac{\partial^2 F}{\partial\omega\partial\omega^\top}(x; \omega_j^*, \tau^*, \eta_j^*) = \left[\frac{1}{\tau^*}\frac{\partial^2 \pi}{\partial\omega\partial\omega^\top}(x, \omega_j^*) + \frac{1}{(\tau^*)^2}\frac{\partial\pi}{\partial\omega}(x, \omega_j^*) \cdot \frac{\partial\pi}{\partial\omega^\top}(x, \omega_j^*)\right]F(x; \omega_j^*, \tau^*, \eta_j^*),$$

$$\frac{\partial^2 F}{\partial\tau^2}(x; \omega_j^*, \tau^*, \eta_j^*) = \left[2\frac{\pi(x, \omega_j^*)}{(\tau^*)^3} + \frac{\pi^2(x, \omega_j^*)}{(\tau^*)^4}\right]F(x; \omega_j^*, \eta_j^*, \tau^*),$$

$$\frac{\partial^2 F}{\partial\eta\partial\eta^\top}(x; \omega_j^*, \tau^*, \eta_j^*) = \exp\left(\frac{\pi(x, \omega_j^*)}{\tau^*}\right)\mathcal{E}(x, \eta_j^*) := F_2(x; \omega_j^*, \tau^*, \eta_j^*),$$

$$\frac{\partial^2 F}{\partial\omega\partial\tau}(x; \omega_j^*, \tau^*, \eta_j^*) = \left[-\frac{1}{(\tau^*)^2}\frac{\partial\pi}{\partial\omega}(x, \omega_j^*) - \frac{\pi(x, \omega_j^*)}{(\tau^*)^3}\frac{\partial\pi}{\partial\omega}(x, \omega_j^*)\right]F(x; \omega_j^*, \tau^*, \eta_j^*),$$

$$\frac{\partial^2 F}{\partial\omega\partial\eta^\top}(x; \omega_j^*, \tau^*, \eta_j^*) = \frac{1}{\tau^*}\frac{\partial\pi}{\partial\omega}(x, \omega_j^*)[F_1(x; \omega_j^*, \tau^*, \eta_j^*)]^\top,$$

$$\frac{\partial^2 F}{\partial\tau\partial\eta}(x; \omega_j^*, \tau^*, \eta_j^*) = -\frac{\pi(x, \omega_j^*)}{(\tau^*)^2}F_1(x; \omega_j^*, \tau^*, \eta_j^*).$$

Then, we can represent $A_{n,2}(x)$ as

$$A_{n,2}(x) = \sum_{j:|\mathcal{A}_j|>1} [A_{n,1,j}(x) + A_{n,2,j}(x)] + R_2(x), \tag{61}$$

where

$$A_{n,2,j}(x) := \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)\left\{\left[(\Delta\omega_{ij}^n)^\top M_d \odot \left(\frac{1}{\tau^*}\frac{\partial^2\pi}{\partial\omega\partial\omega^\top}(x, \omega_j^*) + \frac{1}{(\tau^*)^2}\frac{\partial\pi}{\partial\omega}(x, \omega_j^*) \cdot \frac{\partial\pi}{\partial\omega^\top}(x, \omega_j^*)\right)(\Delta\omega_{ij}^n)\right.\right.$$

$$+ \left(\frac{\pi(x, \omega_j^*)}{(\tau^*)^3} + \frac{\pi^2(x, \omega_j^*)}{2(\tau^*)^4}\right)(\Delta\tau^n)^2 - (\Delta\omega_{ij}^n)^\top\left(\frac{1}{(\tau^*)^2}\frac{\partial\pi}{\partial\omega}(x, \omega_j^*) + \frac{\pi(x, \omega_j^*)}{(\tau^*)^3}\frac{\partial\pi}{\partial\omega}(x, \omega_j^*)\right)(\Delta\tau^n)\right] \cdot F(x; \omega_j^*, \tau^*, \eta_j^*)$$

$$+ \left[(\Delta\omega_{ij}^n)^\top\frac{1}{\tau^*}\frac{\partial\pi}{\partial\omega}(x, \omega_j^*)[F_1(x; \omega_j^*, \tau^*, \eta_j^*)]^\top(\Delta\eta_{ij}^n) - (\Delta\eta_{ij}^n)^\top\frac{\pi(x, \omega_j^*)}{(\tau^*)^2}F_1(x; \omega_j^*, \tau^*, \eta_j^*)(\Delta\tau^n)\right]$$

$$+ (\Delta\eta_{ij}^n)^\top M_d \odot F_2(x; \omega_j^*, \tau^*, \eta_j^*)(\Delta\eta_{ij}^n)\bigg\},$$

with M_d being an $d \times d$ matrix whose diagonal entries are $\frac{1}{2}$ while other entries are 1.

Step 1B - Decompose $B_n(x)$. Next, we apply the same strategy of decomposing $A_n(x)$ for partitioning the term $B_n(x)$. In particular, we first have

$$B_n(x) = \sum_{j:|\mathcal{A}_j|=1} \sum_{i\in\mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)\left[H(x;\omega_i^n,\tau^n) - H(x;\omega_j^*,\tau^*)\right]$$

$$+ \sum_{j:|\mathcal{A}_j|>1} \sum_{i\in\mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)\left[H(x;\omega_i^n,\tau^n) - H(x;\omega_j^*,\tau^*)\right]$$

$$:= B_{n,1}(x) + B_{n,2}(x).$$

Regarding the term $B_{n,1}(x)$, by employing the first-order Taylor expansion, we have

$$B_{n,1}(x) = \sum_{j:|\mathcal{A}_j|=1} \sum_{i\in\mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right) \sum_{|\alpha|=1} (\Delta\omega_{ij}^n)^{\alpha_1}(\Delta\tau^n)^{\alpha_2} \cdot \frac{\partial H}{\partial\omega^{\alpha_1}\partial\tau^{\alpha_2}}(x;\omega_j^*,\tau^*) + R_3(x)$$

$$= \sum_{j:|\mathcal{A}_j|=1} \sum_{i\in\mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)\left[(\Delta\omega_{ij}^n)^\top\frac{\partial\pi}{\partial\omega}(x,\omega_j^*) - (\Delta\tau^n)\frac{\pi(x,\omega_j^*)}{(\tau^*)^2}\right] \cdot H(x;\omega_j^*,\tau^*) + R_3(x)$$

$$= \sum_{j:|\mathcal{A}_j|=1} B_{n,1,j}(x) + R_3(x), \tag{62}$$

where $R_3(x)$ is a Taylor remainder such that $R_3(x)/\mathcal{L}_{4n} \to 0$ as $n\to\infty$, and

$$B_{n,1,j}(x) := \sum_{i\in\mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)\left[\frac{(\Delta\omega_{ij}^n)^\top}{\tau^*}\frac{\partial\pi}{\partial\omega}(x,\omega_j^*) - (\Delta\tau^n)\frac{\pi(x,\omega_j^*)}{(\tau^*)^2}\right] \cdot H(x;\omega_j^*,\tau^*).$$

Similarly, by means of the second-order Taylor expansion, we get

$$B_{n,2}(x) = \sum_{j:|\mathcal{A}_j|>1} \sum_{i\in\mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right) \sum_{|\alpha|=1}^{2} (\Delta\omega_{ij}^n)^{\alpha_1}(\Delta\tau^n)^{\alpha_2} \cdot \frac{\partial H}{\partial\omega^{\alpha_1}\partial\tau^{\alpha_2}}(x;\omega_j^*,\tau^*) + R_4(x)$$

$$= \sum_{j:|\mathcal{A}_j|>1} [B_{n,1,j}(x) + B_{n,2,j}(x)] + R_4(x), \tag{63}$$

where $R_4(x)$ is a Taylor remainder such that $R_4(x)/\mathcal{L}_{4n} \to 0$ as $n\to\infty$, and

$$B_{n,2,j}(x) := \sum_{i\in\mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)\left[(\Delta\omega_{ij}^n)^\top M_d \odot \left(\frac{1}{\tau^*}\frac{\partial^2\pi}{\partial\omega\partial\omega^\top}(x,\omega_j^*) + \frac{1}{(\tau^*)^2}\frac{\partial\pi}{\partial\omega}(x,\omega_j^*) \cdot \frac{\partial\pi}{\partial\omega^\top}(x,\omega_j^*)\right)(\Delta\omega_{ij}^n)\right.$$

$$\left.+ \left(\frac{\pi(x,\omega_j^*)}{(\tau^*)^3} + \frac{\pi^2(x,\omega_j^*)}{2(\tau^*)^4}\right)(\Delta\tau^n)^2 - (\Delta\omega_{ij}^n)^\top\left(\frac{1}{(\tau^*)^2}\frac{\partial\pi}{\partial\omega}(x,\omega_j^*) + \frac{\pi(x,\omega_j^*)}{(\tau^*)^3}\frac{\partial\pi}{\partial\omega}(x,\omega_j^*)\right)(\Delta\tau^n)\right] \cdot H(x;\omega_j^*).$$

Putting the above results together, we see that $[A_n(x) - R_1(x) - R_2(x)]/\mathcal{L}_{4n}$, $[B_n(x) - R_3(x) -$

$R_4(x)]/\mathcal{L}_{4n}$ and $C_n(x)/\mathcal{L}_{4n}$ can be written as a combination of elements from the following set

$$\Big\{ F_j, \pi(x, \omega_j^*) F_j, \pi^2(x, \omega_j^*) F_j, \frac{\partial \pi}{\partial \omega^{(u)}}(x, \omega_j^*) F_j, \pi(x, \omega_j^*) \frac{\partial \pi}{\partial \omega^{(u)}}(x, \omega_j^*) F_j, \frac{\partial \pi}{\partial \omega^{(u)}}(x, \omega_j^*) \frac{\partial \pi}{\partial \omega^{(v)}}(x, \omega_j^*) F_j, $$

$$\frac{\partial^2 \pi}{\partial \omega^{(u)} \partial \omega^{(v)}}(x, \omega_j^*) F_j : j \in [k_*], u, v \in [d] \Big\}$$

$$\cup \Big\{ (F_{1j})^{(u)}, \pi(x, \omega_j^*)(F_{1j})^{(u)}, \frac{\partial \pi}{\partial \omega^{(u)}}(x, \omega_j^*)(F_{1j})^{(v)} : j \in [k_*], u, v \in [d] \Big\}$$

$$\cup \Big\{ (F_{2j})^{(uv)} : j \in [k_*], u, v \in [d] \Big\}$$

$$\cup \Big\{ H_j, \pi(x, \omega_j^*) H_j, \pi^2(x, \omega_j^*) H_j, \frac{\partial \pi}{\partial \omega^{(u)}}(x, \omega_j^*) H_j, \pi(x, \omega_j^*) \frac{\partial \pi}{\partial \omega^{(u)}}(x, \omega_j^*) H_j, \frac{\partial \pi}{\partial \omega^{(u)}}(x, \omega_j^*) \frac{\partial \pi}{\partial \omega^{(v)}}(x, \omega_j^*) H_j, $$

$$\frac{\partial^2 \pi}{\partial \omega^{(u)} \partial \omega^{(v)}}(x, \omega_j^*) H_j : j \in [k_*], u, v \in [d] \Big\},$$

where we denote

$$F_j := F(x; \omega_j^*, \tau^*, \eta_j^*),$$

$$F_{1j} := \exp\Big(\frac{\pi(x, \omega_j^*)}{\tau^*} \Big) \frac{\partial \mathcal{E}}{\partial \eta}(x, \eta_j^*),$$

$$F_{2j} := \exp\Big(\frac{\pi(x, \omega_j^*)}{\tau^*} \Big) \frac{\partial^2 \mathcal{E}}{\partial \eta \partial \eta^\top}(x, \eta_j^*).$$

Step 2 - Non-vanishing coefficients. Moving to this step, we utilize the proof by contradiction method to show that not all coefficients in the representations of $[A_n - R_1(x) - R_2(x)]/\mathcal{L}_{4n}$, $[B_n - R_3(x) - R_4(x)]/\mathcal{L}_{4n}$ and $C_n(x)/\mathcal{L}_{4n}$ converge to zero as $n \to \infty$. Indeed, assume by contrary that all of them go to zero and consider the coefficients of the term

- $F(x; \omega_j^*, \tau^*, \eta_j^*)$ for $j \in [k_*]$, we get that

$$\frac{1}{\mathcal{L}_{4n}} \sum_{j=1}^{k_*} \Big| \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_i^n}{\tau^n} \Big) - \exp\Big(\frac{\beta_j^*}{\tau^*} \Big) \Big| \to 0;$$

- $\frac{\partial \pi}{\partial \omega^{(u)}}(x, \omega_j^*) F(x; \omega_j^*, \tau^*, \eta_j^*)$ for $u \in [d]$ and $j \in [k_*] : |\mathcal{A}_j| = 1$, we get that

$$\frac{1}{\mathcal{L}_{4n}} \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_i^n}{\tau^n} \Big) \|\Delta \omega_{ij}^n\|_1 \to 0.$$

Due to the equivalency between the ℓ_1-norm and the ℓ_2-norm, we deduce that

$$\frac{1}{\mathcal{L}_{4n}} \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_i^n}{\tau^n} \Big) \|\Delta \omega_{ij}^n\| \to 0.$$

- $\pi(x, \omega_j^*) F(x; \omega_j^*, \tau^*, \eta_j^*)$ for $u \in [d]$ and $j : |\mathcal{A}_j| = 1$, we get that

$$\frac{1}{\mathcal{L}_{4n}} \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp\Big(\frac{\beta_i^n}{\tau^n} \Big) \|\Delta \tau^n\| \to 0.$$

- $[F_1(x; \omega_j^*, \tau^*, \eta_j^*)]^{(u)}$ for $u \in [d]$ and $j : |\mathcal{A}_j| = 1$, we get that

$$\frac{1}{\mathcal{L}_{4n}} \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right) \|\Delta \eta_{ij}^n\| \to 0.$$

- $\frac{\partial^2 \pi}{\partial \omega^{(u)} \partial \omega^{(u)}}(x, \omega_j^*) F(x; \omega_j^*, \tau^*, \eta_j^*)$ for $u \in [d]$ and $j : |\mathcal{A}_j| > 1$, we get that

$$\frac{1}{\mathcal{L}_{4n}} \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right) \|\Delta \omega_{ij}^n\|^2 \to 0.$$

- $\pi^2(x, \omega_j^*) F(x; \omega_j^*, \tau^*, \eta_j^*)$ for $u \in [d]$ and $j : |\mathcal{A}_j| > 1$, we get that

$$\frac{1}{\mathcal{L}_{4n}} \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right) \|\Delta \tau^n\|^2 \to 0.$$

- $[F_2(x; \omega_j^*, \tau^*, \eta_j^*)]^{(uu)}$ for $u \in [d]$ and $j : |\mathcal{A}_j| > 1$, we get that

$$\frac{1}{\mathcal{L}_{4n}} \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right) \|\Delta \eta_{ij}^n\|^2 \to 0.$$

Combine all the above limits, we arrive at $1 = \mathcal{L}_{4n}/\mathcal{L}_{4n} \to 0$ as $n \to \infty$, which is a contradiction. Therefore, not all the coefficients in the representations of $[A_n(x) - R_1(x) - R_2(x)]/\mathcal{L}_{4n}$, $[B_n(x) - R_3(x) - R_4(x)]/\mathcal{L}_{4n}$ and $C_n(x)/\mathcal{L}_{4n}$ go to zero as $n \to \infty$.

Step 3 - Application of the Fatou's lemma. Now, we attempt to exhibit a contradiction to the conclusion of Step 2. Let m_n be the maximum of the absolute values of the coefficients in the representations of $[A_n(x) - R_1(x) - R_2(x)]/\mathcal{L}_{4n}$, $[B_n(x) - R_3(x) - R_4(x)]/\mathcal{L}_{4n}$ and $C_n(x)/\mathcal{L}_{4n}$. As at least one among those coefficients does not go zero, we have $1/m_n \not\to \infty$.

Recall from the hypothesis in equation (57) that $\|g_{G_n} - g_{G_*}\|_{L^2(\mu)}/\mathcal{L}_{4n} \to 0$ as $n \to \infty$. This result also implies that $\|g_{G_n} - g_{G_*}\|_{L^1(\mu)}/\mathcal{L}_{4n} \to 0$. Then, by applying the Fatou's lemma, we get

$$0 = \lim_{n \to \infty} \frac{\|g_{G_n} - g_{G_*}\|_{L^1(\mu)}}{m_n \mathcal{L}_{4n}} \geq \int \liminf_{n \to \infty} \frac{|g_{G_n}(x) - g_{G_*}(x)|}{m_n \mathcal{L}_{4n}} d\mu(x) \geq 0.$$

As a result, we have $[g_{G_n}(x) - g_{G_*}(x)]/[m_n \mathcal{L}_{4n}] \to 0$ for almost every x. Since the parameter Θ is compact, the term $\sum_{j=1}^{k_*} \exp\left(\frac{\pi(x, \omega_j^*)}{\tau^*} + \beta_j^*\right)$ is bounded. Thus, it follows that $Q_n(x)/[m_n \mathcal{L}_{4n}] \to 0$, or equivalently,

$$\frac{1}{m_n \mathcal{L}_{1n}} \cdot \Big[(A_{n,1}(x) - R_1(x) + A_{n,2}(x) - R_2(x)) - (B_{n,1}(x) - R_3(x) + B_{n,2}(x) - R_4(x)) + C_n(x) \Big] \to 0.$$

(64)

For ease of presentation, we denote

$$\frac{1}{m_n \mathcal{L}_{4n}} \cdot \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)(\Delta\omega_{ij}^n) \to \phi_{1,j}, \quad \frac{1}{m_n \mathcal{L}_{4n}} \cdot \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)(\Delta\omega_{ij}^n)(\Delta\omega_{ij}^n)^\top \to \phi_{2,j},$$

$$\frac{1}{m_n \mathcal{L}_{4n}} \cdot \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)(\Delta\tau^n) \to \chi_{1,j}, \quad \frac{1}{m_n \mathcal{L}_{4n}} \cdot \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)(\Delta\tau^n)^2 \to \chi_{2,j},$$

$$\frac{1}{m_n \mathcal{L}_{4n}} \cdot \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)(\Delta\eta_{ij}^n) \to \varphi_{1,j}, \quad \frac{1}{m_n \mathcal{L}_{4n}} \cdot \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)(\Delta\eta_{ij}^n)(\Delta\eta_{ij}^n)^\top \to \varphi_{2,j},$$

$$\frac{1}{m_n \mathcal{L}_{4n}} \cdot \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)(\Delta\omega_{ij}^n)(\Delta\tau^n) \to \zeta_{1,j}, \quad \frac{1}{m_n \mathcal{L}_{4n}} \cdot \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)(\Delta\omega_{ij}^n)(\Delta\eta_{ij}^n)^\top \to \zeta_{2,j}$$

$$\frac{1}{m_n \mathcal{L}_{4n}} \cdot \sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right)(\Delta\eta_{ij}^n)(\Delta\tau^n) \to \zeta_{3,j}, \quad \frac{1}{m_n \mathcal{L}_{4n}} \cdot \left(\sum_{i \in \mathcal{A}_j} \exp\left(\frac{\beta_i^n}{\tau^n}\right) - \exp\left(\frac{\beta_j^*}{\tau^*}\right)\right) \to \xi_j.$$

Above, at least one among the limits $\phi_{1,j}^{(u)}$, $\phi_{2,j}^{(uu)}$, $\chi_{1,j}$, $\chi_{2,j}$, $\varphi_{1,j}^{(u)}$, $\varphi_{2,j}^{(uu)}$ and ξ_j, for $j \in [k_*]$, is non-zero as a result of the conclusion in Step 2. Additionally, let us denote $F_{\rho j} := F_\rho(x; \omega_j^*, \tau^*, \eta_j^*)$ for $\rho \in \{1, 2\}$ and $H_j = H(x; \omega_j^*, \tau^*)$. From the formulation of

- $A_{n,1}(x)$ in equation (60), we get

$$\frac{A_{n,1}(x) - R_1(x)}{m_n \mathcal{L}_{4n}} \to \sum_{j:|\mathcal{A}_j|=1} \left[\left(\frac{(\phi_{1,j})^\top}{\tau^*}\frac{\partial\pi}{\partial\omega}(x, \omega_j^*) - \chi_{1,j} \cdot \frac{\pi(x, \omega_j^*)}{(\tau^*)^2}\right) \cdot F_j + \varphi_{1,j}^\top F_{1j}\right]. \quad (65)$$

- $A_{n,2}(x)$ in equation (61), we get

$$\frac{A_{n,2}(x) - R_2(x)}{m_n \mathcal{L}_{2n}} \to \sum_{j:|\mathcal{A}_j|>1} \left\{\left[\frac{(\phi_{1,j})^\top}{\tau^*}\frac{\partial\pi}{\partial\omega}(x, \omega_j^*) - \chi_{1,j} \cdot \frac{\pi(x, \omega_j^*)}{(\tau^*)^2}\right.\right.$$

$$+ M_d \odot \left(\frac{1}{\tau^*}\frac{\partial^2\pi}{\partial\omega\partial\omega^\top}(x, \omega_j^*) + \frac{1}{(\tau^*)^2}\frac{\partial\pi}{\partial\omega}(x, \omega_j^*) \cdot \frac{\partial\pi}{\partial\omega^\top}(x, \omega_j^*)\right) \odot \phi_{2,j}$$

$$+ \left(\frac{\pi(x, \omega_j^*)}{(\tau^*)^3} + \frac{\pi^2(x, \omega_j^*)}{2(\tau^*)^4}\right)\chi_{2,j} - \zeta_{1,j}^\top\left(\frac{1}{(\tau^*)^2}\frac{\partial\pi}{\partial\omega}(x, \omega_j^*) + \frac{\pi(x, \omega_j^*)}{(\tau^*)^3}\frac{\partial\pi}{\partial\omega}(x, \omega_j^*)\right)\right] \cdot F_j$$

$$+ \left[\varphi_{1,j}^\top F_{1j} + \left(\frac{1}{\tau^*}F_{1j}\frac{\partial\pi}{\partial\omega^\top}(x, \omega_j^*)\right) \odot \zeta_{2,j} - \frac{\pi(x, \omega_j^*)}{(\tau^*)^2} \cdot \zeta_{3,j}^\top F_{1j}\right] + \left(M_d \odot \varphi_{2,j}\right) \odot F_{2j}\right\}. \quad (66)$$

- $B_{n,1}(x)$ in equation (62), we get

$$\frac{B_{n,1}(x) - R_3(x)}{m_n \mathcal{L}_{2n}} \to \sum_{j:|\mathcal{A}_j|=1} \left[\frac{\phi_{1,j}^\top}{\tau^*}\frac{\partial\pi}{\partial\omega}(x, \omega_j^*) - \chi_{1,j} \cdot \frac{\pi(x, \omega_j^*)}{(\tau^*)^2}\right] \cdot H_j. \quad (67)$$

- $B_{n,2}(x)$ in equation (63), we get

$$
\begin{aligned}
\frac{B_{n,2}(x) - R_4(x)}{m_n \mathcal{L}_{2n}} \to \sum_{j:|\mathcal{A}_j|>1} \Big[& \frac{(\phi_{1,j})^\top}{\tau^*} \frac{\partial \pi}{\partial \omega}(x, \omega_j^*) - \chi_{1,j} \cdot \frac{\pi(x, \omega_j^*)}{(\tau^*)^2} \\
+ M_d \odot \Big(& \frac{1}{\tau^*} \frac{\partial^2 \pi}{\partial \omega \partial \omega^\top}(x, \omega_j^*) + \frac{1}{(\tau^*)^2} \frac{\partial \pi}{\partial \omega}(x, \omega_j^*) \cdot \frac{\partial \pi}{\partial \omega^\top}(x, \omega_j^*) \Big) \odot \phi_{2,j} \\
+ \Big(& \frac{\pi(x, \omega_j^*)}{(\tau^*)^3} + \frac{\pi^2(x, \omega_j^*)}{2(\tau^*)^4} \Big) \chi_{2,j} - \zeta_{1,j}^\top \Big(\frac{1}{(\tau^*)^2} \frac{\partial \pi}{\partial \omega}(x, \omega_j^*) + \frac{\pi(x, \omega_j^*)}{(\tau^*)^3} \frac{\partial \pi}{\partial \omega}(x, \omega_j^*) \Big) \Big] \cdot H_j. \quad (68)
\end{aligned}
$$

- $C_n(x)$ in equation (59), we get

$$
\frac{C_n(x)}{m_n \mathcal{L}_{2n}} \to \sum_{j=1}^{k_*} \xi_j [F_j - H_j]. \tag{69}
$$

Note that from equation (64), it follows that the sum of the limits in equations (65), (66), (67), (68) and (69) is equal to zero.

Subsequently, we aim to demonstrate that the values of $\phi_{1,j}^{(u)}$, $\phi_{2,j}^{(uu)}$, $\chi_{1,j}$, $\chi_{2,j}$, $\varphi_{1,j}^{(u)}$, $\varphi_{2,j}^{(uu)}$ and ξ_j are all zero for $j \in [k_*]$. Indeed, let $P_1, P_2, \ldots, P_\ell$ be the partition of the set $\Big\{ \exp\Big(\frac{\pi(x, \omega_j^*)}{\tau^*} \Big) : j \in [k_*] \Big\}$, for some $\ell \leq k_*$, such that

(i) $\omega_j^* = \omega_{j'}^*$ for any $j, j' \in P_i$ and $i \in [\ell]$;

(ii) $\omega_j^* \neq \omega_{j'}^*$ when j and j' do not belong to the same set P_i for any $i \in [\ell]$.

Then, the set $\Big\{ \exp\Big(\frac{\pi(x, \omega_{j_1}^*)}{\tau^*} \Big), \ldots, \exp\Big(\frac{\pi(x, \omega_{j_\ell}^*)}{\tau^*} \Big) \Big\}$, where $P_i \in P_i$, is linearly independent. Since

the limits in equations (65), (66), (67), (68) and (69) sum up to zero, we get for any $i \in [\ell]$ that

$$\sum_{j \in P_i : |\mathcal{A}_j| = 1} \left[\left(\xi_j + \frac{(\phi_{1,j})^\top}{\tau^*} \frac{\partial \pi}{\partial \omega}(x, \omega_j^*) - \chi_{1,j} \cdot \frac{\pi(x, \omega_j^*)}{(\tau^*)^2} \right) \cdot \mathcal{E}_j + \varphi_{1,j}^\top \mathcal{E}_{1j} \right]$$

$$+ \sum_{j \in P_i : |\mathcal{A}_j| > 1} \left\{ \left[\xi_j + \frac{(\phi_{1,j})^\top}{\tau^*} \frac{\partial \pi}{\partial \omega}(x, \omega_j^*) - \chi_{1,j} \cdot \frac{\pi(x, \omega_j^*)}{(\tau^*)^2} \right. \right.$$

$$+ M_d \odot \left(\frac{1}{\tau^*} \frac{\partial^2 \pi}{\partial \omega \partial \omega^\top}(x, \omega_j^*) + \frac{1}{(\tau^*)^2} \frac{\partial \pi}{\partial \omega}(x, \omega_j^*) \cdot \frac{\partial \pi}{\partial \omega^\top}(x, \omega_j^*) \right) \odot \phi_{2,j}$$

$$+ \left(\frac{\pi(x, \omega_j^*)}{(\tau^*)^3} + \frac{\pi^2(x, \omega_j^*)}{2(\tau^*)^4} \right) \chi_{2,j} - \zeta_{1,j}^\top \left(\frac{1}{(\tau^*)^2} \frac{\partial \pi}{\partial \omega}(x, \omega_j^*) + \frac{\pi(x, \omega_j^*)}{(\tau^*)^3} \frac{\partial \pi}{\partial \omega}(x, \omega_j^*) \right) \right] \cdot \mathcal{E}_j$$

$$+ \left[\varphi_{1,j}^\top \mathcal{E}_{1j} + \left(\frac{1}{\tau^*} \mathcal{E}_{1j} \frac{\partial \pi}{\partial \omega^\top}(x, \omega_j^*) \right) \odot \zeta_{2,j} - \frac{\pi(x, \omega_j^*)}{(\tau^*)^2} \cdot \zeta_{3,j}^\top \mathcal{E}_{1j} \right] + \left(M_d \odot \varphi_{2,j} \right) \odot \mathcal{E}_{2j} \right\}$$

$$+ \sum_{j : |\mathcal{A}_j| = 1} \left[\xi_j + \frac{\phi_{1,j}^\top}{\tau^*} \frac{\partial \pi}{\partial \omega}(x, \omega_j^*) - \chi_{1,j} \cdot \frac{\pi(x, \omega_j^*)}{(\tau^*)^2} \right] g_{G_*}(x)$$

$$+ \sum_{j : |\mathcal{A}_j| > 1} \left[\xi_j + \frac{(\phi_{1,j})^\top}{\tau^*} \frac{\partial \pi}{\partial \omega}(x, \omega_j^*) - \chi_{1,j} \cdot \frac{\pi(x, \omega_j^*)}{(\tau^*)^2} \right.$$

$$+ M_d \odot \left(\frac{1}{\tau^*} \frac{\partial^2 \pi}{\partial \omega \partial \omega^\top}(x, \omega_j^*) + \frac{1}{(\tau^*)^2} \frac{\partial \pi}{\partial \omega}(x, \omega_j^*) \cdot \frac{\partial \pi}{\partial \omega^\top}(x, \omega_j^*) \right) \odot \phi_{2,j}$$

$$+ \left(\frac{\pi(x, \omega_j^*)}{(\tau^*)^3} + \frac{\pi^2(x, \omega_j^*)}{2(\tau^*)^4} \right) \chi_{2,j} - \zeta_{1,j}^\top \left(\frac{1}{(\tau^*)^2} \frac{\partial \pi}{\partial \omega}(x, \omega_j^*) + \frac{\pi(x, \omega_j^*)}{(\tau^*)^3} \frac{\partial \pi}{\partial \omega}(x, \omega_j^*) \right) \right] g_{G_*}(x) = 0,$$

where we denote $\mathcal{E}_j := \mathcal{E}(x, \eta_j^*)$, $\mathcal{E}_{1j} := \frac{\partial \mathcal{E}}{\partial \eta}(x, \eta_j^*)$ and $\mathcal{E}_{2j} := \frac{\partial^2 \mathcal{E}}{\partial \eta \partial \eta^\top}(x, \eta_j^*)$. Recall that the pair of the expert function $\mathcal{E}$ and the router π satisfies the algebraic independence condition in Definition 2, then the following set is linearly independent for almost every x:

$$\left\{ \mathcal{E}_j, \pi(x, \omega_j^*) \mathcal{E}_j, \frac{\partial \pi}{\partial \omega^{(u)}}(x, \omega_j^*) \mathcal{E}_j, \frac{\partial \pi}{\partial \omega^{(u)}}(x, \omega_j^*) \frac{\partial \pi}{\partial \omega^{(v)}}(x, \omega_j^*) \mathcal{E}_j, \frac{\partial^2 \pi}{\partial \omega^{(u)} \partial \omega^{(v)}}(x, \omega_j^*) \mathcal{E}_j, \right.$$

$$\left. \pi(x, \omega_j^*) \frac{\partial \pi}{\partial \omega^{(u)}}(x, \omega_j^*) \mathcal{E}_j, \mathcal{E}_{1j}, \pi(x, \omega_j^*) \mathcal{E}_{1j}, \frac{\partial \pi}{\partial \omega^{(u)}}(x, \omega_j^*) \mathcal{E}_{1j}, \mathcal{E}_{2j} : j \in [k_*], 1 \le u, v \le d \right\}.$$

Thus, it follows that $\xi_j = 0$, $\phi_{1,j} = \varphi_{1,j} = 0_d$ and $\phi_{2,j} = \varphi_{2,j} = \zeta_j = \mathbf{0}_{d \times d}$ for any $j \in P_i$ and $i \in [\ell]$. This contradicts the fact that not all the values of $\phi_{1,j}^{(u)}, \phi_{2,j}^{(uu)}, \chi_{1,j}, \chi_{2,j} \varphi_{1,j}^{(u)}, \varphi_{2,j}^{(uu)}$ and ξ_j, for $j \in [k_*]$, are zero. Therefore, we obtain the local part (56), that is,

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{G}_k(\Theta) : \mathcal{L}_4(G, G_*) \le \varepsilon} \|g_G - g_{G_*}\|_{L^2(\mu)} / \mathcal{L}_4(G, G_*) > 0.$$

Hence, the proof is completed.

B.4 Proof of Theorem 5

Our main goal is to demonstrate that

$$\inf_{G \in \mathcal{G}_{k_1^*, k_2}(\Theta)} \|h_G - h_{G_*}\|_{L^2(\mu)} / \mathcal{L}_5(G, G_*) > 0. \tag{70}$$

By streamlining the same arguments as in Appendix A, it is sufficient to establish the local part of equation (70) provided below as the global part of equation (70) can be achieved similarly:

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{G}_{k_1^*, k_2}(\Theta): \mathcal{L}_5(G, G_*) \leq \varepsilon} \|h_G - h_{G_*}\|_{L^2(\mu)} / \mathcal{L}_5(G, G_*) > 0. \tag{71}$$

Assume by contrary that the above claim does not hold, then we can retrieve a sequence of mixing measures (G_n) in $\mathcal{G}_{k_1^*, k_2}(\Theta)$ such that $\mathcal{L}_{5n} := \mathcal{L}_5(G_n, G_*) \to 0$ and

$$\|h_{G_n} - h_{G_*}\|_{L^2(\mu)} / \mathcal{L}_{5n} \to 0, \tag{72}$$

as $n \to \infty$. Since the Voronoi loss $\mathcal{A}_{j_1}^n := \mathcal{A}_{j_1}(G_n)$ has only one element, we may assume WLOG that $\mathcal{A}_{j_1} = \{j_1\}$ for any $j_1 \in [k_1^*]$.

$$\begin{aligned}
\mathcal{L}_{5n} = & \sum_{j_1=1}^{k_1^*} \left| \exp(\beta_{j_1}^n) - \exp(\beta_j^*) \right| + \sum_{j_1=1}^{k_1^*} \exp(\beta_{j_1}) \|\omega_{j_1}^n - \omega_{j_1}^*\| \\
& + \sum_{j_1=1}^{k_1^*} \exp(\beta_{j_1}^n) \left[\sum_{j_2:|\mathcal{A}_{j_2|j_1}|=1} \exp(\nu_{j_2|j_1}^n)(\|\kappa_{j_2|j_1}^n - \kappa_{j_2|j_1}^*\| + \|\eta_{j_1 j_2}^n - \eta_{j_1 j_2}^*\|) \right. \\
& \left. + \sum_{j_2:|\mathcal{A}_{j_2|j_1}|>1} \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{j_2|j_1}^n)(\|\kappa_{i_2|j_1}^n - \kappa_{j_2|j_1}^*\|^2 + \|\eta_{j_1 i_2}^n - \eta_{j_1 j_2}^*\|^2) \right] \\
& + \sum_{j_1=1}^{k_1^*} \exp(\beta_{j_1}) \sum_{j_2=1}^{k_2^*} \left| \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|j_1}^n) - \exp(\nu_{j_2|j_1}^*) \right|,
\end{aligned} \tag{73}$$

Since $\mathcal{L}_{5n} \to 0$ as $n \to \infty$, we deduce that $\omega_j^n \to \omega_j^*$ and $\exp(\beta_j^n) \to \exp(\beta_j^*)$ as $n \to \infty$.

Step 1 - Decompose the difference between regression functions: In this step, we utilize the Taylor expansion to decompose the following term:

$$Q_n(x) := \left[\sum_{j_1=1}^{k_1^*} \exp((\omega_{j_1}^*)^\top x + \beta_{j_1}^*) \right] [h_{G_n}(x) - h_{G_*}(x)].$$

Let us denote

$$h_{j_1}^n(x) := \sum_{j_2=1}^{k_2^*} \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \mathrm{Softmax}((\kappa_{i_2|j_1}^n)^\top x + \nu_{i_2|j_1}^n) \mathcal{E}(x, \eta_{j_1 i_2}^n),$$

$$h_{j_1}^*(x) := \sum_{j_2=1}^{k_2^*} \mathrm{Softmax}((\kappa_{j_2|j_1}^*)^\top x + \nu_{j_2|j_1}^*) \mathcal{E}(x, \eta_{j_1 j_2}^*),$$

39

then the term $Q_n(x)$ can be decomposed into three terms as

$$Q_n(x) = \sum_{j_1=1}^{k_1^*} \exp(\beta_{j_1}^n) \left[\exp((\omega_{j_1}^n)^\top x) h_{j_1}^n(x) - \exp((\omega_{j_1}^*)^\top x) h_{j_1}^*(x) \right]$$

$$- \sum_{j_1=1}^{k_1^*} \exp(\beta_{j_1}^n) \left[\exp((\omega_{j_1}^n)^\top x) - \exp((\omega_{j_1}^*)^\top x) \right] h_{G_n}(x)$$

$$+ \sum_{j_1=1}^{k_1^*} \left(\exp(\beta_{j_1}^n) - \exp(\beta_{j_1}^*) \right) \exp((\omega_{j_1}^*)^\top x) \left[h_{j_1}^*(x) - h_{G_n}(x) \right]$$

$$:= A_n(x) - B_n(x) + C_n(x). \tag{74}$$

Step 1A - Decompose $A_n(x)$: We proceed to decompose this term as follows:

$$A_n(x) := \sum_{j_1=1}^{k_1^*} \frac{\exp(\beta_{j_1}^n)}{\sum_{j_2'=1}^{k_2^*} \exp((\kappa_{j_2'|j_1}^*)^\top x + \nu_{j_2'|j_1}^*)} [A_{n,j_1,1}(x) - A_{n,j_1,2}(x) + A_{n,j_1,3}(x)],$$

where we define

$$A_{n,j_1,1}(x) := \sum_{j_2=1}^{k_2^*} \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|j_1}^n) \left[\exp((\kappa_{i_2|j_1}^n)^\top x) \exp((\omega_{j_1}^n)^\top x) \mathcal{E}(x, \eta_i^n) \right.$$

$$\left. - \exp((\kappa_{j_2|j_1}^*)^\top x) \exp((\omega_{j_1}^*)^\top x) \mathcal{E}(x, \eta_{j_1 j_2}^*) \right],$$

$$A_{n,j_1,2}(x) := \sum_{j_2=1}^{k_2^*} \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|j_1}^n) \left[\exp((\kappa_{i_2|j_1}^n)^\top x) - \exp((\kappa_{j_2|j_1}^*)^\top x) \right] \exp((\omega_{j_1}^n)^\top x) h_{j_1}^n(x),$$

$$A_{n,j_1,3}(x) := \sum_{j_2=1}^{k_2^*} \left(\sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|j_1}^n) - \exp(\nu_{j_2|j_1}^*) \right) \exp((\kappa_{j_2|j_1}^*)^\top x)$$

$$\times [\exp((\omega_{j_1}^*)^\top x) \mathcal{E}(x, \eta_{j_1 j_2}^*) - \exp((\omega_{j_1}^n)^\top x) h_{j_1}^n(x)].$$

Step 1A.1 - Decompose $A_{n,j_1,1}(x)$: Moreover, we continue to divide the term $A_{n,j_1,1}(x)$ into two parts based on the cardinality of the Voronoi cells $\mathcal{A}_{j_2|j_1}$:

$$A_{n,j_1,1}(x) = \sum_{j_2:|\mathcal{A}_{j_2|j_1}|=1} \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|j_1}^n) \left[\exp((\kappa_{i_2|j_1}^n)^\top x) \exp((\omega_{j_1}^n)^\top x) \mathcal{E}(x, \eta_i^n) \right.$$

$$\left. - \exp((\kappa_{j_2|j_1}^*)^\top x) \exp((\omega_{j_1}^*)^\top x) \mathcal{E}(x, \eta_{j_1 j_2}^*) \right],$$

$$+ \sum_{j_2:|\mathcal{A}_{j_2|j_1}|>1} \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|j_1}^n) \left[\exp((\kappa_{i_2|j_1}^n)^\top x) \exp((\omega_{j_1}^n)^\top x) \mathcal{E}(x, \eta_i^n) \right.$$

$$\left. - \exp((\kappa_{j_2|j_1}^*)^\top x) \exp((\omega_{j_1}^*)^\top x) \mathcal{E}(x, \eta_{j_1 j_2}^*) \right]$$

$$:= A_{n,j_1,1,1}(x) + A_{n,j_1,1,2}(x).$$

By employing the first-order Taylor expansion to the term $A_{n,j_1,1,1}(x)$, we get

$$A_{n,j_1,1,1}(x) = \sum_{j_2:|\mathcal{A}_{j_2|j_1}|=1} \sum_{i_2\in\mathcal{A}_{j_2|j_1}} \sum_{|\alpha|=1} \frac{\exp(\nu_{i_2|j_1}^n)}{\alpha!} (\Delta\kappa_{i_2j_2|j_1}^n)^{\alpha_1} (\Delta\omega_{j_1}^n)^{\alpha_2} (\Delta\eta_{ij}^n)^{\alpha_3}$$

$$\times x^{\alpha_1+\alpha_2} \exp((\kappa_{j_2|j_1}^*)^\top x) \exp((\omega_{j_1}^*)^\top x) \frac{\partial^{|\alpha_3|}\mathcal{E}}{\partial\eta^{\alpha_3}}(x,\eta_{j_1j_2}^*) + R_{n,1,1}(x),$$

$$= \sum_{j_2:|\mathcal{A}_{j_2|j_1}|=1} \sum_{|\rho_1|+|\alpha_3|=1} S_{n,j_2|j_1,\rho_1,\alpha_3} \cdot x^{\rho_1} \exp((\kappa_{j_2|j_1}^*)^\top x) \exp((\omega_{j_1}^*)^\top x) \frac{\partial^{|\alpha_3|}\mathcal{E}}{\partial\eta^{\alpha_3}}(x,\eta_{j_1j_2}^*) + R_{n,1,1}(x),$$

where $R_{n,1,1}(x)$ is a Taylor remainder such that $R_{n,1,1}(x)/\mathcal{L}_{5n} \to 0$ as $n\to\infty$, and

$$S_{n,j_2|j_1,\rho_1,\alpha_3} := \sum_{i_2\in\mathcal{A}_{j_2|j_1}} \sum_{\alpha_1+\alpha_2=\rho_1} \frac{\exp(\nu_{i_2|j_1}^n)}{\alpha!} (\Delta\kappa_{i_2j_2|j_1}^n)^{\alpha_1} (\Delta\omega_{j_1}^n)^{\alpha_2} (\Delta\eta_{ij}^n)^{\alpha_3},$$

for any $j_2 \in [k_2^*]$ and $(\rho_1,\alpha_3) \neq (0_d, 0_d)$.

Similarly, we apply the second-order Taylor expansion to the term $A_{n,j,1,2}(x)$ and get that

$$A_{n,j,1,2}(x) = \sum_{j_2:|\mathcal{A}_{j_2|j_1}|>1} \sum_{|\rho_1|+|\alpha_3|=1}^{2} S_{n,j_2|j_1,\rho_1,\alpha_3} \cdot x^{\rho_1} \exp((\kappa_{j_2|j_1}^*)^\top x) \exp((\omega_{j_1}^*)^\top x) \frac{\partial^{|\alpha_3|}\mathcal{E}}{\partial\eta^{\alpha_3}}(x,\eta_{j_1j_2}^*) + R_{n,1,2}(x),$$

where $R_{n,1,2}(x)$ is a Taylor remainder such that $R_{n,1,2}(x)/\mathcal{L}_{5n} \to 0$ as $n\to\infty$.

Step 1A.2 - Decompose $A_{n,j_1,2}(x)$: Next, we also separate the quantity $A_{n,j_1,2}(x)$ into two following terms:

$$A_{n,j_1,2}(x) = \sum_{j_2:|\mathcal{A}_{j_2|j_1}|=1} \sum_{i_2\in\mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|j_1}^n) \left[\exp((\kappa_{i_2|j_1}^n)^\top x) - \exp((\kappa_{j_2|j_1}^*)^\top x) \right] \exp((\omega_{j_1}^n)^\top x) h_{j_1}^n(x)$$

$$+ \sum_{j_2:|\mathcal{A}_{j_2|j_1}|>1} \sum_{i_2\in\mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|j_1}^n) \left[\exp((\kappa_{i_2|j_1}^n)^\top x) - \exp((\kappa_{j_2|j_1}^*)^\top x) \right] \exp((\omega_{j_1}^n)^\top x) h_{j_1}^n(x)$$

$$:= A_{n,j_1,2,1}(x) + A_{n,j_1,2,2}(x).$$

By means of the first-order Taylor expansion, we have

$$A_{n,j_1,2,1}(x) = \sum_{j_2:|\mathcal{A}_{j_2|j_1}|=1} \sum_{i_2\in\mathcal{A}_{j_2|j_1}} \sum_{|\rho_2|=1} \frac{\exp(\nu_{i_2|j_1}^n)}{\rho_2!} \cdot (\Delta\kappa_{i_2j_2|j_1}^n)^{\rho_2} x^{\rho_2} \exp((\kappa_{j_2|j_1}^*)^\top x) \exp((\omega_{j_1}^n)^\top x) h_{j_1}^n(x) + R_{n,2,1}(x),$$

$$= \sum_{j_2:|\mathcal{A}_{j_2|j_1}|=1} \sum_{|\rho_2|=1} T_{n,j_2|j_1,\rho_2} \cdot x^{\rho_2} \exp((\kappa_{j_2|j_1}^*)^\top x) \exp((\omega_{j_1}^n)^\top x) h_{j_1}^n(x) + R_{n,2,1}(x),$$

where $R_{n,2,1}(x)$ is a Taylor remainder such that $R_{n,2,1}(x)/\mathcal{L}_{5n} \to 0$ as $n\to\infty$ and

$$T_{n,j_2|j_1,\rho_2} := \sum_{i_2\in\mathcal{A}_{j_2|j_1}} \frac{\exp(\nu_{i_2|j_1}^n)}{\rho_2!} \cdot (\Delta\kappa_{i_2j_2|j_1}^n)^{\rho_2},$$

41

for any $j_2 \in [k_2^*]$ and $\rho_2 \neq 0_d$.

Analogously, by means of the second-order Taylor expansion, we have

$$A_{n,j_1,2,1}(x) = \sum_{j_2:|\mathcal{A}_{j_2|j_1}|>1} \sum_{|\rho_2|=1}^{2} T_{n,j_2|j_1,\rho_2} \cdot x^{\rho_2} \exp((\kappa_{j_2|j_1}^*)^\top x) \exp((\omega_{j_1}^n)^\top x) h_{j_1}^n(x) + R_{n,2,2}(x),$$

where $R_{n,2,2}(x)$ is a Taylor remainder such that $R_{n,2,2}(x)/\mathcal{L}_{5n} \to 0$ as $n \to \infty$.

Combining the above decompositions of $A_{n,j_1,1}(x)$ and $A_{n,j_1,2}(x)$, we deduce that

$$
\begin{aligned}
A_n(x) = \sum_{j_1=1}^{k_1^*} \sum_{j_2=1}^{k_2^*} & \frac{\exp(\beta_{j_1}^n)}{\sum_{j_2'=1}^{k_2^*} \exp((\kappa_{j_2'|j_1}^*)^\top x + \nu_{j_2'|j_1}^*)} \Big[R_{n,1,1}(x) + R_{n,1,2}(x) - R_{n,2,1}(x) - R_{n,2,2}(x) \\
& + \sum_{|\rho_1|+|\alpha_3|=0}^{2} S_{n,j_2|j_1,\rho_1,\alpha_3} \cdot x^{\rho_1} \exp((\kappa_{j_2|j_1}^*)^\top x) \exp((\omega_{j_1}^*)^\top x) \frac{\partial^{|\alpha_3|} \mathcal{E}}{\partial \eta^{\alpha_3}}(x, \eta_{j_1 j_2}^*) \\
& - \sum_{|\rho_2|=0}^{2} T_{n,j_2|j_1,\rho_2} \cdot x^{\rho_2} \exp((\kappa_{j_2|j_1}^*)^\top x) \exp((\omega_{j_1}^n)^\top x) h_{j_1}^n(x) \Big],
\end{aligned}
\tag{75}
$$

where we define $S_{n,j_2|j_1,\rho_1,\alpha_3} = T_{n,j_2|j_1,\rho_2} = \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|j_1}^n) - \exp(\nu_{j_2|j_1}^*)$ for any $j_1 \in [k_1^*]$, $j_2 \in [k_2^*]$ and $(\rho_1, \alpha_3, \rho_2) = (0_d, 0_q, 0_d)$.

Step 1B - Decompose $B_n(x)$: Subsequently, by invoking the Taylor expansion of first order, we get

$$B_n(x) = \sum_{j_1=1}^{k_1^*} \exp(\beta_{j_1}^n) \sum_{|\rho_3|=1} (\Delta\omega_{j_1}^n)^{\rho_3} \cdot x^{\rho_3} \exp((\omega_{j_1}^*)^\top x) h_{G_n}(x) + R_{n,3}(x), \tag{76}$$

where $R_{n,3}(x)$ is a Taylor remainder such that $R_{n,3}(x)/\mathcal{L}_{5n} \to 0$ as $n \to \infty$.

From the decompositions in equations (75), (76) and (74), we can view $A_n(x)$, $B_n(x)$ and $C_n(x)$ as a combination of elements from the following set:

$$
\begin{aligned}
& \left\{ \frac{x^{\rho_1} \exp((\kappa_{j_2|j_1}^*)^\top x) \exp((\omega_{j_1}^*)^\top x) \frac{\partial^{|\alpha_3|} \mathcal{E}}{\partial \eta^{\alpha_3}}(x, \eta_{j_1 j_2}^*)}{\sum_{j_2'=1}^{k_2^*} \exp((\kappa_{j_2'|j_1}^*)^\top x + \nu_{j_2'|j_1}^*)} : j_1 \in [k_1^*], j_2 \in [k_2^*], 0 \leq |\rho_1| + |\alpha_3| \leq 2 \right\} \\
& \cup \left\{ \frac{x^{\rho_2} \exp((\kappa_{j_2|j_1}^*)^\top x) \exp((\omega_{j_1}^n)^\top x) h_{j_1}^n(x)}{\sum_{j_2'=1}^{k_2^*} \exp((\kappa_{j_2'|j_1}^*)^\top x + \nu_{j_2'|j_1}^*)} : j_1 \in [k_1^*], j_2 \in [k_2^*], 0 \leq |\rho_2| \leq 2 \right\} \\
& \cup \left\{ x^{\rho_3} \exp((\omega_{j_1}^*)^\top x) h_{G_n}(x) : j_1 \in [k_1^*], 0 \leq |\rho_3| \leq 1 \right\} \cup \left\{ \exp((\omega_{j_1}^*)^\top x) h_{j_1}^*(x) : j_1 \in [k_1^*] \right\}.
\end{aligned}
$$

Step 2 - Non-vanishing coefficients: In this step, we aim to demonstrate that not all the coefficients in the representations of $A_n(x)/\mathcal{L}_{5n}$, $B_n(x)/\mathcal{L}_{5n}$ and $C_n(x)/\mathcal{L}_{5n}$ converge to zero as n tends to infinity. Assume by contrary that all of them go to zero, then by considering the coefficients of the following terms

42

- $\exp((\omega_{j_1}^*)^\top x) h_{j_1}^*(x)$ in $C_n(x)/\mathcal{L}_{5n}$, we get

$$\frac{1}{\mathcal{L}_{5n}} \cdot \sum_{j_1=1}^{k_1^*} \left| \exp(\beta_{j_1}^n) - \exp(\beta_{j_1}^*) \right| \to 0;$$

- $x^{\rho_3} \exp((\omega_{j_1}^*)^\top x) h_{G_n}(x)$ in $B_n(x)/\mathcal{L}_{5n}$ for $\rho_3 = e_{d,u} := (0, \ldots, 0, \underbrace{1}_{u\text{-}th}, 0, \ldots, 0) \in \mathbb{R}^d$, we have

$$\frac{1}{\mathcal{L}_{5n}} \cdot \sum_{j_1=1}^{k_1^*} \exp(\beta_{j_1}^n) \|\omega_{j_1}^n - \omega_{j_1}^*\| \to 0;$$

- $\dfrac{x^{\rho_2} \exp((\kappa_{j_2|j_1}^*)^\top x) \exp((\omega_{j_1}^n)^\top x) h_{j_1}^n(x)}{\sum_{j_2'=1}^{k_2^*} \exp((\kappa_{j_2'|j_1}^*)^\top x + \nu_{j_2'|j_1}^*)}$ in $A_n(x)/\mathcal{L}_{5n}$ for $j_1 \in [k_1^*]$, $j_2 \in [k_2^*] : |\mathcal{A}_{j_2|j_1}| = 1$ and $\rho_2 = e_{d,u}$, we get

$$\frac{1}{\mathcal{L}_{5n}} \cdot \sum_{j_1=1}^{k_1^*} \exp(\beta_{j_1}^n) \sum_{j_2 \in [k_2^*]:|\mathcal{A}_{j_2|j_1}|=1} \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|j_1}^n) \|\kappa_{i_2|j_1}^n - \kappa_{j_2|j_1}^*\|_1 \to 0;$$

Due to the equivalency between the ℓ_1-norm and the ℓ_2-norm, we deduce that

$$\frac{1}{\mathcal{L}_{5n}} \cdot \sum_{j_1=1}^{k_1^*} \exp(\beta_{j_1}^n) \sum_{j_2 \in [k_2^*]:|\mathcal{A}_{j_2|j_1}|=1} \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|j_1}^n) \|\kappa_{i_2|j_1}^n - \kappa_{j_2|j_1}^*\| \to 0;$$

- $\dfrac{\exp((\kappa_{j_2|j_1}^*)^\top x) \exp((\omega_{j_1}^*)^\top x) \frac{\partial^{|\alpha_3|}\mathcal{E}}{\partial \eta^{\alpha_3}}(x, \eta_{j_1 j_2}^*)}{\sum_{j_2'=1}^{k_2^*} \exp((\kappa_{j_2'|j_1}^*)^\top x + \nu_{j_2'|j_1}^*)}$ in $A_n(x)/\mathcal{L}_{5n}$ for $j_1 \in [k_1^*]$, $j_2 \in [k_2^*] : |\mathcal{A}_{j_2|j_1}| = 1$ and $\alpha_3 = e_{q,u} \in \mathbb{R}^q$, we get

$$\frac{1}{\mathcal{L}_{5n}} \cdot \sum_{j_1=1}^{k_1^*} \exp(\beta_{j_1}^n) \sum_{j_2 \in [k_2^*]:|\mathcal{A}_{j_2|j_1}|=1} \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|j_1}^n) \|\eta_{j_1 i_2}^n - \eta_{j_1 j_2}^*\| \to 0;$$

- $\dfrac{x^{\rho_2} \exp((\kappa_{j_2|j_1}^*)^\top x) \exp((\omega_{j_1}^n)^\top x) h_{j_1}^n(x)}{\sum_{j_2'=1}^{k_2^*} \exp((\kappa_{j_2'|j_1}^*)^\top x + \nu_{j_2'|j_1}^*)}$ in $A_n(x)/\mathcal{L}_{5n}$ for $j_1 \in [k_1^*]$, $j_2 \in [k_2^*] : |\mathcal{A}_{j_2|j_1}| > 1$ and $\rho_2 = 2e_{d,u}$, we get

$$\frac{1}{\mathcal{L}_{5n}} \cdot \sum_{j_1=1}^{k_1^*} \exp(\beta_{j_1}^n) \sum_{j_2 \in [k_2^*]:|\mathcal{A}_{j_2|j_1}|>1} \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|j_1}^n) \|\kappa_{i_2|j_1}^n - \kappa_{j_2|j_1}^*\|^2 \to 0;$$

- $\dfrac{\exp((\kappa_{j_2|j_1}^*)^\top x) \exp((\omega_{j_1}^*)^\top x) \frac{\partial^{|\alpha_3|}\mathcal{E}}{\partial \eta^{\alpha_3}}(x, \eta_{j_1 j_2}^*)}{\sum_{j_2'=1}^{k_2^*} \exp((\kappa_{j_2'|j_1}^*)^\top x + \nu_{j_2'|j_1}^*)}$ in $A_n(x)/\mathcal{L}_{5n}$ for $j_1 \in [k_1^*]$, $j_2 \in [k_2^*] : |\mathcal{A}_{j_2|j_1}| > 1$ and $\alpha_3 = 2e_{q,u} \in \mathbb{R}^q$, we get

$$\frac{1}{\mathcal{L}_{5n}} \cdot \sum_{j_1=1}^{k_1^*} \exp(\beta_{j_1}^n) \sum_{j_2 \in [k_2^*]:|\mathcal{A}_{j_2|j_1}|>1} \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|j_1}^n) \|\eta_{j_1 i_2}^n - \eta_{j_1 j_2}^*\|^2 \to 0;$$

43

- $\dfrac{\exp((\kappa^*_{j_2|j_1})^\top x)\exp((\omega^*_{j_1})^\top x)\mathcal{E}(x,\eta^*_{j_1 j_2})}{\sum_{j'_2=1}^{k^*_2}\exp((\kappa^*_{j'_2|j_1})^\top x + \nu^*_{j'_2|j_1})}$ in $A_n(x)/\mathcal{L}_{5n}$, we get

$$\frac{1}{\mathcal{L}_{5n}}\cdot\sum_{j_1=1}^{k^*_1}\exp(\beta^n_{j_1})\sum_{j_2=1}^{k^*_2}\Big|\sum_{i_2\in\mathcal{A}_{j_2|j_1}}\exp(\nu^n_{i_2|j_1})-\exp(\nu^*_{j_2|j_1})\Big|\to 0;$$

By taking the summation of the above limits, then it follows from the formulation of the loss $\mathcal{L}_{5n}$ in equation (73) that $1 = \mathcal{L}_{5n}/\mathcal{L}_{5n}$, which is a contradiction. Hence, we obtain at least one among the coefficients in the representations of $A_n(x)/\mathcal{L}_{5n}$, $B_n(x)/\mathcal{L}_{5n}$ and $C_n(x)/\mathcal{L}_{5n}$ does not go to zero as $n\to\infty$.

Step 3 - Application of the Fatou's lemma: In this step, we point out that all the coefficients in the representations of $A_n(x)/\mathcal{L}_{5n}$, $B_n(x)/\mathcal{L}_{5n}$ and $C_n(x)/\mathcal{L}_{5n}$ converge to zero as $n\to\infty$. In particular, let us denote by m_n the maximum of the absolute values of those coefficients. From the result achieved in Step 2, we deduce that $1/m_n\not\to\infty$. Recall from the hypothesis in equation (72) that $\|h_{G_n}-h_{G_*}\|_{L^2(\mu)}/\mathcal{L}_{5n}\to 0$ as $n\to\infty$, which indicates that $\|h_{G_n}-h_{G_*}\|_{L^1(\mu)}/\mathcal{L}_{5n}\to 0$. Then, according to the Fatou's lemma, we get that

$$0 = \lim_{n\to\infty}\frac{\|h_{G_n}-h_{G_*}\|_{L^1(\mu)}}{m_n\mathcal{L}_{5n}} \geq \int\liminf_{n\to\infty}\frac{|h_{G_n}(x)-h_{G_*}(x)|}{m_n\mathcal{L}_{5n}}\mathrm{d}\mu(x)\geq 0.$$

As a result, it follows that $\frac{1}{m_n\mathcal{L}_{5n}}\cdot[h_{G_n}(x)-h_{G_*}(x)]\to 0$ as $n\to\infty$ for μ-almost surely x. From the formulation of $Q_n(x)$ in equation (74), since the quantity $\sum_{j_1=1}^{K}\exp(-\|\omega^*_{j_1}-x\|+\beta^*_{j_1})$ is bounded, we get $\frac{1}{m_n\mathcal{L}_{5n}}\cdot Q_n(x)\to 0$ for μ-almost surely x.

Let us denote

$$\frac{\exp(\beta^n_{j_1})S_{n,j_2|j_1,\rho_1,\alpha_3}}{m_n\mathcal{L}_{5n}}\to\phi_{j_2|j_1,\rho_1,\alpha_3}, \qquad\qquad \frac{\exp(\beta^n_{j_1})T_{n,j_2|j_1,\rho_2}}{m_n\mathcal{L}_{5n}}\to\varphi_{j_2|j_1,\rho_2},$$

$$\frac{\exp(\beta^n_{j_1})(\Delta\omega^n_{j_1})^{\rho_3}}{m_n\mathcal{L}_{5n}}\to\lambda_{j_1,\rho_3}, \qquad\qquad \frac{\exp(\beta^n_{j_1})-\exp(\beta^*_{j_1})}{m_n\mathcal{L}_{5n}}\to\chi_{j_1},$$

for all $j_1\in[k^*_1]$, $j_2\in[k^*_2]$, $0\leq|\rho_1|+|\alpha_3|\leq 2$, $0\leq|\rho_2|\leq 2$, $|\rho_3|=1$ with a note that at least one among them is different from zero. From the decomposition of $Q_n(x)$ in equation (74), we have

$$\lim_{n\to\infty}\frac{Q_n(x)}{m_n\mathcal{L}_{5n}} = \lim_{n\to\infty}\frac{A_n(x)}{m_n\mathcal{L}_{5n}} - \lim_{n\to\infty}\frac{B_n(x)}{m_n\mathcal{L}_{5n}} + \lim_{n\to\infty}\frac{C_n(x)}{m_n\mathcal{L}_{5n}},$$

44

where

$$\lim_{n\to\infty}\frac{A_n(x)}{m_n\mathcal{L}_{5n}} = \sum_{j_1=1}^{k_1^*}\sum_{j_2=1}^{k_2^*}\left[\sum_{|\rho_1|+|\alpha_3|=0}^{2}\phi_{j_2|j_1,\rho_1,\alpha_3}\cdot x^{\rho_1}\exp((\kappa_{j_2|j_1}^*)^\top x)\exp((\omega_{j_1}^*)^\top x)\frac{\partial^{|\alpha_3|}\mathcal{E}}{\partial\eta^{\alpha_3}}(x,\eta_{j_1j_2}^*)\right.$$

$$\left.-\sum_{|\rho_2|=0}^{2}\varphi_{j_2|j_1,\rho_2}\cdot x^{\rho_2}\exp((\kappa_{j_2|j_1}^*)^\top x)\exp((\omega_{j_1}^*)^\top x)h_{j_1}^*(x)\right]\cdot\frac{1}{\sum_{j_2'=1}^{k_2^*}\exp((\kappa_{j_2'|j_1}^*)^\top x+\nu_{j_2'|j_1}^*)},$$

$$\lim_{n\to\infty}\frac{B_n(x)}{m_n\mathcal{L}_{5n}} = \sum_{j_1=1}^{k_1^*}\sum_{|\rho_3|=1}\lambda_{j_1,\rho_3}\cdot x^{\rho_3}\exp((\omega_{j_1}^*)^\top x)h_{G_*}(x),$$

$$\lim_{n\to\infty}\frac{C_n(x)}{m_n\mathcal{L}_{5n}} = \sum_{j_1=1}^{k_1^*}\chi_{j_1}\exp((\omega_{j_1}^*)^\top x)\left[h_{j_1}^*(x)-h_{G_*}(x)\right].$$

Since the expert function $x\mapsto\mathcal{E}(x,\eta)$ meets the strong identifiability condition in Definition 1, the following set of functions in x is linearly independent:

$$\left\{\frac{x^{\rho_1}\exp((\kappa_{j_2|j_1}^*)^\top x)\exp((\omega_{j_1}^*)^\top x)\frac{\partial^{|\alpha_3|}\mathcal{E}}{\partial\eta^{\alpha_3}}(x,\eta_{j_1j_2}^*)}{\sum_{j_2'=1}^{k_2^*}\exp((\kappa_{j_2'|j_1}^*)^\top x+\nu_{j_2'|j_1}^*)}:j_1\in[k_1^*],j_2\in[k_2^*],0\le|\rho_1|+|\alpha_3|\le 2\right\}$$

$$\cup\left\{\frac{x^{\rho_2}\exp((\kappa_{j_2|j_1}^*)^\top x)\exp((\omega_{j_1}^*)^\top x)h_{j_1}^*(x)}{\sum_{j_2'=1}^{k_2^*}\exp((\kappa_{j_2'|j_1}^*)^\top x+\nu_{j_2'|j_1}^*)}:j_1\in[k_1^*],j_2\in[k_2^*],0\le|\rho_2|\le 2\right\}$$

$$\cup\left\{x^{\rho_3}\exp((\omega_{j_1}^*)^\top x)h_{G_*}(x):j_1\in[k_1^*],0\le|\rho_3|\le 1\right\}\cup\left\{\exp((\omega_{j_1}^*)^\top x)h_{j_1}^*(x):j_1\in[k_1^*]\right\}.$$

As a result, we get that $\phi_{j_2|j_1,\rho_1,\alpha_3}=\varphi_{j_2|j_1,\rho_2}=\lambda_{j_1,\rho_3}=\chi_{j_1}$ for all all $j_1\in[k_1^*]$, $j_2\in[k_2^*]$, $0\le|\rho_1|+|\alpha_3|\le 2$, $0\le|\rho_2|\le 2$, $|\rho_3|=0$, which contradicts to the fact that at least one of them is non-zero. Thus, we achieve the local part (71), that is,

$$\lim_{\varepsilon\to 0}\inf_{G\in\mathcal{G}_{k_1^*,k_2}(\Theta):\mathcal{L}_5(G,G_*)\le\varepsilon}\|h_G-h_{G_*}\|_{L^2(\mu)}/\mathcal{L}_5(G,G_*)>0,$$

Hence, the proof is completed.

B.5   Proof of Theorem 6

Firstly, we will demonstrate that the limit

$$\lim_{\varepsilon\to 0}\inf_{G\in\mathcal{G}_{k_1^*k_2}(\Theta):\mathcal{L}_{6,r}(G,G_*)\le\varepsilon}\frac{\|h_G-h_{G_*}\|_{L^2(\mu)}}{\mathcal{L}_{6,r}(G,G_*)}=0 \tag{77}$$

holds true for any $r\ge 1$. Then, by employing the same arguments for proving the equation (49) in Appendix B.1, we arrive at our desired result

$$\inf_{\overline{G}_n\in\mathcal{G}_{k_1^*k_2}(\Theta)}\sup_{G\in\mathcal{G}_{k_1^*k_2}(\Theta)\setminus\mathcal{G}_{k_1^*(k_2^*-1)}(\Theta)}\mathbb{E}_{h_G}[\mathcal{L}_{6,r}(\overline{G}_n,G)]\gtrsim n^{-1/2}.$$

45

Proof for equation (77): We need to construct a sequence of mixing measures $(G_n) \subset \mathcal{G}_{k_1^* k_2}(\Theta)$ satisfying $\mathcal{L}_{6,r}(G_n, G_*) \to 0$ and

$$\|h_{G_n} - h_{G_*}\|_{L^2(\mu)}/\mathcal{L}_{6,r}(G_n, G_*) \to 0,$$

as $n \to \infty$. For that purpose, let us take into account the mixing measure sequence defined as

$$G_n = \sum_{i_1=1}^{k_1^*} \exp(\beta_{i_1}^n) \sum_{i_2=1}^{k_2^*+1} \exp(\nu_{i_2|i_1}^n) \delta_{(\omega_{j_1}^n, \kappa_{i_2|i_1}^n, a_{i_1 i_2}^n, b_{i_1 i_2}^n)},$$

where

$\exp(\beta_{i_1}^n) = \exp(\beta_{i_1}^*)$, for any $1 \le i_1 \le k_1^*$;

$\omega_{i_1}^n = \omega_{i_1}^*$, for any $1 \le i_1 \le k_1^*$;

$\exp(\nu_{1|i_1}^n) = \exp(\nu_{2|i_1}^n) = \frac{1}{2} \exp(\nu_{1|i_1}^*) + \frac{1}{2n^{r+1}}$ and $\exp(\nu_{i_2|i_1}^n) = \exp(\nu_{i_2-1|i_1}^n)$, for any $1 \le i_1 \le k_1^*$ and $3 \le i_2 \le k_2^* + 1$;

$\kappa_{1|i_1}^n = \kappa_{2|i_1}^n = \kappa_{1|i_1}^*$ and $\kappa_{i_2|i_1}^n = \kappa_{i_2-1|i_1}^n$, for any $1 \le i_1 \le k_1^*$ and $3 \le i_2 \le k_2^* + 1$;

$a_{i_1 1}^n = a_{i_1 2}^n = a_{i_1 1}^*$ and $a_{i_1 i_2}^n = a_{i_1 (i_2-1)}^n$, for any $1 \le i_1 \le k_1^*$ and $3 \le i_2 \le k_2^* + 1$;

$b_{i_1 1}^n = b_{i_1 1}^* + \frac{1}{n}$, $b_{i_1 2}^n = b_{i_1 1}^* - \frac{1}{n}$ and $b_{i_1 i_2}^n = b_{i_1 (i_2-1)}^*$, for any $1 \le i_1 \le k_1^*$ and $3 \le i_2 \le k_2^* + 1$.

Then, the Voronoi loss $\mathcal{L}_{6,r}(G_n, G_*)$ can be rewritten as

$$\mathcal{L}_{6,r}(G_n, G_*) = \frac{1}{n^{r+1}} \sum_{j_1=1}^{k_1^*} \exp(\beta_{j_1}^*) + \sum_{j_1=1}^{k_1^*} \exp(\beta_{j_1}^*)\left(\exp(\nu_{1|j_1}^*) + \frac{1}{n^{r+1}}\right) \cdot \frac{1}{n^r} = \mathcal{O}(n^{-r}). \tag{78}$$

From the above equation, it is clear that $\mathcal{L}_{6,r}(G_n, G_*) \to 0$ as $n \to \infty$. Thus, it suffices to show that

$$\|h_{G_n} - h_{G_*}\|_{L^2(\mu)}/\mathcal{L}_{6,r}(G_n, G_*) \to 0.$$

Let us consider the quantity $Q_n(x) := \left[\sum_{j_1=1}^{k_1^*} \exp((\omega_{j_1}^*)^\top x + \beta_{j_1}^*)\right][h_{G_n}(x) - h_{G_*}(x)]$, which can be decomposed as follows:

$$Q_n(x) = \sum_{j_1=1}^{k_1^*} \exp(\beta_{j_1}^n)\left[\exp((\omega_{j_1}^n)^\top x)h_{j_1}^n(x) - \exp((\omega_{j_1}^*)^\top x)h_{j_1}^*(x)\right]$$

$$- \sum_{j_1=1}^{k_1^*} \exp(\beta_{j_1}^n)\left[\exp((\omega_{j_1}^n)^\top x) - \exp((\omega_{j_1}^*)^\top x)\right] h_{G_n}(x)$$

$$+ \sum_{j_1=1}^{k_1^*} \left(\exp(\beta_{j_1}^n) - \exp(\beta_{j_1}^*)\right) \exp((\omega_{j_1}^*)^\top x)\left[h_{j_1}^*(x) - h_{G_n}(x)\right],$$

46

where we denote

$$h_{j_1}^n(x) := \sum_{j_2=1}^{k_2^*} \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \mathrm{Softmax}((\kappa_{i_2|j_1}^n)^\top x + \nu_{i_2|j_1}^n)((a_{j_1 i_2}^n)^\top x + b_{j_1 i_2}^n),$$

$$h_{j_1}^*(x) := \sum_{j_2=1}^{k_2^*} \mathrm{Softmax}((\kappa_{j_2|j_1}^*)^\top x + \nu_{j_2|j_1}^*)((a_{j_1 j_2}^*)^\top x + b_{j_1 j_2}^*),$$

for all $j_1 \in [k_1^*]$. Recall that we have $\exp(\beta_{i_1}^n) = \exp(\beta_{i_1}^*)$ and $\omega_{i_1}^n = \omega_{i_1}^*$ for all $i_1 \in [k_1^*]$. Then, the second and third terms in the decomposition of $Q_n(x)$ become zero. Thus, we can represent $Q_n(x)$ as

$$Q_n(x) = \sum_{j_1=1}^{k_1^*} \exp(\beta_{j_1}^*) \exp((\omega_{j_1}^*)^\top x) \left[h_{j_1}^n(x) - h_{j_1}^*(x) \right]. \tag{79}$$

Note that we can continue to decompose the term $h_{j_1}^n(x) - h_{j_1}^*(x)$ as

$$\left[\sum_{j_2=1}^{k_2^*} \exp((\kappa_{j_2|j_1}^*)^\top x + \nu_{j_2|j_1}^*) \right] \cdot [h_{j_1}^n(x) - h_{j_1}^*(x)]$$

$$= \sum_{j_2=1}^{k_2^*} \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|j_1}^n) \left[\exp((\kappa_{i_2|j_1}^n)^\top x)((a_{i_2 j_1}^n)^\top x + b_{i_2 j_1}^n) - \exp((\kappa_{j_2|j_1}^*)^\top x)((a_{j_1 j_2}^*)^\top x + b_{j_1 j_2}^*) \right]$$

$$- \sum_{j_2=1}^{k_2^*} \sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|j_1}^n) \left[\exp((\kappa_{i_2|j_1}^n)^\top x) - \exp((\kappa_{j_2|j_1}^*)^\top x) \right] h_{j_1}^n(x)$$

$$+ \sum_{j_2=1}^{k_2^*} \left(\sum_{i_2 \in \mathcal{A}_{j_2|j_1}} \exp(\nu_{i_2|j_1}^n) - \exp(\nu_{j_2|j_1}^*) \right) \exp((\kappa_{j_2|j_1}^*)^\top x)[((a_{j_1 j_2}^*)^\top x + b_{j_1 j_2}^*) - h_{j_1}^n(x)]$$

$$:= A_{n,j_1}(x) - B_{n,j_1}(x) + E_{n,j_1}(x)$$

From the definitions of $\kappa_{i_2|i_1}^n$, $a_{i_1 i_2}^n$ and $b_{i_1 i_2}^n$, we can rewrite $A_{n,j_1}(x)$ as follows:

$$A_{n,j_1}(x) = \sum_{i_2=1}^{2} \frac{1}{2} \exp(\nu_{1|j_1}^n) \exp((\kappa_{1|j_1}^*)^\top x)(b_{j_1 i_2}^n - b_{j_1 1}^*)$$

$$= \frac{1}{2} \exp(\nu_{1|j_1}^n) \exp((\kappa_{1|j_1}^*)^\top x)[(b_{j_1 1}^n - b_{j_1 1}^*) + (b_{j_1 2}^n - b_{j_1 1}^*)]$$

$$= 0.$$

Next, since $\kappa_{i_2|j_1}^n = \kappa_{j_2|j_1}^*$, we can see that $B_{n,j_1}(x) = 0$. Lastly, it can be justified that $E_{n,j_1}(x) = \mathcal{O}(n^{-(r+1)})$, leading to $E_{n,j_1}(x)/\mathcal{L}_{6,r}(G_n, G_*) \to 0$. Combined these results, we deduce that

$$\left[\sum_{j_2=1}^{k_2^*} \exp((\kappa_{j_2|j_1}^*)^\top x + \nu_{j_2|j_1}^*) \right] \cdot [h_{j_1}^n(x) - h_{j_1}^*(x)]/\mathcal{L}_{6,r}(G_n, G_*) \to 0,$$

as $n \to \infty$. Since the term $\sum_{j_2=1}^{k_2^*} \exp((\kappa_{j_2|j_1}^*)^\top x + \nu_{j_2|j_1}^*)$ is bounded, it follows that $[h_{j_1}^n(x) - h_{j_1}^*(x)]/\mathcal{L}_{6,r}(G_n, G_*) \to 0$ for almost every x. Putting this limit and equation (79) together, we get that $Q_n(x)/\mathcal{L}_{6,r}(G_n, G_*) \to 0$ as $n \to \infty$ for almost every x. Again, since the term $\sum_{j_1=1}^{k_1^*} \exp((\omega_{j_1}^*)^\top x + \beta_{j_1}^*)$ is bounded, we deduce that $[h_{G_n}(x) - h_{G_*}(x)]/\mathcal{L}_{6,r}(G_n, G_*) \to 0$ as $n \to \infty$ for almost every x. This result implies that

$$\|h_{G_n} - h_{G_*}\|_{L^2(\mu)}/\mathcal{L}_{6,r}(G_n, G_*) \to 0,$$

as $n \to \infty$. Hence, the proof of claim (77) is completed.

C   Auxiliary Results

C.1   Proof of Proposition 1

For the proof of the theorem, we first introduce some necessary concepts and notations. Firstly, we denote by $\mathcal{F}_k(\Theta)$ the set of regression functions w.r.t all mixing measures in $\mathcal{G}_k(\Theta)$, that is, $\mathcal{F}_k(\Theta) := \{f_G(x) : G \in \mathcal{G}_k(\Theta)\}$. Additionally, for each $\delta > 0$, the L^2 ball centered around the regression function $f_{G_*}(x)$ and intersected with the set $\mathcal{F}_k(\Theta)$ is defined as

$$\mathcal{F}_k(\Theta, \delta) := \left\{ f \in \mathcal{F}_k(\Theta) : \|f - f_{G_*}\|_{L^2(\mu)} \leq \delta \right\}.$$

In order to measure the size of the above set, [44] suggest using the following quantity:

$$\mathcal{J}_B(\delta, \mathcal{F}_k(\Theta, \delta)) := \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(t, \mathcal{F}_k(\Theta, t), \|\cdot\|_{L^2(\mu)}) \, \mathrm{d}t \vee \delta, \tag{80}$$

where $H_B(t, \mathcal{F}_k(\Theta, t), \|\cdot\|_{L^2(\mu)})$ stands for the bracketing entropy [44] of $\mathcal{F}_k(\Theta, u)$ under the L^2-norm, and $t \vee \delta := \max\{t, \delta\}$. By using the similar proof argument of Theorem 7.4 and Theorem 9.2 in [44] with notations being adapted to this work, we obtain the following lemma:

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

$$\mathbb{P}\left(\|f_{\widehat{G}_n} - f_{G_*}\|_{L^2(\mu)} > \delta\right) \leq c \exp\left(-\frac{n\delta^2}{\nu^2}\right),$$

for all $\delta \geq \delta_n$.

Proof overview. In this proof, we first demonstrate that the following bound holds for any $0 < \varepsilon \leq 1/2$:

$$H_B(\varepsilon, \mathcal{F}_k(\Theta), \|.\|_{L^2(\mu)}) \lesssim \log(1/\varepsilon), \tag{81}$$

Then, it follows that

$$\mathcal{J}_B(\delta, \mathcal{F}_k(\Theta, \delta)) = \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(t, \mathcal{F}_k(\Theta, t), \|\cdot\|_{L^2(\mu)}) \, \mathrm{d}t \vee \delta \lesssim \int_{\delta^2/2^{13}}^{\delta} \log(1/t) dt \vee \delta. \tag{82}$$

Let $\Psi(\delta) = \delta \cdot [\log(1/\delta)]^{1/2}$, then $\Psi(\delta)/\delta^2$ is a non-increasing function of δ. Furthermore, equation (82) indicates that $\Psi(\delta) \geq \mathcal{J}_B(\delta, \mathcal{F}_k(\Theta, \delta))$. In addition, let $\delta_n = \sqrt{\log(n)/n}$, then we get that $\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n)$ for some universal constant c. Finally, by applying Lemma 1, we achieve the desired conclusion of Proposition 1, that is,

$$\|f_{\widehat{G}_n} - f_{G_*}\|_{L^2(\mu)} = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}}).$$

Proof of the bound (81). Let $\zeta \leq \varepsilon$ and $\{\pi_1, \ldots, \pi_N\}$ be the ζ-cover under the L^∞-norm of the set $\mathcal{F}_k(\Theta)$ where $N := N(\zeta, \mathcal{F}_k(\Theta), \|\cdot\|_\infty)$ is the η-covering number of the metric space $(\mathcal{F}_k(\Theta), \|\cdot\|_\infty)$. Next, since the expert function $\mathcal{E}(\cdot, \eta)$ is bounded, the regression function $f_G(\cdot)$ is also bounded, that is, $f_G(x) \leq M$ for all x, where $M > 0$ is the bounded constant of the expert function. Then, we construct the brackets of the form $[L_i(x), U_i(x)]$ for all $i \in [N]$ as follows:

$$L_i(x) := \max\{\pi_i(x) - \zeta, 0\},$$
$$U_i(x) := \max\{\pi_i(x) + \zeta, M\}.$$

From the above construction, we can validate that $\mathcal{F}_k(\Theta) \subset \cup_{i=1}^N [L_i(x), U_i(x)]$ and $U_i(x) - L_i(x) \leq \min\{2\zeta, M\}$. Therefore, it follows that

$$\|U_i - L_i\|_{L^2(\mu)} = \left(\int (U_i - L_i)^2 \mathrm{d}\mu(x)\right)^{1/2} \leq \left(\int 4\zeta^2 \mathrm{d}\mu(x)\right)^{1/2} = 2\zeta.$$

By definition of the bracketing entropy, we deduce that

$$H_B(2\zeta, \mathcal{F}_k(\Theta), \|\cdot\|_{L^2(\mu)}) \leq \log N = \log N(\zeta, \mathcal{F}_k(\Theta), \|\cdot\|_\infty). \tag{83}$$

Therefore, we need to provide an upper bound for the covering number N. In particular, we denote $\Delta := \{(\beta, \omega) \in \mathbb{R} \times \mathbb{R}^d : (\beta, \omega, \eta) \in \Theta\}$ and $\Omega := \{\eta \in \mathbb{R}^q : (\beta, \omega, \eta) \in \Theta\}$. Since Θ is a compact set, Δ and Ω are also compact. Therefore, we can find ζ-covers Δ_ζ and Ω_ζ for Δ and Ω, respectively. We can check that

$$|\Delta_\zeta| \leq \mathcal{O}(\zeta^{-(d+1)k}), \quad |\Omega_\zeta| \leq \mathcal{O}(\zeta^{-qk}).$$

For each mixing measure $G = \sum_{i=1}^k \exp(\beta_i)\delta_{(\omega_i, \eta_i)} \in \mathcal{G}_k(\Theta)$, we consider other two mixing measures:

$$\widetilde{G} := \sum_{i=1}^k \exp(\beta_i)\delta_{(\beta_{1i}, \overline{\eta}_i)}, \qquad \overline{G} := \sum_{i=1}^k \exp(\overline{\beta}_i)\delta_{(\overline{\beta}_{1i}, \overline{\eta}_i)}.$$

Here, $\overline{\eta}_i \in \Omega_\zeta$ such that $\overline{\eta}_i$ is the closest to η_i in that set, while $(\overline{\beta}_i, \overline{\omega}_i) \in \Delta_\zeta$ is the closest to (β_i, ω_i)

49

in that set. From the above formulations, we get that

$$
\begin{aligned}
\|f_G - f_{\widetilde{G}}\|_\infty &= \sup_{x \in \mathcal{X}} \left| \sum_{i=1}^{k} \frac{\exp((\omega_i)^\top x + \beta_i)}{\sum_{j=1}^{k} \exp((\omega_j)^\top x + \beta_j)} \cdot [\mathcal{E}(x, \eta_i) - \mathcal{E}(x, \overline{\eta}_i)] \right| \\
&\leq \sum_{i=1}^{k} \sup_{x \in \mathcal{X}} \frac{\exp((\omega_i)^\top x + \beta_i)}{\sum_{j=1}^{k} \exp((\omega_j)^\top x + \beta_j)} \cdot |\mathcal{E}(x, \eta_i) - \mathcal{E}(x, \overline{\eta}_i)| \\
&\leq \sum_{i=1}^{k} \sup_{x \in \mathcal{X}} |\mathcal{E}(x, \eta_i) - \mathcal{E}(x, \overline{\eta}_i)| \\
&\lesssim \sum_{i=1}^{k} \sup_{x \in \mathcal{X}} \|\eta_i - \overline{\eta}_i\| \\
&\lesssim \zeta.
\end{aligned}
$$

Here, the first inequality is according to the triangle inequality, the second inequality occurs as the softmax weight is bounded by one, and the fourth inequality follows from the fact that the expert function $\mathcal{E}(x, \cdot)$ is a Lipschitz function. Next, we have

$$
\begin{aligned}
\|f_{\widetilde{G}} - f_{\overline{G}}\|_\infty &= \sup_{x \in \mathcal{X}} \left| \sum_{i=1}^{k} \left(\frac{\exp((\omega_i)^\top x + \beta_i)}{\sum_{j=1}^{k} \exp((\omega_j)^\top x + \beta_j)} - \frac{\exp((\overline{\omega}_i)^\top x + \overline{\beta}_i)}{\sum_{j=1}^{k} \exp((\overline{\omega}_j)^\top x + \overline{\beta}_j)} \right) \cdot \mathcal{E}(x, \overline{\eta}_i) \right| \\
&\leq \sum_{i=1}^{k} \sup_{x \in \mathcal{X}} \left| \frac{\exp((\omega_i)^\top x + \beta_i)}{\sum_{j=1}^{k} \exp((\omega_j)^\top x + \beta_j)} - \frac{\exp((\overline{\omega}_i)^\top x + \overline{\beta}_i)}{\sum_{j=1}^{k} \exp((\overline{\omega}_j)^\top x + \overline{\beta}_j)} \right| \cdot |\mathcal{E}(x, \overline{\eta}_i)| \\
&\leq \sum_{i=1}^{k} \sup_{x \in \mathcal{X}} \left| \frac{\exp((\omega_i)^\top x + \beta_i)}{\sum_{j=1}^{k} \exp((\omega_j)^\top x + \beta_j)} - \frac{\exp((\overline{\omega}_i)^\top x + \overline{\beta}_i)}{\sum_{j=1}^{k} \exp((\overline{\omega}_j)^\top x + \overline{\beta}_j)} \right| \\
&\lesssim \sum_{i=1}^{k} \sup_{x \in \mathcal{X}} \left(\|\omega_i - \overline{\omega}_i\| \cdot \|x\| + |\beta_i - \overline{\beta}_i| \right) \\
&\lesssim \sum_{i=1}^{k} \sup_{x \in \mathcal{X}} (\zeta \cdot B + \zeta) \\
&\lesssim \zeta.
\end{aligned}
$$

Above, the second inequality occurs as the expert $\mathcal{E}(x, \overline{\eta}_i)$ is bounded. The third inequality happens since the softmax is a Lipschitz function. According to the triangle inequality, we have

$$
\|f_G - f_{\overline{G}}\|_\infty \leq \|f_G - f_{\widetilde{G}}\|_\infty + \|f_{\widetilde{G}} - f_{\overline{G}}\|_\infty \lesssim \zeta.
$$

By definition of the covering number, we deduce that

$$
N(\zeta, \mathcal{F}_k(\Theta), \|\cdot\|_\infty) \leq |\Delta_\zeta| \times |\Omega_\zeta| \leq \mathcal{O}(n^{-(d+1)k}) \times \mathcal{O}(n^{-qk}) \leq \mathcal{O}(n^{-(d+1+q)k}). \tag{84}
$$

Combine equations (83) and (84), we achieve that

$$
H_B(2\zeta, \mathcal{F}_k(\Theta), \|\cdot\|_{L^2(\mu)}) \lesssim \log(1/\zeta).
$$

Let $\zeta = \varepsilon/2$, then we obtain that

$$H_B(\varepsilon, \mathcal{F}_k(\Theta), \|.\|_{L^2(\mu)}) \lesssim \log(1/\varepsilon).$$

Hence, the proof is completed.

C.2 Proof of Proposition 3

In this proof, we will adapt notions from the empirical process theory in Appendix C.1 to the HMoE setting and employ some arguments in that appendix. In particular, we denote $\mathcal{F}_{k_1^* k_2}(\Theta)$ as the set of HMoE-based regression functions w.r.t all mixing measures in $\mathcal{G}_{k_1^* k_2}(\Theta)$, that is, $\mathcal{F}_{k_1^* k_2}(\Theta) := \{h_G(x) : G \in \mathcal{G}_{k_1^* k_2}(\Theta)\}$. Next, an L^2 ball centered around the regression function $h_{G_*}(x)$ and intersected with the set $\mathcal{F}_{k_1^* k_2}(\Theta)$ is given by

$$\mathcal{F}_{k_1^* k_2}(\Theta, \delta) := \left\{ h \in \mathcal{F}_{k_1^* k_2}(\Theta) : \|h - h_{G_*}\|_{L^2(\mu)} \leq \delta \right\}.$$

Lastly, we denote by $H_B(t, \mathcal{F}_{k_1^* k_2}(\Theta, t), \| \cdot \|_{L^2(\mu)})$ the bracketing entropy [44] of $\mathcal{F}_{k_1^* k_2}(\Theta, u)$ under the L^2-norm. By arguing in a similar fashion to Appendix C.1, we can show that in order to reach the desired conclusion, it suffices to demonstrate that

$$H_B(\varepsilon, \mathcal{F}_{k_1^* k_2}(\Theta), \|.\|_{L^2(\mu)}) \lesssim \log(1/\varepsilon), \tag{85}$$

for any $0 < \varepsilon \leq 1/2$. Let $\zeta \leq \varepsilon$ and $\{\pi_1, \ldots, \pi_N\}$ be the ζ-cover under the L^∞-norm of the set $\mathcal{F}_{k_1^* k_2}(\Theta)$ where $N := N(\zeta, \mathcal{F}_{k_1^* k_2}(\Theta), \| \cdot \|_\infty)$ is the ζ-covering number of the metric space $(\mathcal{F}_{k_1^* k_2}(\Theta), \| \cdot \|_\infty)$. By employing the same arguments in Appendix C.1, we obtain that

$$H_B(2\zeta, \mathcal{F}_{k_1^* k_2}(\Theta), \| \cdot \|_{L^2(\mu)}) \leq \log N = \log N(\zeta, \mathcal{F}_{k_1^* k_2}(\Theta), \| \cdot \|_\infty). \tag{86}$$

Thus, it is necessary to bound the covering number N. For that purpose, let us denote $\Delta := \{(\beta, \omega) \in \mathbb{R} \times \mathbb{R}^d : (\beta, \omega, \nu, \kappa, \eta) \in \Theta\}$ and $\Omega := \{(\nu, \kappa, \eta) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^q : (\beta, \omega, \nu, \kappa, \eta) \in \Theta\}$. Recall that the parameter space Θ is compact, then the sets Δ and Ω are also compact. Therefore, there exist ζ-covers for Δ and Ω, which will be denoted as Δ_ζ and Ω_ζ, respectively. Furthermore, it can be justified that

$$|\Delta_\zeta| \leq \mathcal{O}(\zeta^{-(d+1)k_1^*}), \quad |\Omega_\zeta| \leq \mathcal{O}(\zeta^{-(d+q+1)k_1^* k_2}).$$

For a mixing measure $G = \sum_{i_1=1}^{k_1^*} \exp(\beta_{i_1}) \sum_{i_2=1}^{k_2} \exp(\nu_{i_2|i_1}) \delta_{(\omega_{i_1}, \kappa_{i_2|i_1}, \eta_{i_1 i_2})} \in \mathcal{G}_{k_1^* k_2}(\Theta)$, we take into account two additional mixing measures defined as

$$\widetilde{G} := \sum_{i_1=1}^{k_1^*} \exp(\beta_{i_1}) \sum_{i_2=1}^{k_2} \exp(\bar{\nu}_{i_2|i_1}) \delta_{(\omega_{i_1}, \bar{\kappa}_{i_2|i_1}, \bar{\eta}_{i_1 i_2})},$$

$$\overline{G} := \sum_{i_1=1}^{k_1^*} \exp(\bar{\beta}_{i_1}) \sum_{i_2=1}^{k_2} \exp(\bar{\nu}_{i_2|i_1}) \delta_{(\bar{\omega}_{i_1}, \bar{\kappa}_{i_2|i_1}, \bar{\eta}_{i_1 i_2})}.$$

51

Above, $(\bar{\nu}_{i_2|i_1}, \bar{\kappa}_{i_2|i_1}, \bar{\eta}_{i_1i_2}) \in \Omega_\zeta$ such that $(\bar{\nu}_{i_2|i_1}, \bar{\kappa}_{i_2|i_1}, \bar{\eta}_{i_1i_2})$ is the closest to $(\nu_{i_2|i_1}, \kappa_{i_2|i_1}, \eta_{i_1i_2})$ in that set, while $(\bar{\beta}_{i_1}, \bar{\omega}_{i_1}) \in \Delta_\zeta$ is the closest to $(\beta_{i_1}, \omega_{i_1})$ in that set. Additionally, we also denote

$$h_{i_1}(x) := \sum_{i_2=1}^{k_2} \mathrm{Softmax}((\kappa_{i_2|i_1})^\top x + \nu_{i_2|i_1})\mathcal{E}(x, \eta_{i_1i_2}),$$

$$\tilde{h}_{i_1}(x) := \sum_{i_2=1}^{k_2} \mathrm{Softmax}((\kappa_{i_2|i_1})^\top x + \nu_{i_2|i_1})\mathcal{E}(x, \bar{\eta}_{i_1i_2}),$$

$$\bar{h}_{i_1}(x) := \sum_{i_2=1}^{k_2} \mathrm{Softmax}((\bar{\kappa}_{i_2|i_1})^\top x + \bar{\nu}_{i_2|i_1})\mathcal{E}(x, \bar{\eta}_{i_1i_2}),$$

for all $i_1 \in [k_1^*]$. Now, we start providing an upper bound for the term $\|h_G - h_{\widetilde{G}}\|_\infty$ as

$$\|h_G - h_{\widetilde{G}}\|_\infty = \sum_{i_1=1}^{k_1^*} \mathrm{Softmax}((\omega_{i_1})^\top x + \beta_{i_1}) \cdot \|h_{i_1} - \bar{h}_{i_1}\|_\infty \leq \sum_{i_1=1}^{k_1^*} \|h_{i_1} - \bar{h}_{i_1}\|_\infty$$

$$\leq \sum_{i_1=1}^{k_1^*} \left(\|h_{i_1} - \tilde{h}_{i_1}\|_\infty + \|\tilde{h}_{i_1} - \bar{h}_{i_1}\|_\infty \right). \tag{87}$$

The first term in the above right hand side can be bounded as

$$\|h_{i_1} - \tilde{h}_{i_1}\|_\infty \leq \sum_{i_2=1}^{k_2} \sup_{x \in \mathcal{X}} \left| \mathrm{Softmax}((\kappa_{i_2|i_1})^\top x + \nu_{i_2|i_1}) \cdot [\mathcal{E}(x, \eta_{i_1i_2}) - \mathcal{E}(x, \bar{\eta}_{i_1i_2})] \right|$$

$$\leq \sum_{i_2=1}^{k_2} \sup_{x \in \mathcal{X}} \left| \mathcal{E}(x, \eta_{i_1i_2}) - \mathcal{E}(x, \bar{\eta}_{i_1i_2}) \right|$$

$$\lesssim \sum_{i_2=1}^{k_2} \sup_{x \in \mathcal{X}} \left(\|\eta_{i_1i_2} - \bar{\eta}_{i_1i_2}\| \cdot \|x\| \right)$$

$$\leq \sum_{i_2=1}^{k_2} \sup_{x \in \mathcal{X}} (\zeta \cdot B) \lesssim \zeta, \tag{88}$$

where the second last inequality occurs as the input space is bounded, that is, $\|x\| \leq B$ for all x for some constant $B > 0$.

Next, the second term in the right hand side of equation (87) is bounded as

$$\|\tilde{h}_{i_1} - \bar{h}_{i_1}\|_\infty \leq \sum_{i_2=1}^{k_2} \sup_{x \in \mathcal{X}} \left| \mathrm{Softmax}((\kappa_{i_2|i_1})^\top x + \nu_{i_2|i_1}) - \mathrm{Softmax}((\bar{\kappa}_{i_2|i_1})^\top x + \bar{\nu}_{i_2|i_1}) \right| \cdot |\mathcal{E}(x, \bar{\eta}_{i_1 i_2})|$$

$$\lesssim \sum_{i_2=1}^{k_2} \sup_{x \in \mathcal{X}} \left| \mathrm{Softmax}((\kappa_{i_2|i_1})^\top x + \nu_{i_2|i_1}) - \mathrm{Softmax}((\bar{\kappa}_{i_2|i_1})^\top x + \bar{\nu}_{i_2|i_1}) \right|$$

$$\lesssim \sum_{i_2=1}^{k_2} \sup_{x \in \mathcal{X}} \left(\|\kappa_{i_2|i_1} - \bar{\kappa}_{i_2|i_1}\| \cdot \|x\| + |\nu_{i_2|i_1} - \bar{\nu}_{i_2|i_1}| \right)$$

$$\lesssim \sum_{i_2=1}^{k_2} \sup_{x \in \mathcal{X}} (\zeta \cdot B + \zeta) \lesssim \zeta. \tag{89}$$

From equations (87), (88) and (89), we deduce that $\|h_G - h_{\widetilde{G}}\|_\infty \lesssim \zeta$. Furthermore, we also have that

$$\|h_{\widetilde{G}} - h_{\overline{G}}\|_\infty \leq \sum_{i_1=1}^{k_1^*} \sup_{x \in \mathcal{X}} \left| \mathrm{Softmax}((\omega_{i_1})^\top x + \beta_{i_1}) - \mathrm{Softmax}((\bar{\omega}_{i_1})^\top x + \bar{\beta}_{i_1}) \right| \cdot |\bar{h}_{i_1}(x)|$$

$$\lesssim \sum_{i_1=1}^{k_1^*} \sup_{x \in \mathcal{X}} \left| \mathrm{Softmax}((\omega_{i_1})^\top x + \beta_{i_1}) - \mathrm{Softmax}((\bar{\omega}_{i_1})^\top x + \bar{\beta}_{i_1}) \right|$$

$$\lesssim \sum_{i=1}^{k} \sup_{x \in \mathcal{X}} \left(\|\omega_{i_1} - \overline{\omega}_{i_1}\| \cdot \|x\| + |\beta_{i_1} - \overline{\beta}_{i_1}| \right)$$

$$\lesssim \sum_{i=1}^{k} \sup_{x \in \mathcal{X}} (\zeta \cdot B + \zeta) \lesssim \zeta.$$

Then, by the triangle inequality, we have

$$\|h_G - h_{\overline{G}}\|_\infty \leq \|h_G - h_{\widetilde{G}}\|_\infty + \|h_{\widetilde{G}} - h_{\overline{G}}\|_\infty \lesssim \zeta.$$

By definition of the covering number, we deduce that

$$N(\zeta, \mathcal{F}_{k_1^* k_2}(\Theta), \|\cdot\|_\infty) \leq |\Delta_\zeta| \times |\Omega_\zeta| \leq \mathcal{O}(n^{-(d+1)k}) \times \mathcal{O}(n^{-(d+q+1)k}) \leq \mathcal{O}(n^{-(2d+2+q)k}). \tag{90}$$

From equations (86) and (90), we achieve that

$$H_B(2\zeta, \mathcal{F}_{k_1^* k_2}(\Theta), \|\cdot\|_{L^2(\mu)}) \lesssim \log(1/\zeta).$$

By setting $\zeta = \varepsilon/2$, we obtain that

$$H_B(\varepsilon, \mathcal{F}_{k_1^* k_2}(\Theta), \|.\|_{L^2(\mu)}) \lesssim \log(1/\varepsilon).$$

Hence, the proof is completed.

C.3 Identifiability of the Softmax Gating MoE

Proposition 4. *Suppose that the equation $f_G(x) = f_{G_*}(x)$ holds for almost every x, then we obtain $G \equiv G'$.*

Proof of Proposition 4. To start with, let us expand the equation $f_G(x) = f_{G_*}(x)$ as

$$\sum_{i=1}^{k} \text{Softmax}\left((\omega_i)^\top x + \beta_i\right) \cdot \mathcal{E}(x, \eta_i) = \sum_{i=1}^{k_*} \text{Softmax}\left((\omega_i^*)^\top x + \beta_i^*\right) \cdot \mathcal{E}(x, \eta_i^*). \tag{91}$$

Since the expert function $x \mapsto \mathcal{E}(x, \eta)$ is strongly identifiable, the set of functions in x given by $\{\mathcal{E}(x, \eta_i') : i \in [k']\}$, where $\eta_1', \eta_2', \ldots, \eta_{k'}'$ are distinct parameters for $k' \in \mathbb{N}$, is linearly independent. Thus, if k is different from k_*, then we can find $i \in [k]$ such that $\eta_i \neq \eta_j^*$ for any $j \in [k_*]$, implying that $\text{Softmax}((\omega_i)^\top x + \beta_i) = 0$, which is a contradiction as the softmax value cannot be zero. Therefore, it must hold that $k = k_*$ and $\left\{ \text{Softmax}\left((\omega_i)^\top x + \beta_i\right) : i \in [k] \right\} = \left\{ \text{Softmax}\left((\omega_i^*)^\top x + \beta_i^*\right) : i \in [k_*] \right\}$, for almost every x. Without loss of generality, we assume that

$$\text{Softmax}\left((\omega_i)^\top x + \beta_i\right) = \text{Softmax}\left((\omega_i^*)^\top x + \beta_i^*\right), \tag{92}$$

for almost every x for any $i \in [k_*]$. Due to the invariance to translations of the softmax function, equation (92) implies that $\omega_i = \omega_i^* + u_1$ and $\beta_i = \beta_i^* + u_0$ for some $u_1 \in \mathbb{R}^d$ and $u_0 \in \mathbb{R}$. Recall from the assumption (A.3), since $\omega_k = \omega_k^* = 0_d$ and $\beta_k = \beta_k^* = 0$, we get $u_1 = 0_d$ and $u_0 = 0$, leading to $\omega_i = \omega_i^*$ and $\beta_i = \beta_i^*$ for all $i \in [k_*]$. Given these results, we can rewrite equation (91) as

$$\sum_{i=1}^{k_*} \exp(\beta_i) \exp((\omega_i)^\top x) \mathcal{E}(x, \eta_i) = \sum_{i=1}^{k_*} \exp(\beta_i^*) \exp((\omega_i^*)^\top x) \mathcal{E}(x, \eta_i^*), \tag{93}$$

for almost every x. Subsequently, let $P_1, P_2, \ldots, P_m$ be a partition of the index set $[k_*]$, where $m \leq k$, such that

(i) $\exp(\beta_i) = \exp(\beta_{i'}^*)$ for any $i, i' \in P_j$ and $j \in [k_*]$;

(ii) $\exp(\beta_i) \neq \exp(\beta_{i'})$ if i and i' are not in the same set P_j.

Based on the above partition, we rewrite equation (93) as

$$\sum_{j=1}^{m} \sum_{i \in P_j} \exp(\beta_i) \exp\left((\omega_i)^\top x\right) \mathcal{E}(x, \eta_i) = \sum_{j=1}^{m} \sum_{i \in P_j} \exp(\beta_i^*) \exp\left((\omega_i^*)^\top x\right) \mathcal{E}(x, \eta_i^*),$$

for almost every x. Since $\omega_i = \omega_i^*$ and $\beta_i = \beta_i^*$ for al $i \in [k_*]$, we have $\{\eta_i : i \in P_j\} \equiv \{\eta_i^* : i \in P_j\}$, for almost every x for any $j \in [m]$. Therefore, it follows that

$$G = \sum_{j=1}^{m} \sum_{i \in P_j} \exp(\beta_i) \delta_{(\omega_i, \eta_i)} = \sum_{j=1}^{m} \sum_{i \in P_j} \exp(\beta_i) \delta_{(\omega_i^*, \eta_i^*)} = G_*.$$

Hence, we reach the conclusion of this proposition. □

References

[1] Y. Bengio, N. Léonard, and A. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arxiv preprint arxiv 1308.3432*, 2013. (Cited on page 2.)

[2] 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 2.)

[3] T. Chen, X. Chen, X. Du, A. Rashwan, F. Yang, H. Chen, Z. Wang, and Y. Li. Adamv-moe: Adaptive multi-task vision mixture-of-experts. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. (Cited on page 2.)

[4] 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 page 1.)

[5] 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 3.)

[6] K. Cho and Y. Bengio. Exponentially increasing the capacity-to-computation ratio for conditional computation in deep learning. *arxiv preprint arxiv 1406.7362*, 2014. (Cited on page 2.)

[7] 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 2.)

[8] R. Csordás, K. Irie, and J. Schmidhuber. Approximating two-layer feedforward networks for efficient transformers. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 674–692, Singapore, Dec. 2023. Association for Computational Linguistics. (Cited on page 3.)

[9] 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, 4, and 19.)

[10] 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.)

[11] 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.)

[12] S. Gupta, S. Mukherjee, K. Subudhi, E. Gonzalez, D. Jose, A. Awadallah, and J. Gao. Sparsely activated mixture-of-experts are robust multi-task learners. *arXiv preprint arxiv 2204.0768*, 2022. (Cited on page 1.)

[13] 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 2.)

[14] H. Hazimeh, Z. Zhao, A. Chowdhery, M. Sathiamoorthy, Y. Chen, R. Mazumder, L. Hong, and E. Chi. DSelect-k: Differentiable Selection in the Mixture of Experts with Applications to Multi-Task Learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29335–29347. Curran Associates, Inc., 2021. (Cited on page 1.)

[15] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arxiv 1606.08415*, 2023. (Cited on page 7.)

[16] N. Ho, C.-Y. Yang, and M. I. Jordan. Convergence rates for Gaussian mixtures of experts. *Journal of Machine Learning Research*, 23(323):1–81, 2022. (Cited on pages 3 and 7.)

[17] O. Irsoy and E. Alpaydın. Dropout regularization in hierarchical mixture of experts. *Neurocomputing*, 419:148–156, 2021. (Cited on page 3.)

[18] 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.)

[19] E. Jeremiah, L. Marshall, S. A. Sisson, and A. Sharma. Specifying a hierarchical mixture of experts for hydrologic modeling: Gating function variable selection. *Water Resources Research*, 49(5):2926–2939, 2013. (Cited on page 3.)

[20] 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 pages 2 and 19.)

[21] W. Jiang and M. A. Tanner. Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *The Annals of Statistics*, 27(3):987 – 1011, 1999. (Cited on page 3.)

[22] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994. (Cited on page 3.)

[23] B. Li, Y. Shen, J. Yang, Y. Wang, J. Ren, T. Che, J. Zhang, and Z. Liu. Sparse mixture-of-experts are domain generalizable learners. In *The Eleventh International Conference on Learning Representations*, 2023. (Cited on page 2.)

[24] H. Liang, Z. Fan, R. Sarkar, Z. Jiang, T. Chen, K. Zou, Y. Cheng, C. Hao, and Z. Wang. M^3ViT: Mixture-of-Experts Vision Transformer for Efficient Multi-task Learning with Model-Accelerator Co-design. In *NeurIPS*, 2022. (Cited on page 2.)

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

[26] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. (Cited on page 2.)

[27] 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 8.)

[28] E. F. Mendes and W. Jiang. Convergence rates for mixture-of-experts. *arXiv preprint arxiv 1110.2058*, 2011. (Cited on page 3.)

[29] E. Moges, Y. Demissie, and H.-Y. Li. Hierarchical mixture of experts and diagnostic modeling approach to reduce hydrologic model structural uncertainty. *Water Resources Research*, 52(4):2551–2570, 2016. (Cited on page 3.)

[30] H. Nguyen, P. Akbarian, and N. Ho. Is temperature sample efficient for softmax Gaussian mixture of experts? In *Proceedings of the ICML*, 2024. (Cited on page 3.)

[31] 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 pages 1 and 7.)

[32] H. Nguyen, P. Akbarian, T. Pham, T. Nguyen, S. Zhang, and N. Ho. Statistical advantages of perturbing cosine router in mixture of experts. *arXiv preprint arXiv:2405.14131*, 2024. (Cited on page 2.)

[33] H. Nguyen, X. Han, C. W. Harris, S. Saria, and N. Ho. On expert estimation in hierarchical mixture of experts: Beyond softmax gating functions. *arxiv preprint arxiv 2410.02935*, 2024. (Cited on page 3.)

[34] H. Nguyen, N. Ho, and A. Rinaldo. On least square estimation in softmax gating mixture of experts. In *Proceedings of the ICML*, 2024. (Cited on pages 3 and 5.)

[35] 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 3, 7, and 10.)

[36] X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370 – 400, 2013. (Cited on page 7.)

[37] X. Nie, X. Miao, S. Cao, L. Ma, Q. Liu, J. Xue, Y. Miao, Y. Liu, Z. Yang, and B. Cui. Evomoe: An evolutional mixture-of-experts training framework via dense-to-sparse gate. *arXiv preprint arXiv:2112.14397*, 2022. (Cited on page 2.)

[38] J. Oldfield, M. Georgopoulos, G. G. Chrysos, C. Tzelepis, Y. Panagakis, M. A. Nicolaou, J. Deng, and I. Patras. Multilinear mixture of experts: Scalable expert specialization through factorization. In *Advances in Neural Information Processing Systems*, 2024. (Cited on page 4.)

[39] F. Peng, R. Jacobs, and M. Tanner. Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association*, 91(435):953–960, Sept. 1996. (Cited on page 1.)

[40] Y. Qi, J. Klein-Seetharaman, and Z. Bar-Joseph. A mixture of feature experts approach for protein-protein interaction prediction. *BMC Bioinformatics*, 8:S6 – S6, 2007. (Cited on page 1.)

[41] 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 2.)

[42] 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.)

[43] 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.)

[44] S. van de Geer. *Empirical processes in M-estimation*. Cambridge University Press, 2000. (Cited on pages 19, 48, and 51.)

[45] 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 2.)

[46] S. Waterhouse and A. Robinson. Classification using hierarchical mixtures of experts. In *Proceedings of IEEE Workshop on Neural Networks for Signal Processing*, pages 177–186, 1994. (Cited on page 1.)

[47] Z. You, S. Feng, D. Su, and D. Yu. Speechmoe: Scaling to large acoustic models with dynamic routing mixture of experts. In *Interspeech*, 2021. (Cited on page 1.)

[48] B. Yu. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*, pages 423–435, 1997. (Cited on page 28.)

[49] 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 2.)

[50] 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 3.)