On Least Square Estimation in Softmax Gating Mixture of Experts

Huy Nguyen [1] **Nhat Ho** [1 *] **Alessandro Rinaldo** [1 *]

Abstract

Mixture of experts (MoE) model is a statistical machine learning design that aggregates multiple expert networks using a softmax gating function in order to form a more intricate and expressive model. Despite being commonly used in several applications owing to their scalability, the mathematical and statistical properties of MoE models are complex and difficult to analyze. As a result, previous theoretical works have primarily focused on probabilistic MoE models by imposing the impractical assumption that the data are generated from a Gaussian MoE model. In this work, we investigate the performance of the least squares estimators (LSE) under a deterministic MoE model where the data are sampled according to a regression model, a setting that has remained largely unexplored. We establish a condition called strong identifiability to characterize the convergence behavior of various types of expert functions. We demonstrate that the rates for estimating strongly identifiable experts, namely the widely used feed-forward networks with activation functions $\mathrm{sigmoid}(\cdot)$ and $\tanh(\cdot)$, are substantially faster than those of polynomial experts, which we show to exhibit a surprising slow estimation rate. Our findings have important practical implications for expert selection.

1. Introduction

Softmax gating mixture of experts (MoE) is introduced by (Jacobs et al., 1991; Jordan & Jacobs, 1994) as a generalization of classical mixture models (McLachlan & Basford, 1988; Lindsay, 1995) based on an adaptive gating mechanism. More concretely, the MoE model is a weighted sum of expert functions associated with input-dependent weights. Here, each expert is either a regression function (De Veaux, 1989; Faria & Soromenho, 2010) or a classifier (Chen et al., 2022; Nguyen et al., 2023a) that specializes in smaller parts of a larger problem. Meanwhile, the softmax gate is responsible for determining the weight of each expert's output. If one expert consistently outperforms others in some domains of the input space, the softmax gate will assign it a larger weight in those domains. Thanks to its flexibility and adaptability, there has been a surge of interest in using the softmax gating MoE models in several fields, namely large language models (Jiang et al., 2024; Puigcerver et al., 2024; Zhou et al., 2023; Du et al., 2022; Fedus et al., 2022b), computer vision (Riquelme et al., 2021; Liang et al., 2022; Ruiz et al., 2021), multi-task learning (Hazimeh et al., 2021; Gupta et al., 2022) and reinforcement learning (Chow et al., 2023). In those applications, each expert plays an essential role in handling one or a few subproblems. As a consequence, it is of practical importance to study the problem of expert estimation, which can be solved indirectly via the parameter estimation problem.

Despite its widespread use in practice, the theory for parameter estimation of the MoE model has not been fully comprehended. From a probabilistic perspective, (Ho et al., 2022) studied the convergence of maximum likelihood estimation under an input-independent gating Gaussian MoE, which admits the following set-up:

Set-up of a Gaussian MoE model. An i.i.d sample $(X_1, Y_1), \dots, (X_n, Y_n)$ are assumed to be drawn from a softmax gating Gaussian MoE model whose conditional density function $p_{G_*}(y|x)$ is of the form

$$\sum_{i=1}^{k_*} \frac{\exp((\beta_{1i}^*)^\top x + \beta_{0i}^*)}{\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top x + \beta_{0j}^*)} \cdot \pi(y|h(x, \eta_i^*), \nu_i^*),$$

where $\pi(\cdot|\mu, \nu)$ denotes a Gaussian density function with mean μ and variance ν, and $h(\cdot, \eta)$ stands for a mean expert function. Additionally, $G_* := \sum_{i=1}^{k_*} \exp(\beta_{0i}^*)\delta_{(\beta_{1i}^*, \eta_i^*, \nu_i^*)}$ stands for the *mixing measure,* a weighted sum of Dirac measures δ, with unknown parameters $(\beta_{0i}^*, \beta_{1i}^*, \eta_i^*, \nu_i^*)$.

By assuming that the data were generated from that model, they demonstrated that the density estimation rate was parametric on the sample size, while the parameter estimation rates depended on the algebraic independence between expert functions. Subsequently, (Nguyen et al., 2023b) and (Nguyen et al., 2024b) also considered the Gaussian MoE

[*] Co-last authors [1] Department of Statistics and Data Sciences, The University of Texas at Austin, USA. Correspondence to: Huy Nguyen <huynm@utexas.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

models but equipped with a softmax gate and a Gaussian gate, respectively, both of which vary with the input values. Owing to an interaction among gating and expert parameters, they showed that the rates for estimating parameters were determined by the solvability of some systems of polynomial equations. Additionally, (Makkuva et al., 2019) also designed provably consistent algorithms for learning parameters of the softmax gating Gaussian MoE. Next, (Nguyen et al., 2024a) investigated a Top-K sparse gating Gaussian MoE model (Shazeer et al., 2017; Fedus et al., 2022a), which activated only one or a few experts for each input. Their findings suggested that turning on exactly one expert per input would remove the interaction of gating parameters with those of experts, and therefore, accelerate the parameter estimation rates.

While the theoretical advances in MoE modeling from recent years have been remarkable, a persistent and significant limitation of all existing contributions in the literature is the reliance on the strong assumption of a well-specified model, namely that the data are sampled from a (say, Gaussian) MoE model. This is of course, an unrealistic assumption that does not reflect real-world data (Li et al., 2023; Pham et al., 2024). Unfortunately, very little is known about the statistical properties of MoE models in mis-specified but more realistic regression settings.

In this paper, we partially address this gap by introducing and analyzing a more general regression framework for MoE models in which, conditionally on the features, the response variables are not sampled from a gated MoE but are instead noisy realization of an unknown and deterministic gated MoE-type regression function, as described next.

Set-up. We assume that an i.i.d. sample of size n: $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ in $\mathbb{R}^d \times \mathbb{R}$ is generated according to the model

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

where $\varepsilon_1, \dots, \varepsilon_n$ are independent Gaussian noise variables such that $\mathbb{E}[\varepsilon_i | X_i] = 0$ and $\mathrm{Var}(\varepsilon_i | X_i) = \nu$ for all $1 \leq i \leq n$. Note that, the Gaussian assumption is just for the simplicity of proof argument. Furthermore, we assume that $X_1, \dots, X_n$ are i.i.d. samples from some probability distribution μ. Above, the regression function $f_{G_*}(\cdot)$ takes the form of a softmax gating MoE with k_* experts, namely

$$f_{G_*}(x) := \sum_{i=1}^{k_*} \frac{\exp((\beta_{1i}^*)^\top x + \beta_{0i}^*)}{\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top x + \beta_{0j}^*)} \cdot h(x, \eta_i^*), \tag{2}$$

where $(\beta_{0i}^*, \beta_{1i}^*, \eta_i^*)_{i=1}^{k_*}$ are unknown parameters in $\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^q$ and $G_* := \sum_{i=1}^{k_*} \exp(\beta_{0i}^*) \delta_{(\beta_{1i}^*, \eta_i^*)}$ denotes the associated *mixing measure*. The function $h(x, \eta)$ is known as *the expert function,* which we assumed to be of parametric form. We will consider general expert functions as well as the widely used ridge expert functions $h(x; (a, b)) = \sigma(a^\top x + b)$, compositions of a non-linear *activation function* $\sigma(\cdot)$ with an affine function. See Section 2 below for further restrictions on the model. In practice, since the true number of experts k_* is unknown, it is customary to fit a softmax gating MoE model of the form (2) with up to $k > k_*$ experts, where k is a given threshold. We call this setting an *over-specified* setting.

In order to estimate the ground-truth parameters $(\beta_{0i}^*, \beta_{1i}^*, \eta_i^*)_{i=1}^{k_*}$ in the above model, we can no longer rely on maximum likelihood estimation. Instead we will deploy the computationally efficient and popular least squares method (see, e.g., van de Geer, 2000). Formally, the mixing measure is estimated with

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

where $\mathcal{G}_k(\Theta) := \{G = \sum_{i=1}^{k'} \exp(\beta_{0i}) \delta_{(\beta_{1i}, \eta_i)} : 1 \leq k' \leq k,\ (\beta_{0i}, \beta_{1i}, \eta_i) \in \Theta\}$ is the set of all mixing measures with at most k components. The goal of this paper is to investigate the convergence properties of estimator $\widehat{G}_n$ in fixed-dimensional setting. To the best of our knowledge, this is the first statistical analysis of the least squares estimation under the MoE models, as previous works (Mendes & Jiang, 2011; Nguyen et al., 2023b) focus on maximum likelihood methods.

Challenges. We highlight two subtle major challenges in analyzing the regression model (2), which require the formulation of novel identifiability conditions and new techniques. To the best of our knowledge, these issues have not been noted before in the regression literature.

(C.1) Expert characterization. In our analysis (which conforms to the latest approaches to MoE modeling), we represent the discrepancy $f_{\widehat{G}_n}(\cdot) - f_{G_*}(\cdot)$ between the estimated and true regression function as a weighted sum of linearly independent terms by applying Taylor expansions to the function $x \mapsto F(x; \beta_1, \eta) := \exp(\beta_1^\top x) h(x, \eta)$. In order to guarantee good convergence rates, it is necessary that the function F and its derivatives are linearly independent (in the space of squared-integrable functions of the features X). This property will be ensured by formulating novel and non-trivial algebraic condition on the expert functions, which we refer to as *strong identifiability*. The derivation of that condition requires us to adopt new proof techniques since those in previous works (Nguyen et al., 2023b; 2024a) apply only for linear experts.

(C.2) Singularity of polynomial experts. An instance of expert functions that does not satisfy the strong identifiability condition is a polynomial of an affine function. For simplicity, let us consider $h(x, \eta) = a^\top x + b$, where $\eta = (a, b)$.

Then, the function F mentioned in the challenge (C.1) becomes $F(x;\beta_1,a,b)=\exp(\beta_1^\top x)(a^\top x+b)$. Under this seemingly unproblematic settings, we encounter an unexpected phenomenon. Specifically, there exists an interaction between the gating parameter β_1 and the expert parameters a,b, captured by the partial differential equation (PDE)

$$\frac{\partial^2 F}{\partial\beta_1\partial b}(x;\beta_{1i}^*,a_i^*,b_i^*)=\frac{\partial F}{\partial a}(x;\beta_{1i}^*,a_i^*,b_i^*). \quad (4)$$

Complex functional interactions of this form are not new – they have been thoroughly characterized in the softmax gating Gaussian MoE model by (Nguyen et al., 2023b). However, and contrary to the case of data drawn from a well-specified softmax gating Gaussian MoE model, in our setting the above interaction causes the estimation rate of all the parameters β_{1i}^*,a_i^*,b_i^* to be slower than any polynomial rates, and thus, could potentially be $\mathcal{O}_P(1/\log(n))$. It is important to note that this singular, rather surprising, phenomenon takes place as we consider a deterministic MoE model instead of a probabilistic one, which requires us to develop new techniques.

Overall contributions. Our contributions are three-fold and can be summarized as follows (see also Table 1 for a summary of the expert estimation rates):

1. Parametric rate for regression function. In our first main result, Theorem 2.1, we demonstrate a parametric estimation rate for the regression function $f_{G_*}(\cdot)$. In particular, we show that $\|f_{\widehat{G}_n}-f_{G_*}\|_{L^2(\mu)}=\mathcal{O}_P(n^{-1/2})$, where $\|\cdot\|_{L^2(\mu)}$ denotes the L^2 norm with respect to the probability measure μ of the input X. This result will be leveraged to obtain more complex estimation rates for the model parameters.

2. Strongly identifiable experts. We formulate a general strong identifiability condition for expert functions in Definition 3.1 which ensures a faster, even parametric, estimation rates for the model parameters. To that effect, we propose a novel loss function $\mathcal{D}_1$ among parameters in equation (7) and establish in Theorem 3.2 the L^2-lower bound $\|f_G-f_{G_*}\|_{L^2(\mu)}\gtrsim\mathcal{D}_1(G,G_*)$ for any $G\in\mathcal{G}_k(\Theta)$. Given the bound $\|f_{\widehat{G}_n}-f_{G_*}\|_{L^2(\mu)}=\mathcal{O}_P(n^{-1/2})$ in Theorem 2.1, we deduce that the convergence rate of the LSE $\widehat{G}_n$ to the true mixing measure G_* is also parametric on the sample size, i.e. $\mathcal{D}_1(\widehat{G}_n,G_*)=\mathcal{O}_P(n^{-1/2})$. This leads to an expert estimation rate of order at least $\mathcal{O}_P(n^{-1/4})$.

3. Ridge experts: Secondly, we focus ridge expert functions consisting of simple two-layer neural networks, which include a linear layer followed by an activation layer, i.e., $h(x,\eta)=\sigma(a^\top x+b)$, where $\eta=(a,b)$. In these very common settings, we give a condition called *strong independence* in Definition 4.1 to characterize activation functions that induce faster expert estimation rates. Interestingly, under the strongly independent settings of the activation function σ, we demonstrate in Theorem 4.4 that even when the activation function σ is strongly independent, the expert estimation rates are still slower than any polynomial rates and could be as slow as $\mathcal{O}_P(1/\log(n))$ if at lease one among parameters $a_1^*,\ldots,a_{k_*}^*$ vanishes. Otherwise, we show in Theorem 4.2 that the expert estimation rates are no worse than $\mathcal{O}_P(n^{-1/4})$.

Lastly, we consider the settings when the activation function σ is not strongly independent, e.g., polynomial experts of the form $h(x,\eta)=(a^\top x+b)^p$, where $p\in\mathbb{N}$ and $\eta=(a,b)$ (of which linear experts are special cases). This choice can be regarded as an ridge expert associated with the activation function $\sigma(z)=z^p$, which violates the strong independence condition. As a consequence, we come across an unforeseen phenomenon in Theorem 4.6: the rates for estimating experts become universally worse than any polynomial rates due to an intrinsic interaction between gating and expert parameters via the PDE (4).

Practical implications. There are two main practical implications from our theoretical results:

(i) Expert network design. Firstly, based on the the strong identifiability condition provide in Definition 3.1, we can verify that plenty of widely used expert functions, namely feed-forward networks with activation functions $\mathrm{sigmoid}(\cdot)$, $\tanh(\cdot)$ and $\mathrm{GELU}(\cdot)$, are strongly identifiable. Therefore, our findings suggest that such experts enjoy faster estimation rates than others. This indicates that our theory is potentially useful for designing experts in practical applications.

(ii) Sample inefficiency of polynomial experts. Secondly, Theorem 4.6 reveals that a class of polynomial experts, including linear experts, are not good choices of expert functions for MoE models due to its significantly slow estimation rates. This observation aligns with the findings in (Chen et al., 2022) which claims that a mixture of non-linear experts achieves a way better performance than a mixture of linear experts.

Outline. The paper is organized as follows. In Section 2, we obtain a parametric rate for the least squares estimation of softmax gating MoE model $f_{G_*}(\cdot)$ under the L^2-norm. Subsequently, we establish estimation rates for experts that satisfy the strong identifiability condition in Section 3. We then investigate ridge experts, including polynomial experts in Section 4. Finally, we conclude the paper and provide some future directions in Section 5. Rigorous proofs and a simulation study are deferred to the supplementary material.

Notations. We let $[n]$ stand for the set $\{1,2,\ldots,n\}$ for any $n\in\mathbb{N}$. Next, for any set S, we denote $|S|$ as its cardinality. 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 $\|v\|$ denotes its 2-norm value. Lastly, for any two positive

Table 1. Summary of expert estimation rates (up to a logarithmic factor) under the softmax gating mixture of experts. In this work, we analyze three types of expert functions including strongly identifiable experts $h(x,\eta)$, ridge experts $\sigma(a^\top x+b)$ and polynomial experts $(a^\top x+b)^p$. For ridge experts, we consider two complement regimes: all the experts are input-dependent (Regime 1) vs. there exists an input-independent expert (Regime 2). Additionally, the notation $\mathcal{A}_j(\widehat{G}_n)$ stands for the Voronoi cels defined in equation (6).

Expert Index	**Strongly-Identifiable Experts**	**Ridge Experts with Strongly Independent Activation**		**Polynomial Experts**
		Regime 1	Regime 2	
$j:\lvert\mathcal{A}_j(\widehat{G}_n)\rvert=1$	$\mathcal{O}_P(n^{-1/2})$		Slower than $\mathcal{O}_P(n^{-1/2r}),\forall r\geq 1$	
$j:\lvert\mathcal{A}_j(\widehat{G}_n)\rvert>1$	$\mathcal{O}_P(n^{-1/4})$		Slower than $\mathcal{O}_P(n^{-1/2r}),\forall r\geq 1$	

sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, we write $a_n=\mathcal{O}(b_n)$ or $a_n\lesssim b_n$ if $a_n\leq Cb_n$ for all $n\in\mathbb{N}$, where $C>0$ is some universal constant. The notation $a_n=\mathcal{O}_P(b_n)$ indicates that a_n/b_n is stochastically bounded.

2. The Estimation Rate for the Regression Function

In this section, we establish an important result, showing that, under minimal assumptions on the regression function, the least squares plug-in estimator of the regression function $f_{\widehat{G}_n}(\cdot)$ is consistent, and converges to the true regression function $f_{G_*}(\cdot)$ at the rate $1/\sqrt{n}$ with respect to the $L^2(\mu)$-distance, where μ is the feature distribution.

Assumptions. Throughout the paper, we impose the following standard assumptions on the model parameters. We recall that the dimension of the parameter space is fixed.

(A.1) We assume that the parameter space Θ is a compact subset of $\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}^q$, while the input space $\mathcal{X}$ is bounded. These assumptions help guarantee the convergence of least squares estimation.

(A.2) For the experts $h(x,\eta_1^*),\dots,h(x,\eta_{k_*}^*)$ being different from each other, we assume that parameters $\eta_1^*,\dots,\eta_{k_*}^*$ are pair-wise distinct. Furthermore, these experts functions are Lipschitz continuous with respect to their parameters and bounded.

(A.3) In order that the softmax gating MoE $f_{G_*}(\cdot)$ is identifiable, i.e., $f_G(x)=f_{G_*}(x)$ for almost every x implies that $G\equiv G_*$, we let $\beta_{0k_*}^*=0$ and $\beta_{1k_*}^*=0_d$.

(A.4) To ensure that the softmax gate is input-dependent, we assume that at least one among gating parameters $\beta_{11}^*,\dots,\beta_{1k_*}^*$ is non-zero.

Theorem 2.1. *Given a least squares estimator $\widehat{G}_n$ defined in equation (3), the model estimation $f_{\widehat{G}_n}$ admits the following convergence rate:*

$$\|f_{\widehat{G}_n}-f_{G_*}\|_{L^2(\mu)}=\mathcal{O}_P(\sqrt{\log(n)/n}). \tag{5}$$

The proof of Theorem 2.1 is in Appendix A.1. It can be seen from the bound (5) that the rate for estimating the entire softmax gating MoE model $f_{G_*}(\cdot)$ is of order $\mathcal{O}_P(n^{-1/2})$ (up to logarithmic factor), which is parametric on the sample size n. More importantly, this result suggests that if we can construct a loss function among parameters $\mathcal{D}$ such that $\|f_{\widehat{G}_n}-f_{G_*}\|_{L^2(\mu)}\gtrsim\mathcal{D}(\widehat{G}_n,G_*)$, then it follows that $\mathcal{D}(\widehat{G}_n,G_*)=\mathcal{O}_P(n^{-1/2})$. As a consequence, we achieve parameter estimation rates through the previous bound, and therefore, our desired expert estimation rates.

3. Strongly Identifiable Experts

In this section, we derive estimation rates for the parameters of the softmax gating MoE regression function (2) assuming that the class of expert functions satisfy a novel regularity condition which we refer to as *strong identifiability*; see Definition 3.1 below.

Let us recall that in order to establish the expert estimation rates, our approach is to establish the L^2-lower bound $\|f_{\widehat{G}_n}-f_{G_*}\|_{L^2(\mu)}\gtrsim\mathcal{D}(\widehat{G}_n,G_*)$ mentioned in Section 2, where $\mathcal{D}$ is an appropriate loss function to be defined later. For that purpose, a key step is to decompose the quantity $f_{\widehat{G}_n}(x)-f_{G_*}(x)$ into a combination of linearly independent terms, where

$$f_{G_*}(x):=\sum_{i=1}^{k_*}\frac{\exp((\beta_{1i}^*)^\top x+\beta_{0i}^*)}{\sum_{j=1}^{k_*}\exp((\beta_{1j}^*)^\top x+\beta_{0j}^*)}\cdot h(x,\eta_i^*).$$

This can be done by using Taylor expansions to the product of a softmax numerator and an expert denoted by $x\mapsto F(x;\beta_1,\eta)=\exp(\beta_1^\top x)h(x,\eta)$. Therefore, to obtain our desired decomposition, we present in the following definition a condition that ensures the derivatives of F with respect to its parameters are linearly independent.

Definition 3.1 (Strong Identifiability)**.** We say that an expert function $x\mapsto h(x,\eta)$ is strongly identifiable if it is twice differentiable with respect to its parameter η and the following set of functions in x is linearly independent:

$$\Big\{x^{\nu} \cdot \frac{\partial^{|\tau|} h}{\partial \eta^{\tau}}(x, \eta_j) : j \in [k], \nu \in \mathbb{N}^d,\ \tau \in \mathbb{N}^q,\ 0 \leq |\nu| + |\tau| \leq 2\Big\},$$

for almost every x for any $k \geq 1$ and pair-wise distinct parameters $\eta_1, \ldots, \eta_k$.

As indicated in Definition 3.1, the main distinction between the strong identifiability and standard identifiability conditions of the expert function h (Ho et al., 2022) is that we further require the first and second-order derivatives of the expert function h with respect to their parameter are also linearly independent. 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).

Example. It can be checked that the strong identifiability condition holds for several experts used in practice, including feed-forward networks with activations like $\mathrm{sigmoid}$, $\tanh$ and GeLU and non-linear transformed input are strongly identifiable experts. For simplicity, let us consider a 2-layer expert network with normalized input, i.e.

$$h(x, (a, b)) = \sigma\Big(a\frac{x}{\|x\|} + b\Big),$$

where σ is the sigmoid function and $x, a, b, \in \mathbb{R}$ (The argument also holds with layer normalization as well). Then, by taking the derivatives of the expert function $h(\cdot, (a, b))$ w.r.t its parameters up to the second order, we can verify that the set mentioned in Definition 3.1 is linearly independent, which means that the expert $h(\cdot, (a, b))$ is strongly identifiable. The non-linear transformation is involved to ensure the linearly independence between the terms $\frac{\partial h}{\partial a}$ and $x\frac{\partial h}{\partial b}$ mentioned in Definition 3.1. Otherwise, the strong identifiability condition is not satisfied. For instance, the ridge expert $h(x, (a, b)) = \sigma(ax + b)$ is not strongly identifiable due to the PDE $\frac{\partial h}{\partial a} = x\frac{\partial h}{\partial b}$.

Next, to compute the expert estimation rates, we propose a loss function based on the notion of Voronoi cells, put forward by (Manole & Ho, 2022), as follows.

Voronoi loss. Given an arbitrary mixing measure G with $k' \leq k$ components, we partition its components to the following Voronoi cells $\mathcal{A}_j \equiv \mathcal{A}_j(G)$, which are generated by the components of G_*:

$$\mathcal{A}_j := \{i \in [k'] : \|\omega_i - \omega_j^*\| \leq \|\omega_i - \omega_\ell^*\|, \forall \ell \neq j\}, \quad (6)$$

where $\omega_i := (\beta_{1i}, \eta_i)$ and $\omega_j^* := (\beta_{1j}^*, \eta_j^*)$ for any $j \in [k_*]$. Notably, the cardinality of Voronoi cell $\mathcal{A}_j$ is exactly the number of fitted components that approximates ω_j^*. Then, the Voronoi loss function used for our analysis is given by:

$$\mathcal{D}_1(G, G_*) := \sum_{j=1}^{k_*} \Big| \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}) - \exp(\beta_{0j}^*)\Big| + \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i})\Big[\|\Delta\beta_{1ij}\|^2 + \|\Delta\eta_{ij}\|^2\Big] + \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i})\Big[\|\Delta\beta_{1ij}\| + \|\Delta\eta_{ij}\|\Big], \quad (7)$$

where we denote $\Delta\beta_{1ij} := \beta_{1i} - \beta_{1j}^*$ and $\Delta\eta_{ij} := \eta_i - \eta_j^*$. Above, if the Voronoi cell $\mathcal{A}_j$ is empty, then we let the corresponding summation term be zero. Additionally, it can be checked that $\mathcal{D}_1(G, G_*) = 0$ if and only if $G \equiv G_*$. Thus, when $\mathcal{D}_1(G, G_*)$ is sufficiently small, the differences $\Delta\beta_{1ij}$ and $\Delta\eta_{ij}$ are also small. This property indicates that $\mathcal{D}_1(G, G_*)$ is an appropriate loss function for measuring the discrepancy between the LSE $\widehat{G}_n$ and the true mixing measures G_*. However, since the loss $\mathcal{D}_1(G, G_*)$ is not symmetric, it is not a proper metric. Finally, computing the Voronoi loss function $\mathcal{D}_1$ is efficient as its computational complexity is at the order of $\mathcal{O}(k \times k_*)$.

Equipped with the Voronoi loss function $\mathcal{D}_1$, we are now ready to characterize the parameter estimation rates as well as the expert estimation rates in the following theorem.

Theorem 3.2. *Suppose that the expert function $h(x, \eta)$ satisfies the condition in Definition 3.1, then the following L^2-lower bound holds true for any $G \in \mathcal{G}_k(\Theta)$:*

$$\|f_G - f_{G_*}\|_{L^2(\mu)} \gtrsim \mathcal{D}_1(G, G_*).$$

Furthermore, this bound and the result in Theorem 2.1 imply that $\mathcal{D}_1(\widehat{G}_n, G_*) = \mathcal{O}_P(\sqrt{\log(n)/n})$.

The proof of Theorem 3.2 is in Appendix A.2. A few remarks regarding the results of Theorem 3.2 are in order.

(i) Firstly, the parameters $\beta_{1j}^*, \eta_{1j}^*$ that are approximated by more than one component, i.e. those for which $|\mathcal{A}_j(\widehat{G}_n)| > 1$, enjoy the same estimation rate of order $\mathcal{O}_P(n^{-1/4})$. Additionally, since the expert $h(x, \eta)$ is twice differentiable over a bounded domain, it is also a Lipschitz function. Therefore, by denoting $\widehat{G}_n := \sum_{i=1}^{\widehat{k}_n} \exp(\widehat{\beta}_{0i})\delta_{(\widehat{\beta}_{1i}^n, \widehat{\eta}_i^n)}$, we obtain that

$$\sup_x |h(x, \widehat{\eta}_i^n) - h(x, \eta_j^*)| \leq L_1 \|\widehat{\eta}_i^n - \eta_j^*\| \lesssim \mathcal{O}_P(n^{-1/4}), \quad (8)$$

for any $i \in \mathcal{A}_j(\widehat{G}_n)$, where $L_1 \geq 0$ is a Lipschitz constant. Consequently, the rate for estimating a strongly identifiable expert $h(x, \eta_j^*)$ continues to be $\mathcal{O}_P(n^{-1/4})$ as long as it is fitted by more than one expert. On the other hand, when considering the softmax gating Gaussian MoE, (Nguyen et al., 2023b) pointed out that the estimation rates for linear experts could be $\mathcal{O}_P(n^{-1/12})$ when they are fitted by three experts, i.e., $|\mathcal{A}_j(\widehat{G}_n)| = 3$. Moreover, these rates will become even slower if their number of fitted experts increases. This comment highlights how the strong identifiability condition proposed in this paper immediately implies fast estimation rates,

(ii) Secondly, the rates for estimating parameters β^*_{1j}, η^*_j that are fitted by exactly one component, i.e., $|\mathcal{A}_j(\widehat{G}_n)| = 1$, are faster than those in Remark (i), of order $\mathcal{O}_P(n^{-1/2})$. By employing the same arguments as in equation (8), we deduce that the expert $h(x, \eta^*_j)$ admits the estimation rate of order $\mathcal{O}_P(n^{-1/2})$, which matches its counterpart in (Nguyen et al., 2023b).

4. Ridge Experts

In this section, we turn to the softmax gating MoE models with ridge experts, i.e two-layer neural networks comprised of a linear layer and an activation layer of the form

$$h(x, \eta^*_j) = \sigma((a^*_j)^\top x + b^*_j), \tag{9}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is the (usually, nonlinear) activation function and $\eta^*_j = (a^*_j, b^*_j) \in \mathbb{R}^d \times \mathbb{R}$ are expert parameters. Ridge experts are commonly deployed in deep-learning architectures and generative models, and they fail to satisfy the strong identifiability condition from the last section. To overcome this issue, we instead formulate a *strong independence* condition on the activation function itself, which will guarantee fast estimation rates, provided that all the expert parameters are non-zero. Interestingly, when one or more parameters are zero, so that the corresponding experts are constant functions, we show slow, non-polynomial rates in the sample size.

In Section 4.2, we then examine polynomial activation functions, which violates the strong independence condition. In this case we again demonstrate slow rates.

4.1. On Strongly Independent Activation

To begin with, let us recall from Section 3 that our goal is to establish the L^2-lower bound $\|f_{\widehat{G}_n} - f_{G_*}\|_{L^2(\mu)} \gtrsim \mathcal{D}(\widehat{G}_n, G_*)$ where $G_* := \sum_{i=1}^{k_*} \exp(\beta^*_{0i})\delta_{(\beta^*_{1i}, a^*_i, b^*_i)}$,

$$f_{G_*}(x) := \sum_{i=1}^{k_*} \frac{\exp((\beta^*_{1i})^\top x + \beta^*_{0i})}{\sum_{j=1}^{k_*} \exp((\beta^*_{1j})^\top x + \beta^*_{0j})} \times \sigma((a^*_i)^\top x + b^*_i),$$

and $\mathcal{D}$ is a loss function among parameters that will be defined later. In our proof techniques, we first need to represent the term $f_{\widehat{G}_n}(x) - f_{G_*}(x)$ as a weighted sum of linearly independent terms by applying Taylor expansions to the function $F(X; \beta_1, a, b) = \exp(\beta_1^\top x)\sigma(a^\top x + b)$. Nevertheless, we notice that if $a^*_i = 0_d$ for some $i \in [k_*]$, then there is an interaction between gating and expert parameters expressed in the language of PDE as follows:

$$\frac{\partial F}{\partial \beta_1}(x; \beta^*_{1i}, a^*_i, b^*_i) = \sigma'(b^*_i) \cdot \frac{\partial F}{\partial a}(x; \beta^*_{1i}, a^*_i, b^*_i). \tag{10}$$

The above PDE leads to a number of linearly dependent terms in the decomposition of $f_{\widehat{G}_n}(x) - f_{G_*}(x)$, which could negatively affect the expert estimation rates. To understand the effects of the previous interaction better, we split the analysis into two following regimes of parameters a^*_i where the interaction (10) vanishes and occurs, respectively:

- **Regime 1:** All parameters $a^*_1, \ldots, a^*_{k_*}$ are different from 0_d;
- **Regime 2:** At least one among parameters $a^*_1, \ldots, a^*_{k_*}$ is equal to 0_d.

Subsequently, we will conduct an expert convergence analysis in each of the above regimes.

4.1.1. Regime 1: Input-dependent Experts

Under this regime, since all the parameters $a^*_1, \ldots, a^*_{k_*}$ are different from 0_d, the PDE (10) does not hold true, and thus, we do not need to deal with linearly dependent terms induced by this PDE. Instead, we establish a strong independence condition on the activation σ in Definition 4.1 to guarantee that there are no interactions among parameters, i.e. the derivatives of the function $x \mapsto F(x; \beta_1, a, b) = \exp(\beta_1^\top x)\sigma(a^\top x + b)$ and its derivatives up to the second order are linearly independent.

Definition 4.1 (Strong Independence)**.** We say that an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is *strongly independent* if it is twice differentiable and the set of functions in x

$$\Big\{x^\nu \sigma^{(\tau)}(a_j^\top x + b_j) : \nu \in \mathbb{N}^d, \tau \in \mathbb{N}, 0 \le |\nu|, \tau \le 2,\ j \in [k]\Big\},$$

is linearly independent, for almost all x, for any pair-wise distinct parameters $(a_1, b_1), \ldots, (a_k, b_k)$ and $k \ge 1$, where $\sigma^{(\tau)}$ denotes the τ-th derivative of σ.

Example. We can verify that $\mathrm{sigmoid}(\cdot)$ and Gaussian error linear units $\mathrm{GELU}(\cdot)$ (Hendrycks & Gimpel, 2023) are strongly independent activation functions. By contrast, the polynomial activation $\sigma(z) = z^p$ is not strongly independent for any $p \ge 1$.

Just like in the previous section, we construct an appropriate loss function among parameters that is upper bounded by the

$L^2(\mu)$ distance between the corresponding softmax gating MoE regression functions to obtain expert estimation rates.

Voronoi loss. Tailored to the setting of Regime 1, the Voronoi loss of interest is given by

$$\mathcal{D}_2(G,G_*) := \sum_{j=1}^{k_*}\Big|\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i})-\exp(\beta_{0j}^*)\Big| + \sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i})\Big[\|\Delta\beta_{1ij}\|^2+\|\Delta a_{ij}\|^2+|\Delta b_{ij}|^2\Big] + \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i})\Big[\|\Delta\beta_{1ij}\|+\|\Delta a_{ij}\|+|\Delta b_{ij}|\Big],$$

where we denote $\Delta a_{ij} := a_i - a_j^*$ and $\Delta b_{ij} := b_i - b_j^*$.

Theorem 4.2. *Assume that the experts take the form $\sigma(a^\top x+b)$, where the activation function $\sigma(\cdot)$ satisfies the condition in Definition 4.1, then the following L^2-lower bound holds true for any $G\in\mathcal{G}_k(\Theta)$ under the Regime 1:*

$$\|f_G - f_{G_*}\|_{L^2(\mu)} \gtrsim \mathcal{D}_2(G,G_*).$$

Furthermore, this bound and the result in Theorem 2.1 imply that $\mathcal{D}_2(\widehat{G}_n,G_*)=\mathcal{O}_P(\sqrt{\log(n)/n})$.

See Appendix A.3 for a proof. Theorem 4.2 indicates the LSE $\widehat{G}_n$ converges to G_* at the parametric rate $\mathcal{O}_P(n^{-1/2})$ under the loss function $\mathcal{D}_2$. From the formulation of this loss, we deduce the following rates.

(i) For parameters β_{1j}^*, a_j^* and b_j^* fitted by one component, i.e., $|\mathcal{A}_j(\widehat{G}_n)|=1$, the estimation rate is of order $\mathcal{O}_P(n^{-1/2})$. Moreover, as the strongly independent σ is twice differentiable, the function $x\mapsto\sigma(a^\top x+b)$ is Lipschitz continuous with some Lipschitz constant $L_2\geq 0$. Thus, denoting $\widehat{G}_n := \sum_{i=1}^{\widehat{k}_n}\exp(\widehat{\beta}_{0i})\delta_{(\widehat{\beta}_{1i}^n,\widehat{a}_i^n,\widehat{b}_i^n)}$, we have

$$\begin{aligned}\sup_x|\sigma((\widehat{a}_i^n)^\top x+\widehat{b}_i^n)-\sigma((a_j^*)^\top x+b_j^*)| &\leq L_2\cdot\|(\widehat{a}_i^n,\widehat{b}_i^n)-(a_j^*,b_j^*)\|\\ &\leq L_2\cdot(\|\widehat{a}_i^n-a_j^*\|+|\widehat{b}_i^n-b_j^*|)\\ &\lesssim \mathcal{O}_P(n^{-1/2}).\end{aligned} \quad (11)$$

As a consequence, the estimation rate for the expert $\sigma((a_j^*)^\top x+b_j^*)$ is also of order $\mathcal{O}_P(n^{-1/2})$.

(ii) For parameters, say β_{1j}^*, a_j^* and b_j^*, fitted by more than one component, i.e. $|\mathcal{A}_j(\widehat{G}_n)|>1$, the corresponding rates are $\mathcal{O}_P(n^{-1/4})$. By reusing the arguments in equation (11), we deduce that the expert $\sigma((a_j^*)^\top x+b_j^*)$ admits the estimation rate of order $\mathcal{O}_P(n^{-1/4})$.

4.1.2. Regime 2: Input-independent Experts

Recall that under this regime, at least one among parameters $a_1^*,\ldots,a_{k_*}^*$ equal to 0_d. Without loss of generality, we may assume that $a_1^*=0_d$. This means that the the value of the first expert $\sigma((a_1^*)^\top x+b_1^*)$ no longer depends on the input x. In this case, there exists an interaction among the gating parameter β_1 and the expert parameter a captured by the PDE

$$\frac{\partial F}{\partial\beta_1}(x;\beta_{11}^*,a_1^*,b_1^*)=\sigma'(b_1^*)\cdot\frac{\partial F}{\partial a}(x;\beta_{11}^*,a_1^*,b_1^*). \quad (12)$$

The significance of this fact is that, owing to the the above PDE, the following Voronoi loss function among parameters is not majorized by the $L^2(\mu)$ distance between the corresponding expert functions:

$$\mathcal{D}_{3,r}(G,G_*) := \sum_{j=1}^{k_*}\Big|\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i})-\exp(\beta_{0j}^*)\Big| + \sum_{j=1}^{k_*}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i})\Big[\|\Delta\beta_{1ij}\|^r+\|\Delta a_{ij}\|^r+|\Delta b_{ij}|^r\Big], \quad (13)$$

for any $r\geq 1$. This is formalized in the next result, whose proof can be found in Appendix A.4.

Proposition 4.3. *Let the expert function take the form $\sigma(a^\top x+b)$, and suppose that not all the parameters $a_1^*,\ldots,a_{k_*}^*$ are different from 0_d, then we obtain that*

$$\lim_{\varepsilon\to 0}\inf_{\substack{G\in\mathcal{G}_k(\Theta):\\ \mathcal{D}_{3,r}(G,G_*)\leq\varepsilon}}\|f_G-f_{G_*}\|_{L^2(\mu)}/\mathcal{D}_{3,r}(G,G_*)=0,$$

for any $r\geq 1$.

The above proposition, combined with Theorem 2.1, indicates that the parameter estimation rate in this situation ought to be slower than any polynomial of $1/\sqrt{n}$. This is indeed the case, as confirmed by the following minimax lower bound.

Theorem 4.4. *Assume that the experts take the form $\sigma(a^\top x+b)$, then the following minimax lower bound of estimating G_* holds true for any $r\geq 1$ under the Regime 2:*

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

where $\mathbb{E}_{f_G}$ indicates the expectation taken w.r.t the product measure with f_G^n and the infimum is over all estimators taking values in $\mathcal{G}_k$.

The proof of Theorem 4.4 is in Appendix A.5. This result together with the formulation of the Voronoi loss $\mathcal{D}_{3,r}$ in equation (13) leads to a singular and striking phenomenon that, to the best of our knowledge, has never been observed in previous work (Ho et al., 2022; Chen et al., 2022; Nguyen et al., 2023b). Specifically,

(i) The rates for estimating the parameters β_{1j}^*, a_j^* and b_j^* are slower than any polynomial rate $\mathcal{O}_P(n^{-1/2r})$ for any $r \geq 1$. In particular, they could be as slow as $\mathcal{O}_P(1/\log(n))$.

(ii) Recall from equation (11) that

$$\sup_x |\sigma((\widehat{a}_i^n)^\top x + \widehat{b}_i^n) - \sigma((a_j^*)^\top x + b_j^*)| \leq L_2 \cdot (\|\widehat{a}_i^n - a_j^*\| + |\widehat{b}_i^n - b_j^*|). \quad (14)$$

Consequently, the expert estimation rates might also be significantly slow, of order $\mathcal{O}_P(1/\log(n))$ or worse, due to the interaction between gating and expert parameters in equation (12). It is worth noting that these slow rates occur even when the activation function σ meets the strong independence condition in Definition 4.1. This observation suggests that all the expert parameters $a_1^*, \ldots, a_{k_*}^*$ should be different from 0_d. In other words, every expert of the form $\sigma(a^\top x + b)$ in the MoE model should depend on the input value.

4.2. On Polynomial Activation

We now focus on a specific setting in which the activation function σ is formulated as a polynomial, i.e. $\sigma(z) = z^p$, for some $p \in \mathbb{N}$. Concretely, for all $j \in [k^*]$, we set

$$h(x, \eta_j^*) = ((a_j^*)^\top x + b_j^*)^p, \quad x \in \mathcal{X}, \quad (15)$$

and call it a polynomial expert. Notably, it can be verified that this activation function violates the strong independence condition in Definition 4.1 for any $p \in \mathbb{N}$. For simplicity, let us consider only the setting when $p = 1$, i.e., $h(x, \eta_j^*) = (a_j^*)^\top x + b_j^*$, with a note that the results for other settings of p can be argued in a similar fashion.

Since the strong independence condition in Definition 4.1 is not satisfied, we have to deal with an interaction among parameters, capture by following PDE:

$$\frac{\partial^2 F}{\partial \beta_1 \partial b}(x; \beta_{1i}^*, a_i^*, b_i^*) = \frac{\partial F}{\partial a}(x; \beta_{1i}^*, a_i^*, b_i^*), \quad (16)$$

where $F(x; \beta_1, a, b) := \exp(\beta_1^\top x)(a^\top x + b)$ is the product of softmax numerator and the expert function. Though this interaction has already been observed and analyzed in previous work (Nguyen et al., 2023b), its effects on the expert convergence rate in the present settings are totally different as we consider a deterministic MoE model rather than a probabilistic model. In particular, (Nguyen et al., 2023b) argued that the interaction (16) led to polynomial expert estimation rates which were determined by the solvability of a system of polynomial equations. On the other hand, we show in the Proposition 4.5 below that such interaction makes the ratio $\|f_G - f_{G_*}\|_{L^2(\mu)}/\mathcal{D}_{3,r}(G, G_*)$ vanish when the loss $\mathcal{D}_{3,r}(G, G_*)$ goes to zero as shown in the next proposition, whose proof can be found in Appendix A.6.

Proposition 4.5. *Let the expert functions take the form $a^\top x + b$, then the following limit holds for any $r \geq 1$:*

$$\lim_{\varepsilon \to 0} \inf_{\substack{G \in \mathcal{G}_k(\Theta): \\ \mathcal{D}_{3,r}(G, G_*) \leq \varepsilon}} \|f_G - f_{G_*}\|_{L^2(\mu)}/\mathcal{D}_{3,r}(G, G_*) = 0.$$

Just like in the previous section, we arrive at a significantly slow expert estimation rates.

Theorem 4.6. *Assume that the experts take the form $a^\top x + b$, then we achieve the following minimax lower bound of estimating G_*:*

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

for any $r \geq 1$, where $\mathbb{E}_{f_G}$ indicates the expectation taken w.r.t the product measure with f_G^n.

Proof of Theorem 4.6 is in Appendix A.7. A few comments regarding the above theorem are in order (see also Table 2):

(i) Theorem 4.6 reveals that using polynomial experts will result in the same slow rates as using input-independent experts, as described in Theorem 4.4. More specifically, the estimation rates for parameters β_{1i}^*, a_i^* and b_i^* are slower than any polynomial rates, and could be of order $\mathcal{O}_P(1/\log(n))$ because of the interaction in equation (16).

(ii) Additionally, we have that

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

Since the input space $\mathcal{X}$ is bounded, we deduce that the rates for estimating polynomial experts $(a_j^*)^\top x + b_j^*$ could also be as slow as $\mathcal{O}_P(1/\log(n))$. This is remarkable, especially in contrast to the polynomial rates of linear expert established by (Nguyen et al., 2023b) in probabilistic softmax gating experts. Hence, for the expert estimation problem, the performance of a mixture of linear experts cannot compare to that of a mixture of non-linear experts. It is worth noting that this claim aligns with the findings in (Chen et al., 2022).

5. Conclusions

In this paper, we have analyzed the convergence rates of the least squares estimator under a deterministic softmax gating MoE model. We have shown that expert functions that satisfy a novel condition referred to as strong identifiability enjoy estimation rates of polynomial orders. When specializing to experts of the form ridge function, polynomial rates can be guaranteed under another condition, called strongly independent activation, provided that all the expert parameters are non-zero. In contrast, when at least one of the expert parameters vanishes, we have unveiled the

Table 2. Comparison of parameter and expert estimation rates under the probabilistic softmax gating mixture of **linear experts** (Nguyen et al., 2023b) and the deterministic one (Ours). Here, we denote $\bar{r}_j := \bar{r}(|\mathcal{A}_j^n|)$, where the function $\bar{r}(\cdot)$ represents for the solvability of a system of polynomial equations in (Nguyen et al., 2023b). Some specific values of this function are given by: $\bar{r}(2) = 4$ and $\bar{r}(3) = 6$.

Model Type	**Parameters** a_j^*		**Parameters** b_j^*		**Experts** $(a_j^*)^\top x + b_j^*$	
	$j : \lvert\mathcal{A}_j^n\rvert = 1$	$j : \lvert\mathcal{A}_j^n\rvert > 1$	$j : \lvert\mathcal{A}_j^n\rvert = 1$	$j : \lvert\mathcal{A}_j^n\rvert > 1$	$j : \lvert\mathcal{A}_j^n\rvert = 1$	$j : \lvert\mathcal{A}_j^n\rvert > 1$
Probabilistic	$\mathcal{O}_P(n^{-1/2})$	$\mathcal{O}_P(n^{-1/\bar{r}_j})$	$\mathcal{O}_P(n^{-1/2})$	$\mathcal{O}_P(n^{-1/2\bar{r}_j})$	$\mathcal{O}_P(n^{-1/2})$	$\mathcal{O}_P(n^{-1/2\bar{r}_j})$
Deterministic	Slower than $\mathcal{O}_P(n^{-1/2r}), \forall r \geq 1$					

surprising fact that expert estimation rates become slower than any polynomial rates. Furthermore, we also prove that polynomial experts, which violate the strong identifiability condition, also experience such slow rates under any parameter settings.

Limitations. Our analysis has two following limitations:

(L.1) The theoretical results established in the paper are under the assumption that the data are generated from a regression model where the regression function is a softmax gating MoE. This assumption can be violated in real-world settings when the data are not necessarily generated from that model. Under those misspecified settings, the regression function is an arbitrary function g which is not necessarily a mixture of experts. Then, the least square estimator $\widehat{G}_n$ defined in equation (3) converges to the mixing measure $\overline{G} \in \mathcal{G}_k$ that minimizes the L^2-distance between f_G and g. Since the current analysis of the least square estimation under the misspecified settings of statistical models is mostly conducted when the function space is convex (van de Geer, 2000), it is inapplicable to our setting where the space $\mathcal{G}_k$ is non-convex. Therefore, we believe that further techniques should be developed to analyze the misspecified settings, which is beyond the scope of our work.

(L.2) The depth of an expert network has not been considered in capturing the convergence rate of expert estimation. In particular, we demonstrate in Theorem 3.2 that any choice of expert network which satisfies the strong identifiability condition will lead to polynomial expert estimation rates regardless of its depth. Secondly, although ridge experts of the form $h(x, (a, b)) = \sigma(a^\top x + b)$ are not strongly identifiable, we show in Theorem 4.2 that if the activation σ satisfies the strong independence condition in Definition 4.1, then the expert estimation rates are also polynomial. On the other hand, for ridge experts with activation σ violating the strong independence condition, e.g. polynomial experts, we find that increasing the depth of the expert network would not help improve the slow expert estimation rates in Theorem 4.6 due to an intrinsic interaction among parameters of polynomial experts (expressed in the language of partial differential equations). We believe that technical tools need to be further developed to understand the effects of the network depth on the expert estimation problem. As it stays beyond the scope of our work, we leave that direction for future work.

Future directions. There are some potential directions to which our current theory can extend. Firstly, we can leverage our techniques to capture the convergence behavior of different types of experts under the MoE models with other gating functions, namely Top-K sparse gate (Shazeer et al., 2017), dense-to-sparse gate (Nie et al., 2022), cosine similarity gate (Li et al., 2023), Laplace gate (Han et al., 2024), and sigmoid gate (Csordás et al., 2023). Such analysis would enrich the knowledge of expert selection given a specific gating function. Additionally, we can develop our current techniques to provide a comprehensive understanding of more complex MoE models such as hierarchical MoE (Zhao et al., 1994; Jacobs et al., 1997) and multigate MoE (Ma et al., 2018; Liang et al., 2022), which have remained elusive in the literature.

Acknowledgements

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

Impact Statement

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

References

Chen, Z., Deng, Y., Wu, Y., Gu, Q., and Li, Y. Towards understanding the mixture-of-experts layer in deep learning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Informa-*

tion Processing Systems, volume 35, pp. 23049–23062. Curran Associates, Inc., 2022.

Chow, Y., Tulepbergenov, A., Nachum, O., Gupta, D., Ryu, M., Ghavamzadeh, M., and Boutilier, C. A Mixture-of-Expert Approach to RL-based Dialogue Management. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=4FBUihxz5nm.

Csordás, R., Irie, K., and Schmidhuber, J. Approximating two-layer feedforward networks for efficient transformers. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 674–692, Singapore, December 2023. Association for Computational Linguistics.

De Veaux, R. D. Mixtures of linear regressions. *Computational Statistics and Data Analysis*, 8(3):227–245, 1989. ISSN 0167-9473.

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

Faria, S. and Soromenho, G. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, 2010.

Fedus, W., Dean, J., and Zoph, B. A review of sparse expert models in deep learning. *arXiv preprint arxiv 2209.01667*, 2022a.

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

Gupta, S., Mukherjee, S., Subudhi, K., Gonzalez, E., Jose, D., Awadallah, A., and Gao, J. Sparsely activated mixture-of-experts are robust multi-task learners. *arXiv preprint arxiv 2204.0768*, 2022.

Han, X., Nguyen, H., Harris, C., Ho, N., and Saria, S. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. *arXiv preprint arXiv:2402.03226*, 2024.

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

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

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

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

Jacobs, R. A., Peng, F., and Tanner, M. A. A Bayesian Approach to Model Selection in Hierarchical Mixtures-of-Experts Architectures. *Neural Networks*, 10(2):231–241, 1997. ISSN 0893-6080.

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

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

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

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

Lindsay, B. *Mixture models: Theory, geometry and applications*. In NSF-CBMS Regional Conference Series in Probability and Statistics. IMS, Hayward, CA., 1995.

Ma, J., Zhao, Z., Yi, X., Chen, J., Hong, L., and Chi, E. H. Modeling Task Relationships in Multi-Task Learning with Multi-Gate Mixture-of-Experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pp. 1930–1939, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 978-1-4503-5552-0. doi: 10.1145/3219819.3220007. URL https://doi.org/10.1145/3219819.3220007. event-place: London, United Kingdom.

Makkuva, A., Viswanath, P., Kannan, S., and Oh, S. Breaking the gridlock in mixture-of-experts: Consistent and efficient algorithms. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4304–4313. PMLR, 09–15 Jun 2019.

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

McLachlan, G. J. and Basford, K. E. *Mixture models: Inference and Applications to Clustering. Statistics: Textbooks and Monographs.* New York, 1988.

Mendes, E. F. and Jiang, W. Convergence rates for mixture-of-experts. *arXiv preprint arxiv 1110.2058*, 2011.

Nguyen, H., Akbarian, P., Nguyen, T., and Ho, N. A general theory for softmax gating multinomial logistic mixture of experts. *arXiv preprint arXiv:2310.14188*, 2023a.

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

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

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

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

Pham, Q., Do, G., Nguyen, H., Nguyen, T., Liu, C., Sartipi, M., Nguyen, B. T., Ramasamy, S., Li, X., Hoi, S., and Ho, N. Competesmoe – effective training of sparse mixture of experts via competition. *arXiv preprint arXiv:2402.02526*, 2024.

Puigcerver, J., Riquelme, C., Mustafa, B., and Houlsby, N. From sparse to soft mixtures of experts. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Ruiz, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Pinto, A., Keysers, D., and Houlsby, N. Scaling vision with sparse mixture of experts. In *NeurIPS*, 2021.

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

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

Yu, B. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*, pp. 423–435, 1997.

Zhao, Y., Schwartz, R., Sroka, J., and Makhoul, J. Hierarchical mixtures of experts methodology applied to continuous speech recognition. In *Advances in Neural Information Processing Systems*, volume 7, 1994.

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

Supplementary Material for "On Least Square Estimation in Softmax Gating Mixture of Experts"

In this supplementary material, we provide proofs for the main results in the paper in Appendix A, while we leave proofs for the identifiability of the softmax gating mixture of experts in Appendix B. Finally, we run several numerical experiments in Appendix C to empirically justify our theoretical results.

A. Proofs of Main Results

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

A.1. Proof of Theorem 2.1

For the proof of the theorem, we first introduce some notation. 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, Geer et. al. (van de Geer, 2000) suggest using the following quantity:

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

where $H_B(t, \mathcal{F}_k(\Theta, t), \|\cdot\|_{L^2(\mu)})$ stands for the bracketing entropy (van de Geer, 2000) 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 (van de Geer, 2000) with notations being adapted to this work, we obtain the following lemma:

Lemma A.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*

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

for all $\delta \geq \delta_n$.

We now demonstrate that when the expert functions are Lipschitz continuous, the following bound holds:

$$H_B(\varepsilon, \mathcal{F}_k(\Theta), \|.\|_{L^2(\mu)}) \lesssim \log(1/\varepsilon), \tag{18}$$

for any $0 < \varepsilon \leq 1/2$. Indeed, for any function $f_G \in \mathcal{F}_k(\Theta)$, since the expert functions are bounded, we obtain that $f_G(x) \leq M$ for all x where M is bounded constant of the expert functions. Let $\tau \leq \varepsilon$ and $\{\pi_1, \ldots, \pi_N\}$ be the τ-cover under the L^2 norm of the set $\mathcal{F}_k(\Theta)$ where $N := N(\tau, \mathcal{F}_k(\Theta), \|\cdot\|_{L^2(\mu)})$ is the η-covering number of the metric space $(\mathcal{F}_k(\Theta), \|\cdot\|_{L^2(\mu)})$. Then, we construct the brackets of the form $[L_i(x), U_i(x)]$ for all $i \in [N]$ as follows:

$$\begin{aligned} L_i(x) &:= \max\{\pi_i(x) - \tau, 0\}, \\ U_i(x) &:= \max\{\pi_i(x) + \tau, M\}. \end{aligned}$$

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

$$\|U_i - L_i\|_{L^2(\mu)}^2 = \int (U_i - L_i)^2 \mathrm{d}\mu(x) \leq \int 16\tau^2 \mathrm{d}\mu(x) = 16\tau^2,$$

which implies that $\|U_i - L_i\|_{L^2(\mu)} \leq 4\tau$. By definition of the bracketing entropy, we deduce that

$$H_B(4\tau, \mathcal{F}_k(\Theta), \|\cdot\|_{L^2(\mu)}) \leq \log N = \log N(\tau, \mathcal{F}_k(\Theta), \|\cdot\|_{L^2(\mu)}). \tag{19}$$

Therefore, we need to provide an upper bound for the covering number N. In particular, we denote $\Delta := \{(\beta_0, \beta_1) \in \mathbb{R} \times \mathbb{R}^d : (\beta_0, \beta_1, \eta) \in \Theta\}$ and $\Omega := \{\eta \in \mathbb{R}^q : (\beta_0, \beta_1, \eta) \in \Theta\}$. Since Θ is a compact set, Δ and Ω are also compact. Therefore, we can find τ-covers Δ_τ and Ω_τ for Δ and Ω, respectively. We can check that

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

For each mixing measure $G = \sum_{i=1}^{k} \exp(\beta_{0i})\delta_{(\beta_{1i},\eta_i)} \in \mathcal{G}_k(\Theta)$, we consider other two mixing measures:

$$\widetilde{G} := \sum_{i=1}^{k} \exp(\beta_{0i})\delta_{(\beta_{1i},\overline{\eta}_i)}, \qquad \overline{G} := \sum_{i=1}^{k} \exp(\overline{\beta}_{0i})\delta_{(\overline{\beta}_{1i},\overline{\eta}_i)}.$$

Here, $\overline{\eta}_i \in \Omega_\tau$ such that $\overline{\eta}_i$ is the closest to η_i in that set, while $(\overline{\beta}_{0i}, \overline{\beta}_{1i}) \in \Delta_\tau$ is the closest to (β_{0i}, β_{1i}) in that set. From the above formulations, we get that

$$\begin{aligned}
\|f_G - f_{\widetilde{G}}\|^2_{L^2(\mu)} &= \int \left[\sum_{i=1}^{k} \frac{\exp((\beta_{1i})^\top x + \beta_{0i})}{\sum_{j=1}^{k} \exp((\beta_{1j})^\top x + \beta_{0j})} \cdot [h(x,\eta_i) - h(x,\overline{\eta}_i)\right]^2 \mathrm{d}\mu(x) \\
&\leq k \int \sum_{i=1}^{k} \left[\frac{\exp((\beta_{1i})^\top x + \beta_{0i})}{\sum_{j=1}^{k} \exp((\beta_{1j})^\top x + \beta_{0j})} \cdot [h(x,\eta_i) - h(x,\overline{\eta}_i)\right]^2 \mathrm{d}\mu(x) \\
&\leq k \int \sum_{i=1}^{k} \, [h(x,\eta_i) - h(x,\overline{\eta}_i)]^2 \,\mathrm{d}\mu(x) \\
&\leq k \int \sum_{i=1}^{k} \, [L_1 \cdot \|\eta_i - \overline{\eta}_i\|]^2 \,\mathrm{d}\mu(x) \\
&\leq k^2 (L_1 \tau)^2,
\end{aligned}$$

which indicates that $\|f_G - f_{\widetilde{G}}\|_{L^2(\mu)} \leq L_1 k \tau$. Here, the second inequality is according to the Cauchy-Schwarz inequality, the third inequality occurs as the softmax weight is bounded by 1, and the fourth inequality follows from the fact that the expert $h(x,\cdot)$ is a Lipschitz function with Lipschitz constant L_1. Next, we have

$$\begin{aligned}
\|f_{\widetilde{G}} - f_{\overline{G}}\|^2_{L^2(\mu)} &\leq k \int \sum_{i=1}^{k} \left[\left(\frac{\exp((\beta_{1i})^\top x + \beta_{0i})}{\sum_{j=1}^{k} \exp((\beta_{1j})^\top x + \beta_{0j})} - \frac{\exp((\overline{\beta}_{1i})^\top x + \overline{\beta}_{0i})}{\sum_{j=1}^{k} \exp((\overline{\beta}_{1j})^\top x + \overline{\beta}_{0j})}\right) \cdot h(x,\overline{\eta}_i)\right]^2 \mathrm{d}\mu(x) \\
&\leq kM^2L^2 \int \sum_{i=1}^{k} \left[\|\beta_{1i} - \overline{\beta}_{1i}\| \cdot \|x\| + |\beta_{0i} - \overline{\beta}_{0i}|\right]^2 \mathrm{d}\mu(x) \\
&\leq kM^2L^2 \int \sum_{i=1}^{k} (\tau \cdot B + \tau)^2 \mathrm{d}\mu(x) \\
&\leq [kML\tau(B+1)]^2,
\end{aligned}$$

where $L \geq 0$ is a Lipschitz constant of the softmax weight. This result implies that $\|f_{\widetilde{G}} - f_{\overline{G}}\|_{L^2(\mu)} \leq kML(B+1)\tau$. According to the triangle inequality, we have

$$\|f_G - f_{\overline{G}}\|_{L^2(\mu)} \leq \|f_G - f_{\widetilde{G}}\|_{L^2(\mu)} + \|f_{\widetilde{G}} - f_{\overline{G}}\|_{L^2(\mu)} \leq [L_1 k + kML(B+1)] \cdot \tau.$$

By definition of the covering number, we deduce that

$$N(\tau, \mathcal{F}_k(\Theta), \|\cdot\|_{L^2(\mu)}) \leq |\Delta_\tau| \times |\Omega_\tau| \leq \mathcal{O}_P(n^{-(d+1)k}) \times \mathcal{O}(n^{-qk}) \leq \mathcal{O}(n^{-(d+1+q)k}). \tag{20}$$

Combine equations (19) and (20), we achieve that

$$H_B(4\tau, \mathcal{F}_k(\Theta), \|\cdot\|_{L^2(\mu)}) \lesssim \log(1/\tau).$$

Let $\tau = \varepsilon/4$, then we obtain that

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

As a result, it follows that

$$\mathcal{J}_B(\delta, \mathcal{F}_k(\Theta, \delta)) = \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(t, \mathcal{F}_k(\Theta, t), \|\cdot\|_{L^2(\mu)})\, \mathrm{d}t \vee \delta \lesssim \int_{\delta^2/2^{13}}^{\delta} \log(1/t) dt \vee \delta. \tag{21}$$

Let $\Psi(\delta) = \delta \cdot [\log(1/\delta)]^{1/2}$, then $\Psi(\delta)/\delta^2$ is a non-increasing function of δ. Furthermore, equation (21) 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 A.1, we achieve the desired conclusion of the theorem.

A.2. Proof of Theorem 3.2

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

$$\inf_{G \in \mathcal{G}_k(\Theta)} \|f_G - f_{G_*}\|_{L^2(\mu)} / \mathcal{D}_1(G, G_*) > 0. \tag{22}$$

For that purpose, we divide the proof of the above inequality into local and global parts in the sequel.

Local part: In this part, we demonstrate that

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{G}_k(\Theta): \mathcal{D}_1(G, G_*) \leq \varepsilon} \|f_G - f_{G_*}\|_{L^2(\mu)} / \mathcal{D}_1(G, G_*) > 0. \tag{23}$$

Assume by contrary that the above inequality does not hold true, then there exists a sequence of mixing measures $G_n = \sum_{i=1}^{k_*} \exp(\beta_{0i}^n) \delta_{(\beta_{1i}^n, \eta_i^n)}$ in $\mathcal{G}_k(\Theta)$ such that $\mathcal{D}_{1n} := \mathcal{D}_1(G_n, G_*) \to 0$ and

$$\|f_{G_n} - f_{G_*}\|_{L^2(\mu)} / \mathcal{D}_{1n} \to 0, \tag{24}$$

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 $\mathcal{D}_{1n}$ can be represented as

$$\begin{aligned}\mathcal{D}_{1n} := \sum_{j=1}^{k_*} \Big| \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^*) \Big| + \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n) \Big[\|\Delta\beta_{1ij}^n\|^2 + \|\Delta\eta_{ij}^n\|^2 \Big] \\ + \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n) \Big[\|\Delta\beta_{1ij}^n\| + \|\Delta\eta_{ij}^n\| \Big],\end{aligned} \tag{25}$$

where we denote $\Delta\beta_{1ij}^n := \beta_{1i}^n - \beta_{1j}^*$ and $\Delta\eta_{ij}^n := \eta_i^n - \eta_j^*$.

Since $\mathcal{D}_{1n} \to 0$, we get that $(\beta_{1i}^n, \eta_i^n) \to (\beta_{1j}^*, \eta_j^*)$ and $\sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n) \to \exp(\beta_{0j}^*)$ as $n \to \infty$ for any $i \in \mathcal{A}_j$ and $j \in [k_*]$. Now, we divide the proof of local part into three steps as follows:

Step 1. In this step, we decompose the term $Q_n(x) := [\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top x + \beta_{0j}^*)] \cdot [f_{G_n}(x) - f_{G_*}(x)]$ into a combination of linearly independent elements using Taylor expansion. In particular, let us denote $F(x; \beta_1, \eta) := \exp(\beta_1^\top x) h(x, \eta)$ and $H(x; \beta_1) = \exp(\beta_1^\top x) f_{G_n}(x)$, then we have

$$\begin{aligned} Q_n(x) &= \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n) \Big[F(x; \beta_{1i}^n, \eta_i^n) - F(x; \beta_{1j}^*, \eta_j^*) \Big] \\ &- \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n) \Big[H(x; \beta_{1i}^n) - H(x; \beta_{1j}^*) \Big] \\ &+ \sum_{j=1}^{k_*} \Big(\sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^*) \Big) \Big[F(x; \beta_{1j}^*, \eta_j^*) - H(x; \beta_{1j}^*) \Big] \\ &:= A_n(x) - B_n(x) + E_n(x). \end{aligned} \tag{26}$$

Decomposition of A_n. Next, we continue to separate the term A_n into two parts as follows:

$$\begin{aligned}A_n(x) &:= \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\Big[F(x;\beta_{1i}^n,\eta_i^n)-F(x;\beta_{1j}^*,\eta_j^*)\Big]\\&+\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\Big[F(x;\beta_{1i}^n,\eta_i^n)-F(x;\beta_{1j}^*,\eta_j^*)\Big]\\&:= A_{n,1}(x)+A_{n,2}(x).\end{aligned}$$

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

$$A_{n,1}(x)=\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\sum_{|\alpha|=1}(\Delta\beta_{1ij}^n)^{\alpha_1}(\Delta\eta_{ij}^n)^{\alpha_2}\cdot\frac{\partial F}{\partial\beta_1^{\alpha_1}\partial\eta^{\alpha_2}}(x;\beta_{1j}^*,\eta_j^*)+R_1(x),$$

where $R_1(x)$ is a Taylor remainder such that $R_1(x)/\mathcal{D}_{1n}\to 0$ as $n\to\infty$. By taking the first derivatives of F w.r.t its parameters, we get

$$\begin{aligned}\frac{\partial F}{\partial\beta_1}(x;\beta_{1j}^*,\eta_j^*)&=x\exp((\beta_{1j}^*)^\top x)h(x,\eta_j^*)=x\cdot F(x;\beta_{1j}^*,\eta_j^*),\\\frac{\partial F}{\partial\eta}(x;\beta_{1j}^*,\eta_j^*)&=\exp((\beta_{1j}^*)^\top x)\cdot\frac{\partial h}{\partial\eta}(x,\eta_j^*):=F_1(x;\beta_{1j}^*,\eta_j^*).\end{aligned}$$

Thus, we can rewrite $A_{n,1}(x)$ as

$$A_{n,1}(x)=\sum_{j:|\mathcal{A}_j|=1}C_{n,1,j}(x)+R_1(x),\tag{27}$$

where

$$C_{n,1,j}(x)=\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\Big[(\Delta\beta_{1ij}^n)^\top x\cdot F(x;\beta_{1j}^*,\eta_j^*)+(\Delta\eta_{ij}^n)^\top F_1(x;\beta_{1j}^*,\eta_j^*)\Big].$$

Next, by applying the second-order Taylor expansion, $A_{n,2}(x)$ can be represented as

$$A_{n,2}(x)=\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\sum_{|\alpha|=1}^{2}\frac{1}{\alpha!}(\Delta\beta_{1ij}^n)^{\alpha_1}(\Delta\eta_{ij}^n)^{\alpha_2}\cdot\frac{\partial^{|\alpha_1|+|\alpha_2|}F}{\partial\beta_1^{\alpha_1}\partial\eta^{\alpha_2}}(x;\beta_{1j}^*,\eta_j^*)+R_2(x),$$

where $R_2(x)$ is a Taylor remainder such that $R_2(x)/\mathcal{D}_{1n}\to 0$ as $n\to\infty$. The second derivatives of F w.r.t its parameters are given by

$$\begin{aligned}&\frac{\partial^2 F}{\partial\beta\partial\beta^\top}(x;\beta_{1j}^*,\eta_j^*)=xx^\top\cdot F(x;\beta_{1j}^*,\eta_j^*),\quad\frac{\partial^2 F}{\partial\beta\partial\eta^\top}(x;\beta_{1j}^*,\eta_j^*)=x\cdot[F_1(x;\beta_{1j}^*,\eta_j^*)]^\top,\\&\frac{\partial^2 F}{\partial\eta\partial\eta^\top}(x;\beta_{1j}^*,\eta_j^*)=\exp((\beta_{1j}^*)^\top x)\cdot\frac{\partial^2 h}{\partial\eta\partial\eta^\top}:=F_2(x;\beta_{1j}^*,\eta_j^*)\end{aligned}$$

Therefore, the term $A_{n,2}(x)$ becomes

$$A_{n,2}(x)=\sum_{j:|\mathcal{A}_j|>1}[C_{n,1,j}(x)+C_{n,2,j}(x)]+R_2(x),\tag{28}$$

where

$$\begin{aligned}C_{n,2,j}(x)&:=\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\bigg\{\Big[x^\top\Big(M_d\odot(\Delta\beta_{1ij}^n)(\Delta\beta_{1ij}^n)^\top\Big)x\Big]\cdot F(x;\beta_{1j}^*,a_j^*,b_j^*)\\&+\Big[x^\top(\Delta\beta_{1ij}^n)(\Delta\eta_{ij}^n)^\top F_1(x;\beta_{1j}^*,\eta_j^*)\Big]+\Big[(\Delta\eta_{ij}^n)^\top\Big(M_d\odot F_2(x;\beta_{1j}^*,\eta_j^*)\Big)(\Delta\eta_{ij}^n)\Big]\bigg\},\end{aligned}$$

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

Decomposition of B_n. Subsequently, we also divide B_n into two terms based on the Voronoi cells as

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

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

$$B_{n,1}(x)=\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)(\Delta\beta_{1ij}^n)^\top x\cdot H(x;\beta_{1j}^*)+R_3(x), \tag{29}$$

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

$$B_{n,2}(x)=\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\Big[(\Delta\beta_{1ij}^n)^\top x+(\Delta\beta_{1ij}^n)^\top\Big(M_d\odot xx^\top\Big)(\Delta\beta_{1ij}^n)\Big]\cdot H(x;\beta_{1j}^*)+R_4(x), \tag{30}$$

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

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

$$\begin{aligned}&\Big\{F(x;\beta_{1j}^*,\eta_j^*),\ x^{(u)}F(x;\beta_{1j}^*,\eta_j^*),\ x^{(u)}x^{(v)}F(x;\beta_{1j}^*,\eta_j^*):u,v\in[d],\ j\in[k_*]\Big\},\\&\cup\Big\{[F_1(x;\beta_{1j}^*,\eta_j^*)]^{(u)},\ x^{(u)}[F_1(x;\beta_{1j}^*,\eta_j^*)]^{(v)}:u,v\in[d],\ j\in[k_*]\Big\},\\&\cup\Big\{[F_2(x;\beta_{1j}^*,\eta_j^*)]^{(uv)}:u,v\in[d],\ j\in[k_*]\Big\},\\&\cup\Big\{H(x;\beta_{1j}^*),\ x^{(u)}H(x;\beta_{1j}^*),\ x^{(u)}x^{(v)}H(x;\beta_{1j}^*):u,v\in[d],\ j\in[k_*]\Big\}.\end{aligned}$$

Step 2. In this step, we prove by contradiction that at least one among coefficients in the representations of $[A_n-R_1(x)-R_2(x)]/\mathcal{D}_{2n}$, $[B_n-R_3(x)-R_4(x)]/\mathcal{D}_{2n}$ and $E_n(x)/\mathcal{D}_{2n}$ does not go to zero as n tends to infinity. Indeed, assume that all of them converge to zero. Then, by considering the coefficients of

- $F(x;\beta_{1j}^*,\eta_j^*)$ for $j\in[k_*]$, we get that $\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j=1}^{k_*}\Big|\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)-\exp(\beta_{0j}^*)\Big|\to 0$;
- $x^{(u)}F(x;\beta_{1j}^*,\eta_j^*)$ for $u\in[d]$ and $j:|\mathcal{A}_j|=1$, we get that $\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\|\Delta\beta_{1ij}^n\|_1\to 0$;
- $[F_1(x;\beta_{1j}^*,\eta_j^*)]^{(u)}$ for $u\in[d]$ and $j:|\mathcal{A}_j|=1$, we get that $\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\|\Delta\eta_{ij}^n\|_1\to 0$;
- $[x^{(u)}]^2F(x;\beta_{1j}^*,\eta_j^*)$ for $u\in[d]$ and $j:|\mathcal{A}_j|>1$, we get that $\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\|\Delta\beta_{1ij}^n\|^2\to 0$;
- $[F_2(x;\beta_{1j}^*,\eta_j^*)]^{(uu)}$ for $u\in[d]$ and $j:|\mathcal{A}_j|>1$, we get that $\frac{1}{\mathcal{D}_{1n}}\cdot\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\|\Delta\eta_{ij}^n\|^2\to 0$;

By taking the summation of the above limits, we obtain that $1=\mathcal{D}_{1n}/\mathcal{D}_{1n}\to 0$ as $n\to\infty$, which is a contradiction. Therefore, not all the coefficients in the representations of $[A_n(x)-R_1(x)-R_2(x)]/\mathcal{D}_{1n}$, $[B_n(x)-R_3(x)-R_4(x)]/\mathcal{D}_{1n}$ and $E_n(x)/\mathcal{D}_{1n}$ go to zero.

Step 3. In this step, we point out a contradiction following from the result in Step 2. Let us denote by m_n the maximum of the absolute values of the coefficients in the representations of $[A_n(x)-R_1(x)-R_2(x)]/\mathcal{D}_{1n}$, $[B_n(x)-R_3(x)-R_4(x)]/\mathcal{D}_{1n}$ and $E_n(x)/\mathcal{D}_{1n}$. Since at least one among those coefficients does not approach zero, we obtain that $1/m_n\not\to\infty$.

Recall the hypothesis in equation (24) that $\|f_{G_n} - f_{G_*}\|_{L^2(\mu)}/\mathcal{D}_{1n} \to 0$ as $n \to \infty$, which indicates that $\|f_{G_n} - f_{G_*}\|_{L^1(\mu)}/\mathcal{D}_{1n} \to 0$. By means of the Fatou's lemma, we have

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

This result implies that $[f_{G_n}(x) - f_{G_*}(x)]/[m_n\mathcal{D}_{1n}] \to 0$ for almost every x. Since the term $\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top x + \beta_{0j}^*)$ is bounded, we deduce that $Q_n(x)/[m_n\mathcal{D}_{1n}] \to 0$, or equivalently,

$$\frac{1}{m_n\mathcal{D}_{1n}} \cdot \Big[(A_{n,1}(x) - R_1(x) + A_{n,2}(x) - R_2(x)) - (B_{n,1}(x) - R_3(x) + B_{n,2}(x) - R_4(x)) + E_n(x)\Big] \to 0. \tag{31}$$

Let us denote

$$\frac{1}{m_n\mathcal{D}_{1n}} \cdot \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)(\Delta\beta_{1ij}^n) \to \phi_{1,j}, \quad \frac{1}{m_n\mathcal{D}_{1n}} \cdot \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)(\Delta\beta_{1ij}^n)(\Delta\beta_{1ij}^n)^\top \to \phi_{2,j},$$
$$\frac{1}{m_n\mathcal{D}_{1n}} \cdot \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)(\Delta\eta_{ij}^n) \to \varphi_{1,j}, \quad \frac{1}{m_n\mathcal{D}_{1n}} \cdot \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)(\Delta\eta_{ij}^n)(\Delta\eta_{ij}^n)^\top \to \varphi_{2,j},$$
$$\frac{1}{m_n\mathcal{D}_{1n}} \cdot \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)(\Delta\beta_{1ij}^n)(\Delta\eta_{ij}^n)^\top \to \zeta_j, \quad \frac{1}{m_n\mathcal{D}_{1n}} \cdot \Big(\sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^*)\Big) \to \xi_j \quad .$$

Here, at least one among $\phi_{1,j}^{(u)}$, $\phi_{2,j}^{(uu)}$, $\varphi_{1,j}^{(u)}$, $\varphi_{2,j}^{(uu)}$ and ξ_j, for $j \in [k_*]$, is different from zero, which results from Step 2. Additionally, let us denote $F_{\tau j} := F_\tau(x; \beta_{1j}^*, \eta_j^*)$ and $H_j = H(x; \beta_{1j}^*)$ for short, then from the formulation of

- $A_{n,1}$ in equation (27), we get

$$\frac{A_{n,1} - R_1(x)}{m_n\mathcal{D}_{1n}} \to \sum_{j:|\mathcal{A}_j|=1} \Big[\phi_{1,j}^\top x \cdot F_j + \varphi_{1,j}^\top F_{1j}\Big]. \tag{32}$$

- $A_{n,2}$ in equation (28), we get

$$\frac{A_{n,2} - R_2(x)}{m_n\mathcal{D}_{2n}} \to \sum_{j:|\mathcal{A}_j|>1} \Bigg\{\Big[\phi_{1,j}^\top x + x^\top\Big(M_d \odot \phi_{2,j}\Big)x\Big] \cdot F_j + [\varphi_{1,j}^\top + x^\top\zeta_j] \cdot F_{1j} + \Big[M_d \odot \varphi_{2,j}\Big] \odot F_{2j}\Bigg\}. \tag{33}$$

- $B_{n,1}$ in equation (29), we get

$$\frac{B_{n,1} - R_3(x)}{m_n\mathcal{D}_{2n}} \to \sum_{j:|\mathcal{A}_j|=1} [\phi_{1,j}^\top x \cdot H_j]. \tag{34}$$

- $B_{n,2}$ in equation (30), we get

$$\frac{B_{n,2} - R_4(x)}{m_n\mathcal{D}_{2n}} \to \sum_{j:|\mathcal{A}_j|>1} \Big[\phi_{1,j}^\top x + x^\top\Big(M_d \odot \phi_{2,j}\Big)x\Big] \cdot H_j. \tag{35}$$

- $E_n(x)$ in equation (26), we get

$$\frac{E_n(x)}{m_n\mathcal{D}_{2n}} \to \sum_{j=1}^{k_*} \xi_j[F_j - H_j]. \tag{36}$$

Due to the result in equation (31), we deduce that the limits in equations (32), (33), (34), (35) and (36) sum up to zero.

Now, we show that 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 equal to zero. For that purpose, we first denote $J_1, J_2, \ldots, J_\ell$ as the partition of the set $\{\exp((\beta_{1j}^*)^\top x) : j \in [k_*]\}$ for some $\ell \leq k_*$ such that

(i) $\beta^*_{1j} = \beta^*_{1j'}$ for any $j, j' \in J_i$ and $i \in [\ell]$;

(ii) $\beta^*_{1j} \neq \beta^*_{1j'}$ when j and j' do not belong to the same set J_i for any $i \in [\ell]$.

Then, the set $\{\exp((\beta^*_{1j_1})^\top x), \ldots, \exp((\beta^*_{1j_\ell})^\top x)\}$, where $j_i \in J_i$, is linearly independent. Since the limits in equations (32), (33), (34), (35) and (36) sum up to zero, we get for any $i \in [\ell]$ that

$$\sum_{j \in J_i : |\mathcal{A}_j| = 1} \Big[(\xi_j + \phi_{1,j}^\top x) \cdot h_j + \varphi_{1,j}^\top h_{1j}\Big] + \sum_{j \in J_i : |\mathcal{A}_j| > 1} \Bigg\{ \Big[\phi_{1,j}^\top x + x^\top \Big(M_d \odot \phi_{2,j}\Big) x\Big] \cdot h_j$$
$$+ [\varphi_{1,j}^\top + x^\top \zeta_j] \cdot h_{1j} + \Big[M_d \odot \varphi_{2,j}\Big] \odot h_{2j} \Bigg\} - \sum_{j \in J_i : |\mathcal{A}_j| = 1} [(\phi_{1,j}^\top x + \xi_j) \cdot f_{G_*}(x)]$$
$$- \sum_{j \in J_i : |\mathcal{A}_j| > 1} \Big[\xi_j + \phi_{1,j}^\top x + x^\top \Big(M_d \odot \phi_{2,j}\Big) x\Big] \cdot f_{G_*}(x) = 0,$$

where we denote $h_j := h(x, \eta^*_j)$, $h_{1j} := \frac{\partial h}{\partial \eta}(x, \eta^*_j)$ and $h_{2j} := \frac{\partial^2 h}{\partial \eta \partial \eta^\top}(x, \eta^*_j)$. Recall that the expert function h satisfies conditions in Definition 3.1, then the following set is linearly independent

$$\Big\{ x^\nu \cdot \frac{\partial^{|\tau_1| + |\tau_2|} h}{\partial \eta^{\tau_1} \partial \eta^{\tau_2}}(x, \eta^*_j),\ x^\nu \cdot f_{G_*}(x) : \nu \in \mathbb{N}^d,\ \tau_1, \tau_2 \in \mathbb{N}^q,\ 0 \leq |\nu| + |\tau_1| + |\tau_2| \leq 2,\ j \in [k_*] \Big\}.$$

is linearly independent. Therefore, we obtain 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 $j \in J_i$ and $i \in [\ell]$. In other words, those results hold true for any $j \in [k_*]$, which contradicts to the fact that at least one among $\phi_{1,j}^{(u)}, \phi_{2,j}^{(uu)}, \varphi_{1,j}^{(u)}, \varphi_{2,j}^{(uu)}$ and ξ_j, for $j \in [k_*]$, is different from zero. Thus, we achieve the inequality (23), i.e.

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

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

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

Global part: Given the above result, it suffices to demonstrate that

$$\inf_{G \in \mathcal{G}_k(\Theta) : \mathcal{D}_1(G, G_*) > \varepsilon'} \|f_G - f_{G_*}\|_{L^2(\mu)} / \mathcal{D}_1(G, G_*) > 0. \tag{37}$$

Assume by contrary that the inequality (37) does not hold true, then we can find a sequence of mixing measures $G'_n \in \mathcal{G}_k(\Theta)$ such that $\mathcal{D}_1(G'_n, G_*) > \varepsilon'$ and

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

which indicates that $\|f_{G'_n} - f_{G_*}\|_{L^2(\mu)} \to 0$ as $n \to \infty$. Recall that Θ is a compact set, therefore, we can replace the sequence G'_n by one of its subsequences that converges to a mixing measure $G' \in \mathcal{G}_k(\Omega)$. Since $\mathcal{D}_1(G'_n, G_*) > \varepsilon'$, we deduce that $\mathcal{D}_1(G', G_*) > \varepsilon'$.

Next, by invoking the Fatou's lemma, we have that

$$0 = \lim_{n \to \infty} \|f_{G'_n} - f_{G_*}\|^2_{L^2(\mu)} \geq \int \liminf_{n \to \infty} \Big| f_{G'_n}(x) - f_{G_*}(x) \Big|^2 \mathrm{d}\mu(x).$$

Thus, we get that $f_{G'}(x) = f_{G_*}(x)$ for almost every x. From Proposition B.1, we deduce that $G' \equiv G_*$. Consequently, it follows that $\mathcal{D}_1(G', G_*) = 0$, contradicting the fact that $\mathcal{D}_1(G', G_*) > \varepsilon' > 0$.

Hence, the proof is completed.

A.3. Proof of Theorem 4.2

In this proof, we focus on demonstrating the following inequality:

$$\inf_{G\in\mathcal{G}_k(\Theta)} \|f_G - f_{G_*}\|_{L^2(\mu)}/\mathcal{D}_2(G,G_*) > 0. \tag{38}$$

To this end, we divide the proof of the above inequality into local and global parts in the sequel.

Local part: In this part, we show that

$$\lim_{\varepsilon\to 0} \inf_{G\in\mathcal{G}_k(\Theta):\mathcal{D}_2(G,G_*)\leq\varepsilon} \|f_G - f_{G_*}\|_{L^2(\mu)}/\mathcal{D}_2(G,G_*) > 0. \tag{39}$$

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

$$\|f_{G_n} - f_{G_*}\|_{L^2(\mu)}/\mathcal{D}_{2n} \to 0, \tag{40}$$

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 assymptotic, 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 $\mathcal{D}_{2n}$ can be represented as

$$\begin{aligned}\mathcal{D}_{2n} := \sum_{j=1}^{k_*}\Big|\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n) - \exp(\beta_{0j}^*)\Big| + \sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\Big[\|\Delta\beta_{1ij}^n\|^2 + \|\Delta a_{ij}^n\|^2 + |\Delta b_{ij}^n|^2\Big] \\ + \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i})\Big[\|\Delta\beta_{1ij}^n\| + \|\Delta a_{ij}^n\| + |\Delta b_{ij}^n|\Big],\end{aligned} \tag{41}$$

where we denote $\Delta\beta_{1ij}^n := \beta_{1i}^n - \beta_{1j}^*$, $\Delta a_{ij}^n := a_i^n - a_j^*$ and $\Delta b_{ij}^n := b_i^n - b_j^*$.

Since $\mathcal{D}_{2n}\to 0$, we get that $(\beta_{1i}^n, a_i^n, b_i^n) \to (\beta_{1j}^*, a_j^*, b_j^*)$ and $\exp(\beta_{0i}^n)\to\exp(\beta_{0j}^*)$ as $n\to\infty$ for any $i\in\mathcal{A}_j$ and $j\in[k_*]$. Now, we divide the proof of local part into three steps as follows:

Step 1. In this step, we decompose the term $Q_n(x) := [\sum_{j=1}^{k_*}\exp((\beta_{1j}^*)^\top x + \beta_{0j}^*)]\cdot[f_{G_n}(x) - f_{G_*}(x)]$ into a combination of linearly independent elements using Taylor expansion. In particular, let us denote $F(x;\beta_1,a,b) := \exp(\beta_1^\top x)\sigma(a^\top x + b)$ and $H(x;\beta_1) = \exp(\beta_1^\top x) f_{G_n}(x)$, then we have

$$\begin{aligned}Q_n(x) &= \sum_{j=1}^{k_*}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\Big[F(x;\beta_{1i}^n,a_i^n,b_i^n) - F(x;\beta_{1j}^*,a_j^*,b_j^*)\Big] \\ &- \sum_{j=1}^{k_*}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\Big[H(x;\beta_{1i}^n) - H(x;\beta_{1j}^*)\Big] \\ &+ \sum_{j=1}^{k_*}\Big(\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n) - \exp(\beta_{0j}^*)\Big)\Big[F(x;\beta_{1j}^*,a_j^*,b_j^*) - H(x;\beta_{1j}^*)\Big] \\ &:= A_n(x) - B_n(x) + E_n(x).\end{aligned} \tag{42}$$

Decomposition of $A_n(x)$. Next, we continue to separate the term $A_n(x)$ into two parts as follows:

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

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

$$A_{n,1} = \sum_{j:|\mathcal{A}_j|=1} \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n) \sum_{|\alpha|=1} (\Delta\beta_{1ij}^n)^{\alpha_1} (\Delta a_{ij}^n)^{\alpha_2} (\Delta b_{ij}^n)^{\alpha_3} \cdot \frac{\partial F}{\partial\beta_1^{\alpha_1}\partial a^{\alpha_2}\partial b^{\alpha_3}}(x;\beta_{1j}^*,a_j^*,b_j^*) + R_1(x),$$

where $R_1(x)$ is a Taylor remainder such that $R_1(x)/\mathcal{D}_{2n} \to 0$ as $n\to\infty$. By taking the first derivatives of F w.r.t its parameters, we get

$$\begin{aligned}
\frac{\partial F}{\partial \beta_1}(x;\beta_{1j}^*,a_j^*,b_j^*) &= x\exp((\beta_{1j}^*)^\top x)\cdot\sigma((a_j^*)^\top x + b_j^*) = x\cdot F(x;\beta_{1j}^*,a_j^*,b_j^*),\\
\frac{\partial F}{\partial a}(x;\beta_{1j}^*,a_j^*,b_j^*) &= x\exp((\beta_{1j}^*)^\top x)\cdot\sigma^{(1)}((a_j^*)^\top x + b_j^*) = x\cdot F_1(x;\beta_{1j}^*,a_j^*,b_j^*),\\
\frac{\partial F}{\partial b}(x;\beta_{1j}^*,a_j^*,b_j^*) &= \exp((\beta_{1j}^*)^\top x)\cdot\sigma^{(1)}((a_j^*)^\top x + b_j^*) = F_1(x;\beta_{1j}^*,a_j^*,b_j^*),
\end{aligned}$$

where we denote $F_\tau(x;\beta_{1j}^*,a_j^*,b_j^*) = \exp((\beta_{1j}^*)^\top x)\cdot\sigma^{(\tau)}((a_j^*)^\top x + b_j^*)$. Thus, we can rewrite $A_{n,1}$ as

$$A_{n,1} = \sum_{j:|\mathcal{A}_j|=1} C_{n,1,j}(x) + R_1(x), \tag{43}$$

where

$$C_{n,1,j}(x) = \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)\Big[(\Delta\beta_{1ij}^n)^\top x\cdot F(x;\beta_{1j}^*,a_j^*,b_j^*) + ((\Delta a_{ij}^n)^\top x + (\Delta b_{ij}^n))\cdot F_1(x;\beta_{1j}^*,a_j^*,b_j^*)\Big].$$

Next, by applying the second-order Taylor expansion, $A_{n,2}$ can be represented as

$$A_{n,2} = \sum_{j:|\mathcal{A}_j|>1} \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n) \sum_{|\alpha|=1}^{2} \frac{1}{\alpha!} (\Delta\beta_{1ij}^n)^{\alpha_1} (\Delta a_{ij}^n)^{\alpha_2} (\Delta b_{ij}^n)^{\alpha_3} \cdot \frac{\partial^{|\alpha_1|+|\alpha_2|+\alpha_3} F}{\partial\beta_1^{\alpha_1}\partial a^{\alpha_2}\partial b^{\alpha_3}}(x;\beta_{1j}^*,a_j^*,b_j^*) + R_2(x),$$

where $R_2(x)$ is a Taylor remainder such that $R_2(x)/\mathcal{D}_{2n}\to 0$ as $n\to\infty$. The second derivatives of F w.r.t its parameters are given by

$$\begin{aligned}
\frac{\partial^2 F}{\partial\beta\partial\beta^\top}(x;\beta_{1j}^*,a_j^*,b_j^*) = xx^\top\cdot F(x;\beta_{1j}^*,a_j^*,b_j^*), &\quad \frac{\partial^2 F}{\partial\beta_1\partial a^\top} = xx^\top\cdot F_1(x;\beta_{1j}^*,a_j^*,b_j^*),\\
\frac{\partial^2 F}{\partial\beta_1\partial b} = x\cdot F_1(x;\beta_{1j}^*,a_j^*,b_j^*), &\quad \frac{\partial^2 F}{\partial a\partial a^\top} = xx^\top\cdot F_2(x,\beta_{1j}^*,a_j^*,b_j^*),\\
\frac{\partial^2 F}{\partial a\partial b} = x\cdot F_2(x,\beta_{1j}^*,a_j^*,b_j^*), &\quad \frac{\partial^2 F}{\partial b^2} = F_2(x,\beta_{1j}^*,a_j^*,b_j^*).
\end{aligned}$$

Therefore, the term $A_{n,2}$ becomes

$$A_{n,2} = \sum_{j:|\mathcal{A}_j|>1} [C_{n,1,j}(x) + C_{n,2,j}(x)] + R_2(x), \tag{44}$$

where

$$\begin{aligned}
C_{n,2,j}(x) := \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)\Bigg\{ &\Big[x^\top\Big(M_d\odot(\Delta\beta_{1ij}^n)(\Delta\beta_{1ij}^n)^\top\Big)x\Big]\cdot F(x;\beta_{1j}^*,a_j^*,b_j^*)\\
+ \Big[x^\top\Big(M_d\odot(\Delta a_{ij}^n)(\Delta a_{ij}^n)^\top\Big)x + (\Delta b_{ij}^n)(\Delta\beta_{1ij}^n)^\top x &+ x^\top(\Delta\beta_{1ij}^n)(\Delta a_{ij}^n)^\top x\Big]\cdot F_1(x;\beta_{1j}^*,a_j^*,b_j^*)\\
&+ \Big[\frac{1}{2}(\Delta b_{ij}^n)^2 + (\Delta b_{ij}^n)(\Delta a_{ij}^n)^\top x\Big]\cdot F_2(x;\beta_{1j}^*,a_j^*,b_j^*)\Bigg\}.
\end{aligned}$$

Decomposition of $B_n(x)$. Subsequently, we also divide $B_n(x)$ into two terms based on the Voronoi cells as

$$B_n(x) = \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\Big[H(x;\beta_{1i}^n)-H(x;\beta_{1j}^*)\Big]$$
$$+\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\Big[H(x;\beta_{1i}^n)-H(x;\beta_{1j}^*)\Big]$$
$$:= B_{n,1}+B_{n,2}.$$

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

$$B_{n,1}=\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)(\Delta\beta_{1ij}^n)^\top x\cdot H(x;\beta_{1j}^*)+R_3(x), \tag{45}$$

where $R_3(x)$ is a Taylor remainder such that $R_3(x)/\mathcal{D}_{2n}\to 0$ as $n\to\infty$. Meanwhile, by applying the second-order Taylor expansion, we get

$$B_{n,2}=\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\Big[(\Delta\beta_{1ij}^n)^\top x+x^\top\Big(M_d\odot(\Delta\beta_{1ij}^n)(\Delta\beta_{1ij}^n)^\top\Big)x\Big]\cdot H(x;\beta_{1j}^*)+R_4(x), \tag{46}$$

where $R_4(x)$ is a Taylor remainder such that $R_4(x)/\mathcal{D}_{2n}\to 0$ as $n\to\infty$.

Putting the above results together, we see that $[A_n(x)-R_1(x)-R_2(x)]/\mathcal{D}_{2n}$, $[B_n(x)-R_3(x)-R_4(x)]/\mathcal{D}_{2n}$ and $E_n(x)/\mathcal{D}_{2n}$ can be written as a combination of elements from set $\mathcal{S}:=\cup_{\tau=0}^3\mathcal{S}_\tau$ in which

$$\mathcal{S}_0:=\Big\{F(x;\beta_{1j}^*,a_j^*,b_j^*),\ x^{(u)}F(x;\beta_{1j}^*,a_j^*,b_j^*),\ x^{(u)}x^{(v)}F(x;\beta_{1j}^*,a_j^*,b_j^*):u,v\in[d],\ j\in[k_*]\Big\},$$
$$\mathcal{S}_1:=\Big\{F_1(x;\beta_{1j}^*,a_j^*,b_j^*),\ x^{(u)}F_1(x;\beta_{1j}^*,a_j^*,b_j^*),\ x^{(u)}x^{(v)}F_1(x;\beta_{1j}^*,a_j^*,b_j^*):u,v\in[d],\ j\in[k_*]\Big\},$$
$$\mathcal{S}_2:=\Big\{F_2(x;\beta_{1j}^*,a_j^*,b_j^*),\ x^{(u)}F_2(x;\beta_{1j}^*,a_j^*,b_j^*),\ x^{(u)}x^{(v)}F_2(x;\beta_{1j}^*,a_j^*,b_j^*):u,v\in[d],\ j\in[k_*]\Big\},$$
$$\mathcal{S}_3:=\Big\{H(x;\beta_{1j}^*),\ x^{(u)}H(x;\beta_{1j}^*),\ x^{(u)}x^{(v)}H(x;\beta_{1j}^*):u,v\in[d],\ j\in[k_*]\Big\}.$$

Step 2. In this step, we prove by contradiction that at least one among coefficients in the representations of $[A_n(x)-R_1(x)-R_2(x)]/\mathcal{D}_{2n}$, $[B_n(x)-R_3(x)-R_4(x)]/\mathcal{D}_{2n}$ and $E_n(x)/\mathcal{D}_{2n}$ does not go to zero as n tends to infinity. Indeed, assume that all of them converge to zero. Then, by considering the coefficients of

- $F(x;\beta_{1j}^*,a_j^*,b_j^*)$ for $j\in[k_*]$, we get that $\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{j=1}^{k_*}\Big|\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)-\exp(\beta_{0j}^*)\Big|\to 0$;
- $x^{(u)}F(x;\beta_{1j}^*,a_j^*,b_j^*)$ for $u\in[d]$ and $j:|\mathcal{A}_j|=1$, we get that $\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\|\Delta\beta_{1ij}^n\|_1\to 0$;
- $x^{(u)}F_1(x;\beta_{1j}^*,a_j^*,b_j^*)$ for $u\in[d]$ and $j:|\mathcal{A}_j|=1$, we get that $\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\|\Delta a_{ij}^n\|_1\to 0$;
- $F_1(x;\beta_{1j}^*,a_j^*,b_j^*)$ for $j:|\mathcal{A}_j|=1$, we get that $\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)|\Delta b_{ij}^n|_1\to 0$;
- $[x^{(u)}]^2F(x;\beta_{1j}^*,a_j^*,b_j^*)$ for $u\in[d]$ and $j:|\mathcal{A}_j|>1$, we get that $\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\|\Delta\beta_{1ij}^n\|^2\to 0$;
- $[x^{(u)}]^2F_2(x;\beta_{1j}^*,a_j^*,b_j^*)$ for $u\in[d]$ and $j:|\mathcal{A}_j|>1$, we get that $\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\|\Delta a_{ij}^n\|^2\to 0$;
- $F_2(x;\beta_{1j}^*,a_j^*,b_j^*)$ for $j:|\mathcal{A}_j|>1$, we get that $\frac{1}{\mathcal{D}_{2n}}\cdot\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\beta_{0i}^n)\|\Delta b_{ij}^n\|^2\to 0$.

By taking the summation of the above limits, we obtain that $1=\mathcal{D}_{2n}/\mathcal{D}_{2n}\to 0$ as $n\to\infty$, which is a contradiction. Therefore, not all the coefficients in the representations of $[A_n(x)-R_1(x)-R_2(x)]/\mathcal{D}_{2n}$, $[B_n(x)-R_3(x)-R_4(x)]/\mathcal{D}_{2n}$ and $E_n(x)/\mathcal{D}_{2n}$ go to zero.

Step 3. In this step, we point out a contradiction following from the result in Step 2. Let us denote by m_n the maximum of the absolute values of the coefficients in the representations of $[A_n(x)-R_1(x)-R_2(x)]/\mathcal{D}_{2n}$, $[B_n(x)-R_3(x)-R_4(x)]/\mathcal{D}_{2n}$ and $E_n(x)/\mathcal{D}_{2n}$. Since at least one among those coefficients does not approach zero, we obtain that $1/m_n\not\to\infty$.

Recall the hypothesis in equation (40) that $\|f_{G_n} - f_{G_*}\|_{L^2(\mu)}/\mathcal{D}_{2n} \to 0$ as $n \to \infty$, which indicates that $\|f_{G_n} - f_{G_*}\|_{L^1(\mu)}/\mathcal{D}_{2n} \to 0$. By means of the Fatou's lemma, we have

$$0 = \lim_{n\to\infty} \frac{\|f_{G_n} - f_{G_*}\|_{L^1(\mu)}}{m_n\mathcal{D}_{2n}} \geq \int \liminf_{n\to\infty} \frac{|f_{G_n}(x) - f_{G_*}(x)|}{m_n\mathcal{D}_{2n}} \mathrm{d}\mu(x) \geq 0.$$

This result implies that $[f_{G_n}(x) - f_{G_*}(x)]/[m_n\mathcal{D}_{2n}]$ for almost every x. Since the term $\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top x + \beta_{0j}^*)$ is bounded, we deduce that $Q_n(x)/[m_n\mathcal{D}_{2n}] \to 0$, or equivalently,

$$\lim_{n\to\infty} \frac{1}{m_n\mathcal{D}_{2n}} \cdot \Big[(A_{n,1} - R_1(x) + A_{n,2} - R_2(x)) - (B_{n,1} - R_3(x) + B_{n,2} - R_4(x)) + E_n(x)\Big] \to 0. \tag{47}$$

Let us denote

$$\begin{aligned}
\frac{1}{m_n\mathcal{D}_{2n}} \cdot \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)(\Delta\beta_{1ij}^n) \to \phi_{1,j}, &\quad \frac{1}{m_n\mathcal{D}_{2n}} \cdot \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)(\Delta\beta_{1ij}^n)(\Delta\beta_{1ij}^n)^\top \to \phi_{2,j},\\
\frac{1}{m_n\mathcal{D}_{2n}} \cdot \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)(\Delta a_{ij}^n) \to \varphi_{1,j}, &\quad \frac{1}{m_n\mathcal{D}_{2n}} \cdot \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)(\Delta a_{ij}^n)(\Delta a_{ij}^n)^\top \to \varphi_{2,j},\\
\frac{1}{m_n\mathcal{D}_{2n}} \cdot \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)(\Delta b_{ij}^n) \to \kappa_{1,j}, &\quad \frac{1}{m_n\mathcal{D}_{2n}} \cdot \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)(\Delta b_{ij}^n)^2 \to \kappa_{2,j},\\
\frac{1}{m_n\mathcal{D}_{2n}} \cdot \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)(\Delta\beta_{1ij}^n)(\Delta a_{ij}^n)^\top \to \zeta_{1,j}, &\quad \frac{1}{m_n\mathcal{D}_{2n}} \cdot \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)(\Delta b_{ij}^n)(\Delta\beta_{1ij}^n) \to \zeta_{2,j},\\
\frac{1}{m_n\mathcal{D}_{2n}} \cdot \sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n)(\Delta b_{ij}^n)(\Delta a_{ij}^n) \to \zeta_{3,j}, &\quad \frac{1}{m_n\mathcal{D}_{2n}} \cdot \Big(\sum_{i\in\mathcal{A}_j} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^*)\Big) \to \xi_j .
\end{aligned}$$

Here, at least one among $\phi_{1,j}^{(u)}, \phi_{2,j}^{(uu)}, \varphi_{1,j}^{(u)}, \varphi_{2,j}^{(uu)}, \kappa_{1,j}, \kappa_{2,j}$ and ξ_j, for $j \in [k_*]$, is different from zero, which results from Step 2. Additionally, let us denote $F_{\tau j} := F_\tau(x; \beta_{1j}^*, a_j^*, b_j^*)$ and $H_j = H(x; \beta_{1j}^*)$ for short, then from the formulation of

- $A_{n,1}$ in equation (43), we get

$$\frac{A_{n,1} - R_1(x)}{m_n\mathcal{D}_{2n}} \to \sum_{j:|\mathcal{A}_j|=1} \Big[\phi_{1,j}^\top x \cdot F_j + (\kappa_{1,j} + \varphi_{1,j}^\top x) \cdot F_{1j}\Big]. \tag{48}$$

- $A_{n,2}$ in equation (44), we get

$$\begin{aligned}
\frac{A_{n,2} - R_2(x)}{m_n\mathcal{D}_{2n}} \to \sum_{j:|\mathcal{A}_j|>1} \bigg\{ \Big[\phi_{1,j}^\top x + x^\top \Big(M_d \odot \phi_{2,j}\Big) x\Big] \cdot F_j + [\kappa_{1,j} + (\varphi_{1,j} + \zeta_{2,j})^\top x + x^\top \zeta_{1,j} x] \cdot F_{1j}\\
+ \Big[\frac{1}{2}\kappa_{2,j} + \zeta_{3,j}^\top x + x^\top \Big(M_d \odot \varphi_{2,j}\Big) x\Big] \cdot F_{2j}
\end{aligned} \tag{49}$$

- $B_{n,1}$ in equation (45), we get

$$\frac{B_{n,1} - R_3(x)}{m_n\mathcal{D}_{2n}} \to \sum_{j:|\mathcal{A}_j|=1} [\phi_{1,j}^\top x \cdot H_j]. \tag{50}$$

- $B_{n,2}$ in equation (46), we get

$$\frac{B_{n,2} - R_4(x)}{m_n\mathcal{D}_{2n}} \to \sum_{j:|\mathcal{A}_j|>1} \Big[\phi_{1,j}^\top x + x^\top \Big(M_d \odot \phi_{2,j}\Big) x\Big] \cdot H_j. \tag{51}$$

- $E_n(x)$ in equation (42), we get

$$\frac{E_n(x)}{m_n\mathcal{D}_{2n}} \to \sum_{j=1}^{k_*}\xi_j[F_j - H_j]. \tag{52}$$

Due to the result in equation (47), we deduce that the limits in equations (48), (49), (50), (51) and (52) sum up to zero.

Now, we show that all the values of $\phi_{1,j}^{(u)}$, $\phi_{2,j}^{(uu)}$, $\varphi_{1,j}^{(u)}$, $\varphi_{2,j}^{(uu)}$, $\kappa_{1,j}$, $\kappa_{2,j}$ and ξ_j, for $j\in[k_*]$, are equal to zero. For that purpose, we first denote $J_1, J_2, \ldots, J_\ell$ as the partition of the set $\{\exp((\beta_{1j}^*)^\top x): j\in[k_*]\}$ for some $\ell \leq k_*$ such that

(i) $\beta_{0j}^* = \beta_{0j'}^*$ for any $j,j'\in J_i$ and $i\in[\ell]$;

(ii) $\beta_{0j}^* \neq \beta_{0j'}^*$ when j and j' do not belong to the same set J_i for any $i\in[\ell]$.

Then, the set $\{\exp((\beta_{0j_1}^*)^\top x),\ldots,\exp((\beta_{0j_\ell}^*)^\top x)\}$, where $j_i\in J_i$, is linearly independent. Since the limits in equations (48), (49), (50), (51) and (52) sum up to zero, we get for any $i\in[\ell]$ that

$$\begin{aligned}
&\sum_{j\in J_i:|\mathcal{A}_j|=1}\Big[(\phi_{1,j}^\top x+\xi_j)\cdot\sigma_j + (\kappa_{1,j}+\varphi_{1,j}^\top x)\cdot\sigma_j^{(1)}\Big] + \sum_{j\in J_i:|\mathcal{A}_j|>1}\Bigg\{\Big[\xi_j+\phi_{1,j}^\top x + x^\top\Big(M_d\odot\phi_{2,j}\Big)x\Big]\cdot\sigma_j\\
&+[\kappa_{1,j}+(\varphi_{1,j}+\zeta_{2,j})^\top x + x^\top\zeta_{1,j}x]\cdot\sigma_j^{(1)} + \Big[\frac{1}{2}\kappa_{2,j}+\zeta_{3,j}^\top x + x^\top\Big(M_d\odot\varphi_{2,j}\Big)x\Big]\cdot\sigma_j^{(2)}\Bigg\}\\
&-\sum_{j\in J_i:|\mathcal{A}_j|=1}[(\phi_{1,j}^\top x+\xi_j)\cdot f_{G_*}(x)] - \sum_{j\in J_i:|\mathcal{A}_j|>1}\Big[\xi_j+\phi_{1,j}^\top x + x^\top\Big(M_d\odot\phi_{2,j}\Big)x\Big]\cdot f_{G_*}(x) = 0,
\end{aligned}$$

where we denote $\sigma_j^{(\tau)} := \sigma^{(\tau)}((a_j^*)^\top x + b_j^*)$. Additionally, as $(a_1^*,b_1^*),\ldots,(a_{k_*}^*,b_{k_*}^*)$ are pairwise distinct, the experts $(a_1^*)^\top x+b_1^*,\ldots,(a_{k_*}^*)^\top x + b_{k_*}^*$ are also pairwise distinct. Recall that the function σ satisfies conditions in Definition 4.1, then the following set is linearly independent

$$\Big\{x^\nu\sigma_j^{(\tau)}, x^\nu f_{G_*}(x): \nu\in\mathbb{N}^d, \tau\in\mathbb{N},\ 0\leq|\nu|,\tau\leq 2,\ j\in[k_*]\Big\}.$$

is linearly independent. Therefore, we obtain that $\kappa_{1,j}=\kappa_{2,j}=\xi_j=0$, $\phi_{1,j}=\varphi_{1,j}=\zeta_{2,j}=\zeta_{3,j}=0_d$ and $\phi_{2,j}=\varphi_{2,j}=\zeta_{1,j}=\mathbf{0}_{d\times d}$ for any $j\in J_i$ and $i\in[\ell]$. In other words, those results hold true for any $j\in[k_*]$, which contradicts to the fact that at least one among $\phi_{1,j}^{(u)}$, $\phi_{2,j}^{(uu)}$, $\varphi_{1,j}^{(u)}$, $\varphi_{2,j}^{(uu)}$, $\kappa_{1,j}$, $\kappa_{2,j}$ and ξ_j, for $j\in[k_*]$, is different from zero. Thus, we achieve the inequality (39), i.e.

$$\lim_{\varepsilon\to 0}\inf_{G\in\mathcal{G}_k(\Theta):\mathcal{D}_2(G,G_*)\leq\varepsilon}\|f_G - f_{G_*}\|_{L^2(\mu)}/\mathcal{D}_2(G,G_*) > 0.$$

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

$$\inf_{G\in\mathcal{G}_k(\Theta):\mathcal{D}_2(G,G_*)\leq\varepsilon'}\|f_G - f_{G_*}\|_{L^2(\mu)}/\mathcal{D}_2(G,G_*) > 0.$$

Global part: Given the above result, it suffices to demonstrate that

$$\inf_{G\in\mathcal{G}_k(\Theta):\mathcal{D}_2(G,G_*)>\varepsilon'}\|f_G - f_{G_*}\|_{L^2(\mu)}/\mathcal{D}_2(G,G_*) > 0. \tag{53}$$

Assume by contrary that the inequality (53) does not hold true, then we can find a sequence of mixing measures $G_n'\in\mathcal{G}_k(\Theta)$ such that $\mathcal{D}_2(G_n',G_*)>\varepsilon'$ and

$$\lim_{n\to\infty}\frac{\|f_{G_n'} - f_{G_*}\|_{L^2(\mu)}}{\mathcal{D}_2(G_n',G_*)} = 0,$$

which indicates that $\|f_{G'_n} - f_{G_*}\|_{L^2(\mu)} \to 0$ as $n \to \infty$. Recall that Θ is a compact set, therefore, we can replace the sequence G'_n by one of its subsequences that converges to a mixing measure $G' \in \mathcal{G}_k(\Omega)$. Since $\mathcal{D}_2(G'_n, G_*) > \varepsilon'$, we deduce that $\mathcal{D}_2(G', G_*) > \varepsilon'$.

Next, by invoking the Fatou's lemma, we have that

$$0 = \lim_{n\to\infty} \|f_{G'_n} - f_{G_*}\|^2_{L^2(\mu)} \geq \int \liminf_{n\to\infty} \Big|f_{G'_n}(x) - f_{G_*}(x)\Big|^2 \mathrm{d}\mu(x).$$

Thus, we get that $f_{G'}(x) = f_{G_*}(x)$ for almost every x. From Proposition B.1, we deduce that $G' \equiv G_*$. Consequently, it follows that $\mathcal{D}_2(G', G_*) = 0$, contradicting the fact that $\mathcal{D}_2(G', G_*) > \varepsilon' > 0$.

Hence, the proof is completed.

A.4. Proof of Proposition 4.3

It is sufficient to show that the following limit holds true for any $r \geq 1$:

$$\lim_{\varepsilon\to 0} \inf_{G\in\mathcal{G}_k(\Theta):\mathcal{D}_{3,r}(G,G_*)\leq\varepsilon} \frac{\|f_G - f_{G_*}\|_{L^2(\mu)}}{\mathcal{D}_{3,r}(G, G_*)} = 0. \tag{54}$$

To this end, we need to construct a sequence of mixing measures (G_n) that satisfies $\mathcal{D}_{3,r}(G_n, G_*) \to 0$ and

$$\frac{\|f_{G_n} - f_{G_*}\|_{L^2(\mu)}}{\mathcal{D}_{3,r}(G_n, G_*)} \to 0,$$

as $n \to \infty$. Recall that under the Regime 2, at least one among parameters $a^*_1, \ldots, a^*_{k_*}$ is equal to 0_d. Without loss of generality, we may assume that $a^*_1 = 0_d$. Next, let us take into account the sequence $G_n = \sum_{i=1}^{k_*+1} \exp(\beta^n_{0i})\delta_{(\beta^n_{1i}, a^n_i, b^n_i)}$ in which

- $\exp(\beta^n_{01}) = \exp(\beta^n_{02}) = \frac{1}{2}\exp(\beta^*_{01})$ and $\exp(\beta^n_{0i}) = \exp(\beta^*_{0(i-1)})$ for any $3 \leq i \leq k_* + 1$;
- $\beta^n_{11} = \beta^n_{12} = \beta^*_{11}$ and $\beta^n_{1i} = \beta^*_{1(i-1)}$ for any $3 \leq i \leq k_* + 1$;
- $a^n_1 = a^n_2 = a^*_1 = 0_d$ and $a^n_i = a^*_{i-1}$ for any $3 \leq i \leq k_* + 1$;
- $b^n_1 = b^*_1 + \frac{c}{n}$, $b^n_2 = b^*_1 + \frac{2c}{n}$ and $b^n_i = b^*_{i-1}$ for any $3 \leq i \leq k_* + 1$,

where $c \in \mathbb{R}$ will be chosen later. Consequently, we get that

$$\mathcal{D}_{3,r}(G_n, G_*) = \frac{1}{2}\exp(\beta^*_{01})\Big[\frac{c^r}{n^r} + \frac{(2c)^r}{n^r}\Big] = \mathcal{O}(n^{-r}).$$

Next, we demonstrate that $\|f_{G_n} - f_{G_*}\|_{L^2(\mu)}/\mathcal{D}_{3,r}(G_n, G_*) \to 0$. To this end, consider the quantity $Q_n(x) := [\sum_{j=1}^{k_*} \exp((\beta^*_{1j})^\top x + \beta^*_{0j})] \cdot [f_{G_n}(x) - f_{G_*}(x)]$, and decompose it as follows:

$$\begin{aligned}
Q_n(x) = &\sum_{j=1}^{k_*}\sum_{i\in\mathcal{A}_j} \exp(\beta^n_{0i})\Big[\exp((\beta^n_{1i})^\top x)\sigma((a^n_i)^\top x + b^n_i) - \exp((\beta^*_{1j})^\top x)\sigma((a^*_j)^\top x + b^*_j)\Big] \\
&- \sum_{j=1}^{k_*}\sum_{i\in\mathcal{A}_j} \exp(\beta^n_{0i})\Big[\exp((\beta^n_{1i})^\top x) f_{G_n}(x) - \exp((\beta^*_{1j})^\top x) f_{G_n}(x)\Big] \\
&+ \sum_{j=1}^{k_*}\Big(\sum_{i\in\mathcal{A}_j} \exp(\beta^n_{0i}) - \exp(\beta^*_{0j})\Big)\Big[\exp((\beta^*_{1j})^\top x)\sigma((a^*_j)^\top x + b^*_j) - \exp((\beta^*_{1j})^\top x) f_{G_n}(x)\Big] \\
&:= A_n(x) - B_n(x) + E_n(x).
\end{aligned}$$

From the definitions of β^n_{1i}, a^n_i and b^n_i, we can verify that $B_n(x) = E_n(x) = 0$. Additionally, we can represent $A_n(x)$ as

$$A_n(x) = \sum_{i=1}^{2} \exp(\beta^*_{01}) \exp((\beta^*_{11})^\top x) \Big[\sigma(b^n_i) - \sigma(b^*_1)\Big].$$

When r is odd: By applying the Taylor expansion up to order r-th, we get that

$$\begin{aligned} A_n(x) &= \sum_{i=1}^{2} \exp(\beta^*_{01}) \exp((\beta^*_{11})^\top x) \sum_{\alpha=1}^{r} \frac{(b^n_i - b^*_1)^\alpha}{\alpha!} \cdot \sigma^{(\alpha)}(b^*_1) + R_1(x) \\ &= \Big[\sum_{\alpha=1}^{r} \frac{(1+2^\alpha)\sigma^{(\alpha)}(b^*_1)}{\alpha! n^r} \cdot c^\alpha\Big] \exp((\beta^*_{11})^\top x + \beta^*_{01}) + R_1(x), \end{aligned}$$

where $R_1(x)$ is a Taylor remainder such that $R_1(x)/\mathcal{D}_{3,r}(G_n, G_*) \to 0$. Note that $\Big[\sum_{\alpha=1}^{r} \frac{(1+2^\alpha)\sigma^{(\alpha)}(b^*_1)}{\alpha! n^\alpha} \cdot c^\alpha\Big]$ is an odd-order polynomial of c. Thus, we can choose c as a root of this polynomial, which leads to the fact that $A_n(x) = 0$. From the above results, we deduce that $Q_n(x)/\mathcal{D}_{3,r}(G_n, G_*) \to 0$, or equivalently, $[f_{G_n}(x) - f_{G_*}(x)]/\mathcal{D}_{3,r}(G_n, G_*) \to 0$ as $n \to \infty$ for almost every x. As a consequence, we achieve that $\|f_{G_n} - f_{G_*}\|_{L^2(\mu)}/\mathcal{D}_{3,r}(G_n, G_*) \to 0$.

When r is even: By means of the Taylor expansion of order $(r+1)$-th, we have

$$\begin{aligned} A_n(x) &= \sum_{i=1}^{2} \exp(\beta^*_{01}) \exp((\beta^*_{11})^\top x) \sum_{\alpha=1}^{r+1} \frac{(b^n_i - b^*_1)^\alpha}{\alpha!} \cdot \sigma^{(\alpha)}(b^*_1) + R_2(x) \\ &= \Big[\sum_{\alpha=1}^{r+1} \frac{(1+2^\alpha)\sigma^{(\alpha)}(b^*_1)}{\alpha! n^r} \cdot c^\alpha\Big] \exp((\beta^*_{11})^\top x + \beta^*_{01}) + R_2(x), \end{aligned}$$

where $R_2(x)$ is a Taylor remainder such that $R_2(x)/\mathcal{D}_{3,r}(G_n, G_*) \to 0$. Since $\Big[\sum_{\alpha=1}^{r+1} \frac{(1+2^\alpha)\sigma^{(\alpha)}(b^*_1)}{\alpha! n^\alpha} \cdot c^\alpha\Big]$ is an odd-degree polynomial of variable c, we can argue in a similar fashion to the scenario when r is odd to obtain that $\|f_{G_n} - f_{G_*}\|_{L^2(\mu)}/\mathcal{D}_{3,r}(G_n, G_*) \to 0$.

Combine results from the above two cases of r, we reach the conclusion of claim (54).

A.5. Proof of Theorem 4.4

Based on the result of Proposition 4.3, we demonstrate that the following minimax lower bound holds true for any $r \geq 1$:

$$\inf_{\overline{G}_n \in \mathcal{G}_k(\Theta)} \sup_{G \in \mathcal{G}_k(\Theta) \setminus \mathcal{G}_{k_*-1}(\Theta)} \mathbb{E}_{f_G}[\mathcal{D}_{3,r}(\overline{G}_n, G)] \gtrsim n^{-1/2}. \tag{55}$$

Indeed, from the Gaussian assumption on the noise variables, we obtain that $Y_i|X_i \sim \mathcal{N}(f_{G_*}(x_i), \sigma^2)$ for all $i \in [n]$. Now, from Proposition 4.3, for sufficiently small $\varepsilon > 0$ and a fixed constant $C_1 > 0$ that we will choose later, we can find a mixing measure $G'_* \in \mathcal{G}_k(\Theta)$ such that $\mathcal{D}_{3,r}(G'_*, G_*) = 2\varepsilon$ and $\|f_{G'_*} - f_{G_*}\|_{L^2(\mu)} \leq C_1\varepsilon$. From Le Cam's lemma (Yu, 1997), as the Voronoi loss function $\mathcal{D}_{3,r}$ satisfies the weak triangle inequality, we obtain that

$$\begin{aligned} \inf_{\overline{G}_n \in \mathcal{G}_k(\Theta)} \sup_{G \in \mathcal{G}_k(\Theta) \setminus \mathcal{G}_{k_*-1}(\Theta)} \mathbb{E}_{f_G}[\mathcal{D}_{3,r}(\overline{G}_n, G)] &\gtrsim \frac{\mathcal{D}_{3,r}(G'_*, G_*)}{8} \exp(-n\mathbb{E}_{X\sim\mu}[\mathrm{KL}(\mathcal{N}(f_{G'_*}(x), \sigma^2), \mathcal{N}(f_{G_*}(x), \sigma^2))]) \\ &\gtrsim \varepsilon \cdot \exp(-n\|f_{G'_*} - f_{G_*}\|^2_{L^2(\mu)}), \\ &\gtrsim \varepsilon \cdot \exp(-C_1 n\varepsilon^2), \end{aligned} \tag{56}$$

where the second inequality is due to the fact that $\mathrm{KL}(\mathcal{N}(f_{G'_*}(x), \sigma^2), \mathcal{N}(f_{G_*}(x), \sigma^2)) = \dfrac{(f_{G'_*}(x) - f_{G_*}(x))^2}{2\sigma^2}$.

By choosing $\varepsilon = n^{-1/2}$, we obtain that $\varepsilon \cdot \exp(-C_1 n\varepsilon^2) = n^{-1/2}\exp(-C_1)$. As a consequence, we achieve the desired minimax lower bound in equation (55).

A.6. Proof of Proposition 4.5

We need to prove that the following limit holds true for any $r \geq 1$:

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{G}_k(\Theta): \mathcal{D}_{3,r}(G,G_*) \leq \varepsilon} \frac{\|f_G - f_{G_*}\|_{L^2(\mu)}}{\mathcal{D}_{3,r}(G,G_*)} = 0. \tag{57}$$

For that purpose, it suffices to build a sequence of mixing measures (G_n) such that both $\mathcal{D}_{3,r}(G_n, G_*) \to 0$ and

$$\frac{\|f_{G_n} - f_{G_*}\|_{L^2(\mu)}}{\mathcal{D}_{3,r}(G_n,G_*)} \to 0,$$

as $n \to \infty$. To this end, we consider the sequence $G_n = \sum_{i=1}^{k_*+1} \exp(\beta_{0i}^n)\delta_{(\beta_{1i}^n, a_i^n, b_i^n)}$, where

- $\exp(\beta_{01}^n) = \exp(\beta_{02}^n) = \frac{1}{2}\exp(\beta_{01}^*) + \frac{1}{2n^{r+1}}$ and $\exp(\beta_{0i}^n) = \exp(\beta_{0(i-1)}^n)$ for any $3 \leq i \leq k_* + 1$;
- $\beta_{11}^n = \beta_{12}^n = \beta_{11}^*$ and $\beta_{1i}^n = \beta_{1(i-1)}^n$ 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 $b_i^n = b_{i-1}^*$ for any $3 \leq i \leq k_* + 1$.

As a result, the loss function $\mathcal{D}_{3,r}(G_n, G_*)$ is reduced to

$$\mathcal{D}_{3,r}(G_n,G_*) = \frac{1}{n^{r+1}} + \Big[\exp(\beta_{01}^*) + \frac{1}{n^{r+1}}\Big] \cdot \frac{1}{n^r} = \mathcal{O}(n^{-r}). \tag{58}$$

which indicates indicates that $\mathcal{D}_{3,r}(G_n,G_*) \to 0$ as $n \to \infty$. Now, we prove that $\|f_{G_n} - f_{G_*}\|_{L^2(\mu)}/\mathcal{D}_{3,r}(G_n,G_*) \to 0$. For that purpose, let us consider the quantity $Q_n(x) := [\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top x + \beta_{0j}^*)] \cdot [f_{G_n}(x) - f_{G_*}(x)]$. Then, we decompose $Q_n(x)$ as follows:

$$\begin{aligned}
Q_n(x) &= \sum_{j=1}^{k_*}\sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n)\Big[\exp((\beta_{1i}^n)^\top x)((a_i^n)^\top x + b_i^n) - \exp((\beta_{1j}^*)^\top x)((a_j^*)^\top x + b_j^*)\Big] \\
&- \sum_{j=1}^{k_*}\sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n)\Big[\exp((\beta_{1i}^n)^\top x) f_{G_n}(x) - \exp((\beta_{1j}^*)^\top x) f_{G_n}(x)\Big] \\
&+ \sum_{j=1}^{k_*}\Big(\sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^*)\Big)\Big[\exp((\beta_{1j}^*)^\top x)((a_j^*)^\top x + b_j^*) - \exp((\beta_{1j}^*)^\top x) f_{G_n}(x)\Big] \\
&:= A_n(x) - B_n(x) + E_n(x).
\end{aligned}$$

From the definitions of β_{1i}^n, a_i^n and b_i^n, we can rewrite $A_n(x)$ as follows:

$$A_n(x) = \sum_{i=1}^{2} \frac{1}{2}\exp(\beta_{01}^n)\exp((\beta_{11}^*)^\top x)(b_i^n - b_1^*) = \frac{1}{2}\exp(\beta_{01}^n)\exp((\beta_{11}^*)^\top x)[(b_1^n - b_1^*) + (b_2^n - b_1^*)] = 0.$$

Additionally, it can also be checked that $B_n(x) = 0$. Next, we have $E_n(x) = \mathcal{O}(n^{-(r+1)})$, therefore, it follows that $E_n(x)/\mathcal{D}_{3,r}(G_n,G_*) \to 0$. As a consequence, $Q_n(x)/\mathcal{D}_{3,r}(G_n,G_*) \to 0$ as $n \to \infty$ for almost every x. Since the term $\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top x + \beta_{0j}^*)$ is bounded, we deduce that $[f_{G_n}(x) - f_{G_*}(x)]/\mathcal{D}_{3,r} \to 0$ for almost every x. This result indicates that $\|f_{G_n} - f_{G_*}\|_{L^2(\mu)}/\mathcal{D}_{3,r} \to 0$ as $n \to \infty$. Hence, the proof of claim (57) is completed.

A.7. Proof of Theorem 4.6

By leveraging the result of Proposition 4.5 and the arguments for Theorem 4.4 in Appendix A.5, we achieve the following minimax lower bound for any $r \geq 1$:

$$\inf_{\overline{G}_n \in \mathcal{G}_k(\Theta)} \sup_{G \in \mathcal{G}_k(\Theta) \setminus \mathcal{G}_{k_*-1}(\Theta)} \mathbb{E}_{f_G}[\mathcal{D}_{3,r}(\overline{G}_n, G)] \gtrsim n^{-1/2}. \tag{59}$$

B. Identifiability of the Softmax Gating Mixture of Experts

Proposition B.1. *If $f_G(x) = f_{G_*}(x)$ holds true for almost every x, then we get that $G \equiv G'$.*

Proof of Proposition B.1. Since $f_G(x) = f_{G_*}(x)$ for almost every x, we have

$$\sum_{i=1}^{k} \mathrm{Softmax}\Big((\beta_{1i})^\top x + \beta_{0i}\Big) \cdot h(x, \eta_i) = \sum_{i=1}^{k_*} \mathrm{Softmax}\Big((\beta_{1i}^*)^\top x + \beta_{0i}^*\Big) \cdot h(x, \eta_i^*). \tag{60}$$

As the expert function h satisfies the conditions in Definition 3.1, the set $\{h(x, \eta_i') : i \in [k']\}$, where $\eta_1', \ldots, \eta_{k'}'$ are distinct vectors for some $k' \in \mathbb{N}$, is linearly independent. If $k \neq k_*$, then there exists some $i \in [k]$ such that $\eta_i \neq \eta_j^*$ for any $j \in [k_*]$. This implies that $\mathrm{Softmax}((\beta_{1i})^\top x + \beta_{0i}) = 0$, which is a contradiction. Thus, we must have that $k = k_*$. As a result,

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

for almost every x. WLOG, we may assume that

$$\mathrm{Softmax}\Big((\beta_{1i})^\top x + \beta_{0i}\Big) = \mathrm{Softmax}\Big((\beta_{1i}^*)^\top x + \beta_{0i}^*\Big), \tag{61}$$

for almost every x for any $i \in [k_*]$. It is worth noting that the $\mathrm{Softmax}$ function is invariant to translations, then equation (61) indicates that $\beta_{1i} = \beta_{1i}^* + v_1$ and $\beta_{0i} = \beta_{0i}^* + v_0$ for some $v_1 \in \mathbb{R}^d$ and $v_0 \in \mathbb{R}$. However, from the assumptions $\beta_{1k} = \beta_{1k}^* = 0_d$ and $\beta_{0k} = \beta_{0k}^* = 0$, we deduce that $v_1 = 0_d$ and $v_0 = 0$. Consequently, we get that $\beta_{1i} = \beta_{1i}^*$ and $\beta_{0i} = \beta_{0i}^*$ for any $i \in [k_*]$. Then, equation (60) can be rewritten as

$$\sum_{i=1}^{k_*} \exp(\beta_{0i}) \exp\Big((\beta_{1i})^\top x\Big) h(x, \eta_i) = \sum_{i=1}^{k_*} \exp(\beta_{0i}^*) \exp\Big((\beta_{1i}^*)^\top x\Big) h(x, \eta_i^*), \tag{62}$$

for almost every x. Next, we denote $P_1, P_2, \ldots, P_m$ as a partition of the index set $[k_*]$, where $m \leq k$, such that $\exp(\beta_{0i}) = \exp(\beta_{0i'}^*)$ for any $i, i' \in P_j$ and $j \in [k_*]$. On the other hand, when i and i' do not belong to the same set P_j, we let $\exp(\beta_{0i}) \neq \exp(\beta_{0i'})$. Thus, we can reformulate equation (62) as

$$\sum_{j=1}^{m} \sum_{i \in P_j} \exp(\beta_{0i}) \exp\Big((\beta_{1i})^\top x\Big) h(x, \eta_i) = \sum_{j=1}^{m} \sum_{i \in P_j} \exp(\beta_{0i}^*) \exp\Big((\beta_{1i}^*)^\top x\Big) h(x, \eta_i^*),$$

for almost every x. Recall that $\beta_{1i} = \beta_{1i}^*$ and $\beta_{0i} = \beta_{0i}^*$ for any $i \in [k_*]$, then the above leads to

$$\{\eta_i : i \in P_j\} \equiv \{\eta_i^* : i \in P_j\},$$

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

$$G = \sum_{j=1}^{m} \sum_{i \in P_j} \exp(\beta_{0i}) \delta_{(\beta_{1i}, \eta_i)} = \sum_{j=1}^{m} \sum_{i \in P_j} \exp(\beta_{0i}) \delta_{(\beta_{1i}^*, \eta_i^*)} = G_*.$$

Hence, we reach the conclusion of this proposition. □

C. Numerical Experiments

In this section, we conduct a simulation study to empirically demonstrate that the convergence rates of least square estimation under the softmax gating MoE model with ridge experts $h_1(x, (a, b)) = \mathrm{sigmoid}(ax + b)$ are significantly faster than those obtained when using linear experts $h_2(x, (a, b)) = ax + b$. We conduct those experiments under both the exact-specified setting (*when the true number of experts k_* is known*) and the over-specified setting (*when the true number of experts k_* is unknown*).

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

$$\beta_{01}^* = 0.0, \quad \beta_{11}^* = 1.0, \quad a_1^* = -1.0, \quad b_1^* = 2.0,$$

$$\beta_{02}^* = 0.0, \quad \beta_{12}^* = 0.0, \quad a_2^* = 1.0, \quad b_2^* = 2.0.$$

Then, we generate i.i.d samples $\{(X_i, Y_i)\}_{i=1}^n$ by first sampling X_i's from the uniform distribution $\mathrm{Uniform}[0,1]$ and then sampling Y_i's from the regression equation

$$Y_i = f_{G_*}(X_i) + \varepsilon_i,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent Gaussian noise variables such that $\mathbb{E}[\varepsilon_i|X_i] = 0$ and $\mathrm{Var}(\varepsilon_i|X_i) = 1$.

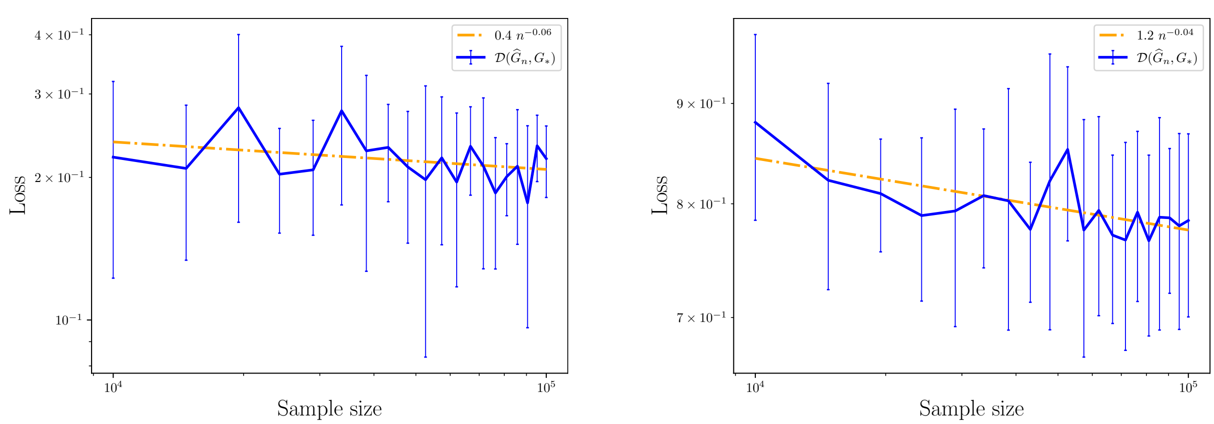
Initialization. For each $k \in \{k_*, k_* + 1\}$, we randomly distribute elements of the set $\{1, 2, ..., k\}$ into k_* different Voronoi cells $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_{k_*}$, each contains at least one element. Moreover, we repeat this process for each replication. Subsequently,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

Training. We use the stochastic gradient descent algorithm to minimize the mean square losses. We conduct 20 sample generations for each configuration, across a spectrum of 20 different sample sizes n ranging from 10^4 to 10^5. Finally, we generate log-log scaled plots for the Voronoi loss functions. For ridge experts, we use the Voronoi loss $\mathcal{D}_2$ given in Section 4.1.1, while for linear experts, we use the Voronoi loss $\mathcal{D}_{3,r}$ in Section 4.2.

- *Exact-specified setting:* Under this setting, as the true number of experts k_* is known, we set $k = k_* = 2$.
- *Over-specified setting:* Under this setting, as k_* is unknown, we over-specified the true model by setting $k = 3$.

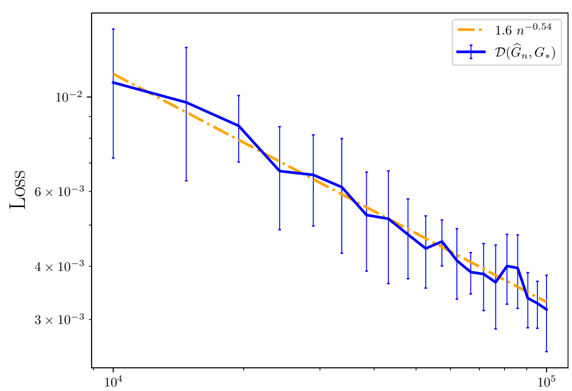
Remark. From Figure 1, it can be seen that under the exact-specified and over-specified settings, the convergence rates of least square estimators $\widehat{G}_n$ when using linear experts are significantly slow, at orders $\mathcal{O}(n^{-0.06})$ and $\mathcal{O}(n^{-0.04})$, respectively. This observation totally aligns with our theoretical result in Theorem 4.6.

On the other hand, Figure 2 indicates that when using ridge experts, the least square estimator $\widehat{G}_n$ converges to G_* at much faster rates, at order $\mathcal{O}(n^{-0.54})$ under the exact-specified setting, and at order $\mathcal{O}(n^{-0.57})$ under the over-specified setting. These empirical rates match the theoretical rate $\mathcal{O}(n^{-0.5})$ captured in Theorem 4.2.

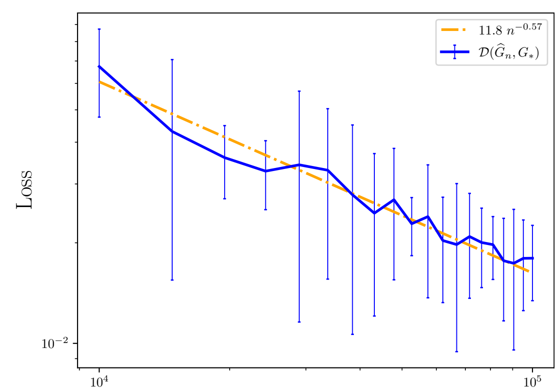


(a) Exact-specified setting (b) Over-specified setting

Figure 1. Log-log scaled plots illustrating empirical convergence rates of parameter estimation in the softmax gating mixture of **linear experts** under the exact-specified setting (Figure 1a) and the over-specified setting (Figure 1b). The blue curves depict the mean discrepancy between the least squares estimator $\widehat{G}_n$ and the true mixing measure G_* under the loss $\mathcal{D}_{3,r}$, accompanied by error bars signifying two empirical standard deviations. Additionally, an orange dash-dotted line represents the least-squares fitted linear regression line for these data points.



(a) Exact-specified setting



(b) Over-specified setting

Figure 2. Log-log scaled plots illustrating empirical convergence rates of parameter estimation in the softmax gating mixture of **ridge experts** with the sigmoid activation under the exact-specified setting (Figure 2a) and the over-specified setting (Figure 2b). The blue curves depict the mean discrepancy between the least squares estimator $\widehat{G}_n$ and the true mixing measure G_* under the loss $\mathcal{D}_2$, accompanied by error bars signifying two empirical standard deviations. Additionally, an orange dash-dotted line represents the least-squares fitted linear regression line for these data points.