A General Theory for Softmax Gating Multinomial Logistic Mixture of Experts

Huy Nguyen [1] **Pedram Akbarian** [2] **TrungTin Nguyen** [3,4] **Nhat Ho** [1]

Abstract

Mixture-of-experts (MoE) model incorporates the power of multiple submodels via gating functions to achieve greater performance in numerous regression and classification applications. From a theoretical perspective, while there have been previous attempts to comprehend the behavior of that model under the regression settings through the convergence analysis of maximum likelihood estimation in the Gaussian MoE model, such analysis under the setting of a classification problem has remained missing in the literature. We close this gap by establishing the convergence rates of density estimation and parameter estimation in the softmax gating multinomial logistic MoE model. Notably, when part of the expert parameters vanish, these rates are shown to be slower than polynomial rates owing to an inherent interaction between the softmax gating and expert functions via partial differential equations. To address this issue, we propose using a novel class of modified softmax gating functions which transform the input before delivering them to the gating functions. As a result, the previous interaction disappears and the parameter estimation rates are significantly improved.

1. Introduction

Mixture of experts (MoE) (Jacobs et al., 1991; Jordan & Jacobs, 1994) is a statistical machine learning model which aggregates several submodels represented by expert networks associated with standard softmax gating functions. Due to its modular and flexible structure, there has been a surge of interests in leveraging the MoE model in various fields, including large language models (Shazeer et al.,

2017; Lepikhin et al., 2021; Zhou et al., 2022; Du et al., 2022; Fedus et al., 2022; Zhou et al., 2023; Do et al., 2023), computer vision (Deleforge et al., 2015; Riquelme et al., 2021; Dosovitskiy et al., 2021; Bao et al., 2022; Liang et al., 2022), speech recognition (Peng et al., 1996; Gulati et al., 2020; You et al., 2022), reinforcement learning (Ren et al., 2021; Chow et al., 2023) and multi-task learning (Hazimeh et al., 2021; Gupta et al., 2022; Chen et al., 2023). Despite its recent progress in the aforementioned applications, the theoretical comprehension of the MoE model has been found relatively restricted in the literature.

In general, that model can be formulated as either a regression problem, namely when the distribution of MoE outputs is continuous, or a classification problem, i.e., when the MoE outputs follow a discrete distribution. Under the setting of a regression problem, there are some previous works attempting to theoretically understand the MoE model through the convergence analysis of maximum likelihood estimation (MLE) by focusing on Gaussian MoE equipped with different types of gating functions. First, (Ho et al., 2022) established the convergence rates of parameter estimation in the covariate-free gating Gaussian MoE model using the generalized Wasserstein loss function (Villani, 2003; 2008). In that paper, they discovered that those parameter estimation rates were significantly slow owing to an interaction between expert parameters via some partial differential equations (PDEs). Next, (Nguyen et al., 2024c) considered the Gaussian density gating Gaussian MoE model. Since that gating function depended on the covariates, there was another interaction between the parameters of the gating function and the expert function. To this end, the authors proposed Voronoi loss functions to capture those interactions and demonstrate that the convergence rates of parameter estimation were determined by the solvability of a system of polynomial equations. Subsequently, the Gaussian MoE models with softmax gating function and top-K sparse softmax gating function were investigated in (Nguyen et al., 2023) and (Nguyen et al., 2024b), respectively. Due to the non-linearity and sophisticated structures of those gating functions, the parameter estimation rates were shown to vary with the solvability of a more complex system of polynomial equations than that in (Nguyen et al., 2024c), and became substantially slow when the number of fitted experts increased.

[1]Department of Statistics and Data Sciences [2]Department of Electrical and Computer Engineering, The University of Texas at Austin [3]School of Mathematics and Physics, The University of Queensland [4]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK. Correspondence to: Huy Nguyen <huynm@utexas.edu>.

On the other hand, such theoretical analysis of the MoE model under the setting of a classification problem has remained poorly understood, to the best of our knowledge. Therefore, the main objective of our paper is to provide new insights on the convergence behavior of MLE in the softmax gating multinomial logistic MoE model (Chen et al., 1999; Yuksel & Gader, 2010; Huynh & Chamroukhi, 2019; Pham & Chamroukhi, 2022). Before going into further details, it is necessary to introduce the formulation of that model and associated assumptions.

Problem Setup. In this paper, we assume that the output $Y \in \{1, 2, \ldots, K\}$ is a discrete response variable, where $K \in \mathbb{N}$, while $X \in \mathcal{X}$ is an covariate vector having an effect on Y, in which $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$. Next, the data points $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are independently drawn from the standard softmax gating multinomial logistic mixture of experts of order k_*, which admits the conditional probability function $g_{G_*}(Y = s|X)$ defined for any $s \in \{1, 2, \ldots, K\}$ as follows:

$$\sum_{i=1}^{k_*} \text{Softmax}((\beta_{1i}^*)^\top X + \beta_{0i}^*) \times f(Y = s|X; a_i^*, b_i^*)$$

$$:= \sum_{i=1}^{k_*} \frac{\exp((\beta_{1i}^*)^\top X + \beta_{0i}^*)}{\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)}$$
$$\times \frac{\exp(a_{is}^* + (b_{is}^*)^\top X)}{\sum_{\ell=1}^{K} \exp(a_{i\ell}^* + (b_{i\ell}^*)^\top X)}. \quad (1)$$

Here, each expert $f(\cdot|X; a_i^*, b_i^*)$ is a multinomial logistic regression with parameters $a_i^* := (a_{i1}^*, \ldots, a_{iK}^*) \in \mathbb{R}^K$ and $b_i^* := (b_{i1}^*, \ldots, b_{iK}^*) \in \mathbb{R}^{d \times K}$. Meanwhile, $\beta_{1i}^* \in \mathbb{R}^d$ and $\beta_{0i}^* \in \mathbb{R}$ are referred to as gating parameters. Additionally, $G_* := \sum_{i=1}^{k_*} \exp(\beta_{0i}^*)\delta_{(\beta_{1i}^*, a_i^*, b_i^*)}$ denotes a true yet unknown *mixing measure*, that is, a combination of Dirac measures δ associated with true parameters $(\beta_{0i}^*, \beta_{1i}^*, a_i^*, b_i^*) \in \Theta$, where Θ is a bounded subset of $\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^K \times \mathbb{R}^{d \times K}$. Lastly, we define for any vector $v = (v_1, \ldots, v_{k_*}) \in \mathbb{R}^{k_*}$ that $\text{Softmax}(v_i) := \exp(v_i)/\sum_{j=1}^{k_*} \exp(v_j)$.

It is worth noting that if we translate β_{1i}^* to $\beta_{1i}^* + t_1$ and β_{0i}^* to $\beta_{0i}^* + t_0$, then the values of the standard softmax gating function do not change. This implies that gating parameters $\beta_{1i}^*, \beta_{0i}^*$ are only identifiable up to some translation. To alleviate this problem, we assume without loss of generality (WLOG) that $\beta_{1k_*}^* = \mathbf{0}_d$ and $\beta_{0k_*}^* = 0$. Similarly, we also assume that $a_{iK}^* = 0$ and $b_{iK}^* = \mathbf{0}_d$ for any $i \in \{1, 2, \ldots, k_*\}$. Furthermore, at least one among $\beta_{11}^*, \beta_{12}^*, \ldots, \beta_{1k_*}^*$ must be non-zero so that the softmax gating function depends on the covariate X. Lastly, we let (a_i^*, b_i^*), for $i \in \{1, 2, \ldots, k_*\}$, be pairwise distinct to ensure that the epxert functions are different from each other.

Maximum Likelihood Estimation. Regarding the parameter estimation problem in the standard softmax gating multi-

nomial logistic MoE, we propose using the maximum likelihood method in this work. However, as the true number of experts k_* is generally unknown in practice, it is necessary to consider an *over-specified* setting where we fit the true model with a mixture of k multinomial logistic experts, where $k > k_*$. In particular, the maximum likelihood estimation (MLE) of the true mixing measure G_* is given by:

$$\widehat{G}_n \in \arg\max_{G \in \mathcal{O}_k(\Theta)} \sum_{i=1}^{n} \log(g_G(Y_i|X_i)), \quad (2)$$

where $\mathcal{O}_k(\Theta)$ denotes the set of all mixing measures with at most k components of the form $G = \sum_{i=1}^{k'} \exp(\beta_{0i})\delta_{(\beta_{1i}, a_i, b_i)}$, in which $1 \leq k' \leq k$ and $(\beta_{0i}, \beta_{1i}, a_i, b_i) \in \Theta$.

Main Challenges. To characterize the parameter estimation rates, we need to decompose the difference $g_{\widehat{G}_n}(Y|X) - g_{G_*}(Y|X)$ into a linear combination of linearly independent terms using Taylor expansions. Then, when the density estimation $g_{\widehat{G}_n}(Y|X)$ converges to the true density $g_{G_*}(Y|X)$, all the associated coefficients, which involve the discrepancies between true parameters and their estimations, also tend to zero. Consequently, we achieve our desired parameter estimation rates based on the density estimation rate. However, when part of the expert parameters vanishes, there is an interaction between the numerator of the standard softmax gating function and the expert function, which induces several challenges in the density decomposition. In particular, let us denote $u(Y = s|X; \beta_{1i}, a_i, b_i) := \exp((\beta_{1i})^\top X)f(Y = s|X; a_i, b_i)$. If there exists an index $i \in [k_*]$ such that $b_{i\ell}^* = \mathbf{0}_d$ for any $\ell \in [K]$, then the aforementioned interaction is expressed via the following PDE:

$$\frac{\partial u}{\partial \beta_{1i}}(Y = s|X; \beta_{1i}^*, a_i^*, b_i^*)$$
$$= C_{a_i^*} \cdot \frac{\partial u}{\partial b_{is}}(Y = s|X; \beta_{1i}^*, a_i^*, b_i^*), \quad (3)$$

where $C_{a_i^*} > 0$ is a constant depending only on a_i^*. The above PDE shows that there are a number of linearly dependent derivative terms in the Taylor expansion. Then, we have to incorporate these terms by taking the summation of their associated coefficients in order to form a linear combination of linearly independent terms. Therefore, the structure of resulting coefficients becomes complex, which makes the parameter estimation rates slower than polynomial rates. This finding indicates that the standard softmax gating function might hurt the performance of the model in equation (1) despite its widespread use in the literature (Jacobs et al., 1991; Jordan & Jacobs, 1994; Nguyen et al., 2023).

Contribution. Following the above challenge discussion, we construct a generic Voronoi-based metric $\mathcal{D}_r$ among

Table 1. Summary of density estimation rates and parameter estimation rates in the multinomial logistic MoE model with standard and modified softmax gating functions. Here, we refer to exact-specified parameters $\beta_{1j}^*, a_j^*, b_j^*$ as those fitted by exactly one component (i.e. $|\mathcal{C}_j| = 1$), while over-specified parameters are approximated by more than one component (i.e. $|\mathcal{C}_j| > 1$), where $\mathcal{C}_j$ is a Voronoi cell defined in Section 2.

Gating	Parameter Setting	Density	Exact-specified Parameters	Over-specified Parameters
Standard Softmax	Regime 1	$\widetilde{\mathcal{O}}(n^{-1/2})$	$\widetilde{\mathcal{O}}(n^{-1/2})$	$\widetilde{\mathcal{O}}(n^{-1/4})$
	Regime 2	$\widetilde{\mathcal{O}}(n^{-1/2})$	slower than $\widetilde{\mathcal{O}}(n^{-1/2r})$ for any $r \geq 1$	
Modified Softmax	Any regime	$\widetilde{\mathcal{O}}(n^{-1/2})$	$\widetilde{\mathcal{O}}(n^{-1/2})$	$\widetilde{\mathcal{O}}(n^{-1/4})$

parameters in equation (5) to capture the convergence behavior of the MLE $\widehat{G}_n$ to the true mixing measure G_* in the standard softmax gating multinomial logistic MoE model. Moreover, we also design a novel class of modified softmax gating functions which improves the convergence rate of the MLE. Our contributions are two-fold and can be summarized as follows (see also Table 1):

1. Standard Softmax Gating Function. With the standard softmax gating function, we demonstrate that the density estimation $g_{\widehat{G}_n}$ converges to the true density g_{G_*} under the Hellinger distance h at the parametric rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$. Next, we consider the parameter estimation problem under two following complement regimes of the expert parameters b_i^*:

(i) Regime 1: For any $i \in \{1, 2, \ldots, k_*\}$, we can find an index $\ell \in [K - 1]$ that satisfies $b_{i\ell}^* \neq \mathbf{0}_d$. Under this regime, the PDE in equation (3) does not hold and there is no interaction between the standard softmax gating and the expert functions. By deriving the Hellinger lower bound $\mathbb{E}_X[h(g_G(\cdot|X), g_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_2(G, G_*)$ for any mixing measure $G \in \mathcal{O}_k(\Theta)$ in Theorem 3.1, we obtain that the rates for estimating over-specified parameters β_{1i}^*, a_i^* and b_i^*, i.e. those whose Voronoi cells have more than one element, are identical of order $\widetilde{\mathcal{O}}(n^{-1/4})$. At the same time, the estimation rates for their exact-specified parameters, namely those whose Voronoi cells have exactly one element, are substantially faster of order $\widetilde{\mathcal{O}}(n^{-1/2})$.

(ii) Regime 2: There exists an index $i \in \{1, 2, \ldots, k_*\}$ such that $b_{i\ell}^* = \mathbf{0}_d$ for any $\ell \in [K - 1]$. Since the PDE (3) holds true under this regime, an interaction between the standard softmax gating and the expert functions occurs. Then, we demonstrate in Theorem 3.3 that the minimax lower bound $\inf_{\overline{G}_n \in \mathcal{O}_k(\Theta)} \sup_{G \in \mathcal{O}_k(\Theta) \setminus \mathcal{O}_{k_*-1}(\Theta)} \mathbb{E}_{p_G}[\mathcal{D}_r(\overline{G}_n, G)] \gtrsim n^{-1/2}$ holds for any $r \geq 1$. This bound together with the formulation of $\mathcal{D}_r$ indicate that true parameters β_{1i}^*, a_i^* and b_i^* enjoy much worse rates than those in Regime 1. Remarkably, those for over-specified parameters are even slower than polynomial rates.

2. Modified Softmax Gating Function. To resolve the aforementioned downsides of the standard softmax gating function towards the convergence rates of parameter estimation, we propose using the following modified softmax gating functions:

$$\mathrm{Softmax}((\beta_{1i}^*)^\top M(X) + \beta_{0i}^*)$$
$$:= \frac{\exp((\beta_{1i}^*)^\top M(X) + \beta_{0i}^*)}{\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top M(X) + \beta_{0j}^*)},$$

for any $i \in \{1, 2, \ldots, k_*\}$ where $M : \mathbb{R}^d \to \mathbb{R}^d$ is a bounded function such that the set $\left\{X^p[M(X)]^q : p, q \in \mathbb{N}^d, 0 \leq |p| + |q| \leq 2\right\}$ is linearly independent for almost surely X. These assumptions guarantee that the PDE in equation (3) does not hold for any values of expert parameters b_i^*, and therefore, the previous interaction between the gating and expert functions disappears. As a consequence, we show in Theorem 4.4 that the parameter estimations under the modified softmax gating multinomial logistic MoE model share the same convergence behavior as those in Regime 1 regardless of the values of expert parameters b_i^*.

Practical Implications. Below are three practical implications from our convergence analysis of the softmax gating multinomial logistic MoE:

1. Expert parameter collapse: It is worth noting that the parameter estimation rates when the expert parameters collapse are significantly slow, and could be of order $O(1/\log(n))$ (see Table 1). Therefore, during the training process, if ones observe that the model convergence becomes abnormally slow, or the updated loss values almost remain unchanged, then it is highly likely that the expert parameter collapse occurs.

2. Model design: Despite the widespread use of the standard softmax gate in the practical applications of MoE models, the insights from our theory indicates that this gate is not always beneficial for the model performance, particularly when the expert parameter collapse happens. Therefore,

our analysis suggests using a novel modified softmax gate which helps stabilize the training process regardless of the parameter collapse.

3. Expert selection: The practical problem of selecting important experts can benefit from the modified softmax gate. In particular, in many applications of MoE models, there are some experts which do not play an essential role in learning, and even become redundant. In such scenario, we would like the gate to put more weights on the important experts. However, if the input magnitude is huge, then the weight distribution will become uniform, which is undesirable. To address this issue, practitioners can use an input normalization function, i.e. $M(X) = \frac{X}{||X||}$, in the softmax weights so that the input magnitude remains unchanged of one. Another possible option is $M(X) = sigmoid(X)$, which allows the magnitude of the input to vary between 0 and 1.

Outline. The remainder of the paper is organized in the following way. In Section 2, we verify the identiability of the standard softmax gating multinomial logistic MoE model and establish the parametric density estimation rate under that model prior to introducing a novel class of Voronoi loss functions used for the parameter estimation problem. Subsequently, we characterize the parameter estimation rates under the multinomial logistic MoE model equipped with the standard softmax gating function and the modified softmax gating function in Section 3 and Section 4, respectively. Finally, we conclude the paper in Section 5. Meanwhile, full proofs, additional results and a simulation study are relegated to the Appendices.

Notations. First, we let $[n]$ stand for the set $\{1, 2, \ldots, n\}$ for any number $n \in \mathbb{N}$. Next, for any vector $u, v \in \mathbb{N}^d$, we denote $|v| := v_1 + v_2 + \ldots + v_d$, $v^u := v_1^{u_1} v_2^{u_2} \ldots v_d^{u_d}$ and $v! := v_1! v_2! \ldots v_d!$, whereas $\|v\|$ represents for its 2-norm value. Meanwhile, for any set S, we denote $|S|$ as its cardinality. Additionally, for any two probability density functions p, q dominated by the Lebesgue measure μ, we define $V(p, q) := \frac{1}{2} \int |p - q| \mathrm{d}\mu$ as the Total Variation distance between them, and $h(p, q) := \left(\frac{1}{2} \int |\sqrt{p} - \sqrt{q}|^2 \mathrm{d}\mu \right)^{1/2}$ as their Hellinger distance. Lastly, for any two sequences (s_n) and (t_n) in $\mathbb{R}_+$, the notations $s_n = \mathcal{O}(t_n)$ and $s_n \lesssim t_n$ means there exists some constant $C > 0$ independent of n such that $s_n \leq C t_n$ for sufficiently large $n \in \mathbb{N}$, while the notation $s_n = \widetilde{\mathcal{O}}(t_n)$ suggests that the aforementioned inequality might hold up to some logarithmic function of n.

2. Preliminaries

In this section, we study the identifiability of the standard softmax gating multinomial logistic MoE model and the convergence behavior of density estimation under that model. Then, we further highlight the necessity for Voronoi-based

loss functions to determine parameter estimation rates accurately.

Firstly, we show that the standard softmax gating multinomial logistic MoE model is identifiable.

Proposition 2.1 (Identifiability). *Given two mixing measures G and G' in $\mathcal{O}_k(\Theta)$, if $g_G(Y|X) = g_{G'}(Y|X)$ holds true for almost surely (X, Y), then $G \equiv G'$.*

The proof of Proposition 2.1 is in Appendix C.1. This result ensures the convergence of the MLE $\widehat{G}_n$ to the true mixing measure G_* as long as the conditional density $g_{\widehat{G}_n}$ converges to g_{G_*} for almost surely (X, Y). Next, we characterize the density estimation rate under the Hellinger distance in the following proposition:

Proposition 2.2 (Density Estimation Rate). *With the MLE $\widehat{G}_n$ defined in equation (2), the convergence rate of the density estimation $g_{\widehat{G}_n}$ to the true density g_{G_*} is given by:*

$$\mathbb{P}\Big(\mathbb{E}_X[h(g_{\widehat{G}_n}(\cdot|X), g_{G_*}(\cdot|X))] > C_0 \sqrt{\log(n)/n} \Big)$$
$$\lesssim \exp(-c_0 \log(n)), \quad (4)$$

where C_0 and c_0 are universal positive constants that depend only on Θ.

The proof of Proposition 2.2 is in Appendix B.1. It is evident from the bound in equation (4) that the rate for estimating g_{G_*} is parametric of order $\widetilde{\mathcal{O}}(n^{-1/2})$. Therefore, if we are able to construct a metric $\mathcal{D}$ among parameters that satisfies the inequality $\mathbb{E}_X[h(g_G(\cdot|X), g_{G_*}(\cdot|X))] \gtrsim \mathcal{D}(G, G_*)$ for any mixing measure $G \in \mathcal{O}_k(\Theta)$, then the MLE $\widehat{G}_n$ will also converge to the true mixing measure G_* at the parametric rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$.

Voronoi-based Loss: As the MLE $\widehat{G}_n$ is constrained to have more components than its true counterpart, there exists a true component approximated by at least two components while others could be fitted by exactly one component. This possibly leads to different estimation rates among the true parameters, which requires us to design novel loss functions tailored to this observation. To this end, let us recall a notion of Voronoi cells, which was previously studied in (Manole & Ho, 2022). In particular, for any mixing measure $G \in \mathcal{O}_k(\Theta)$, a Voronoi cell of G generated by $\theta_j^* := (\beta_{1j}^*, a_j^*, b_j^*)$ is defined as $\mathcal{C}_j \equiv \mathcal{C}_j(G) := \{i \in [k] : \|\theta_i - \theta_j^*\| \leq \|\theta_i - \theta_\ell^*\|, \forall \ell \neq j\}$ for any $j \in [k_*]$, where $\theta_i := (\beta_{1i}, a_i, b_i)$. Here, the cardinality of each Voronoi cell $\mathcal{C}_j$ indicates the number of fitted components approximating the true component θ_j^*. Based on these Voronoi cells, we define a generic Voronoi-based metric of order $r \geq 1$ between any two mixing measures,

denoted by $\mathcal{D}_r(G, G_*)$, as follows:

$$\mathcal{D}_r(G, G_*) := \sum_{j=1}^{k_*} \left| \sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}) - \exp(\beta_{0j}^*) \right|,$$

$$+ \sum_{\substack{j:|\mathcal{C}_j|>1, \\ i \in \mathcal{C}_j}} \exp(\beta_{0i}) \Big[\|\Delta\beta_{1ij}\|^r + \sum_{\ell=1}^{K-1} (|\Delta a_{ij\ell}|^r + \|\Delta b_{ij\ell}\|^r) \Big]$$

$$+ \sum_{\substack{j:|\mathcal{C}_j|=1, \\ i \in \mathcal{C}_j}} \exp(\beta_{0i}) \Big[\|\Delta\beta_{1ij}\| + \sum_{\ell=1}^{K-1} (|\Delta a_{ij\ell}| + \|\Delta b_{ij\ell}\|) \Big],$$

$$(5)$$

where $\Delta\beta_{1ij} := \beta_{1i} - \beta_{1j}^*$, $\Delta a_{ij\ell} := a_{i\ell} - a_{j\ell}^*$ and $\Delta b_{ij\ell} := b_{i\ell} - b_{j\ell}^*$. The above Voronoi loss function allows us to capture precisely the distinct convergence behaviors of parameter estimations under the multinomial logistic MoE model with (modified) softmax gating function.

3. Standard Softmax Gating Multinomial Logistic MoE

In this section, we present the convergence analysis of parameter estimation in the standard softmax gating multinomial logistic MoE model.

Given the parametric density estimation rate in Proposition 2.2, we observe that if the Hellinger lower bound

$$\mathbb{E}_X[h(g_{\widehat{G}_n}(\cdot|X), g_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_r(\widehat{G}_n, G_*),$$

holds true for some $r \geq 1$, then the convergence rate of the MLE $\widehat{G}_n$ to the true mixing measure G_* would also be parametric of order $\widetilde{\mathcal{O}}(n^{-1/2})$. Therefore, our goal is to establish the above bound. Since the Hellinger distance is lower bounded by the Total Variation distance, i.e., $h \geq V$, it is sufficient to show that

$$\mathbb{E}_X[V(g_{\widehat{G}_n}(\cdot|X), g_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_r(\widehat{G}_n, G_*). \quad (6)$$

Then, we consider the term $V(g_{\widehat{G}_n}(\cdot|X), g_{G_*}(\cdot|X))$ which involves the density discrepancy $g_{\widehat{G}_n}(Y|X) - g_{G_*}(Y|X)$. We aim to decompose this discrepancy into a linear combination of linearly independent terms using Taylor expansion so that when $g_{\widehat{G}_n}$ converges to g_{G_*}, the associated coefficients involving the parameter differences $\widehat{\beta}_{1i}^n - \beta_{1j}^*$, $\widehat{a}_i^n - a_j^*$ and $\widehat{b}_i^n - b_j^*$ also tend to zero, where $(\widehat{\beta}_{1i}^n, \widehat{a}_i^n, \widehat{b}_i^n)$ are components of $\widehat{G}_n$. For that purpose, we first rewrite the previous density discrepancy in terms of

$$u(Y|X; \widehat{\beta}_{1i}^n, \widehat{a}_i^n, \widehat{b}_i^n) - u(Y|X; \beta_{1j}^*, a_j^*, b_j^*),$$

where $u(Y|X; \beta_{1i}, a_i, b_i) = \exp((\beta_{1i})^\top X) f(Y|X; a_i, b_i)$. Next, we apply the Taylor expansion to the function

$u(Y|X; \widehat{\beta}_{1i}^n, \widehat{a}_i^n, \widehat{b}_i^n)$ about the point $(\beta_{1j}^*, a_j^*, b_j^*)$. In this step, we notice that if there exists $j \in [k_*]$ such that $b_{j\ell}^* = \mathbf{0}_d$ for any $\ell \in [K-1]$, then the gating parameters interact with the expert parameters via the following PDEs:

$$\frac{\partial u}{\partial \beta_{1j}}(Y = s|X; \beta_{1j}^*, a_j^*, b_j^*)$$

$$= C_{a_j^*} \cdot \frac{\partial u}{\partial b_{js}}(Y = s|X; \beta_{1j}^*, a_j^*, b_j^*), \quad (7)$$

for any $s \in [K]$, where $C_{a_j^*} > 0$ is some constant depending only on a_j^*. This interaction indicates that there are several linearly dependent derivative terms in the previous Taylor expansion, which induces a number of challenges in representing the density discrepancy $g_{\widehat{G}_n}(Y|X) - g_{G_*}(Y|X)$ as a linear combination of linearly independent terms. On the other hand, if for any $j \in [k_*]$, we can find an index $\ell \in [K-1]$ such that $b_{j\ell}^* \neq \mathbf{0}_d$, then the PDEs inequation (7) no longer hold true and the previous interaction does not occur, which facilitates the decomposition of the density discrepancy.

For those reasons, we will characterize the parameter estimation rates under two following complement regimes of expert parameters b_j^* in Section 3.1 and Section 3.2, respectively:

- **Regime 1:** For any $j \in [k_*]$, there exists an index $\ell \in [K-1]$ such that $b_{j\ell}^* \neq \mathbf{0}_d$;

- **Regime 2:** There exists an index $j \in [k_*]$ such that $b_{j\ell}^* = \mathbf{0}_d$ for any $\ell \in [K-1]$.

3.1. Regime 1 of Expert Parameters

Under this regime, it can be verified that the PDEs in equation (7) are no longer valid for any $j \in [k_*]$. Then, we provide in the following theorem the convergence rate of the MLE $\widehat{G}_n$ to its true counterpart G_* under the Voronoi loss function $\mathcal{D}_2$ in equation (5).

Theorem 3.1 (Parameter Estimation Rate)**.** *Suppose that the assumption of Regime 1 holds true, then we achieve the Hellinger lower bound*

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

for any mixing measure $G \in \mathcal{O}_k(\Theta)$. Therefore, combining this bound with Proposition 2.2 leads to the following convergence rate of the MLE $\widehat{G}_n$:

$$\mathbb{P}\Big(\mathcal{D}_2(\widehat{G}_n, G_*) > C_1 \sqrt{\log(n)/n} \Big) \lesssim \exp(-c_1 \log(n)),$$

where $C_1 > 0$ is a constant depending on Θ and G_*, while the constant $c_1 > 0$ depends only on Θ.

Proof of Theorem 3.1 is in Appendix A.1. A few comments regarding this result are in order.

(i) Theorem 3.1 suggests that the MLE $\widehat{G}_n$ converges to the true mixing measure G_* under the loss function $\mathcal{D}_2$ at the parametric rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$. Based on the formulation of $\mathcal{D}_2$ in equation (5), it follows that exact-specified parameters $\beta_{1j}^*, a_j^*, b_j^*$, whose Voronoi cells $\mathcal{C}_j$ have exactly one element, share the same estimation rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$.

(ii) On the other hand, for over-specified parameters $\beta_{1j}^*, a_j^*, b_j^*$ whose Voronoi cells $\mathcal{C}_j$ have more than one element, the rates for estimating them are slower, standing at order $\widetilde{\mathcal{O}}(n^{-1/4})$. Nevertheless, these rates are substantially faster than their counterparts under the softmax gating Gaussian MoE model studied in (Nguyen et al., 2023), which decreases monotonically with the cardinality of Voronoi cells.

3.2. Regime 2 of Expert Parameters

Under this regime, we can find an index $j \in [k_*]$ that satisfies $b_{j\ell}^* = 0$ for any $\ell \in [K-1]$. By simple calculations, we can validate that the PDEs in equation (7) hold true. This result leads to so many linearly dependent terms among the derivatives of the function u w.r.t its parameters that the density discrepancy $g_{\widehat{G}_n}(Y|X) - g_{G_*}(Y|X)$ cannot be decomposed into a linear combination of linearly independent elements as our expectation. As a consequence, we demonstrate in the following proposition that the Total Variation lower bound in equation (6) no longer holds under Regime 2.

Proposition 3.2. *Suppose that the assumption of Regime 2 is satisfied, then we obtain that*

$$\inf_{\substack{G \in \mathcal{O}_k(\Theta), \\ \mathcal{D}_r(G,G_*) \leq \varepsilon}} \mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_r(G, G_*) \to 0,$$

as $\varepsilon \to 0$, for any $r \geq 1$.

Proof of Proposition 3.2 is in Appendix A.2.2. Based on the above result, we establish a minimax lower bound for estimating the true mixing measure G_* in Theorem 3.3.

Theorem 3.3 (Minimax Lower Bound). *Under Regime 2, the following minimax lower bound of estimating G_* holds true for any $r \geq 1$:*

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

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

Proof of Theorem 3.3 can be found in Appendix A.2. This result together with the formulation of $\mathcal{D}_r$ in equation (5)

indicates that the rates for estimating exact-specified parameters $\beta_{1j}^*, a_j^*, b_j^*$ are not better than $\widetilde{\mathcal{O}}(n^{-1/2})$, while those for their over-specified counterparts are even slower than any polynomial rates.

3.3. Proof Sketches

In this section, we provide proof sketches for both Theorem 3.1 and Theorem 3.3.

3.3.1. Proof Sketch of Theorem 3.1

Due to the inequality $h \geq V$, it suffices to show the Total Variation lower bound $\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X)] \gtrsim \mathcal{D}_2(G, G_*)$ for any $G \in \mathcal{O}_k(\Theta)$. We divide this problem into two parts as follows:

Local structure. In this part, we need to show that

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

Assume by contrary that there exists $G_n \in \mathcal{O}_k(\Theta)$ such that $\mathbb{E}_X[V(g_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_2(G_n, G_*)$ and $\mathcal{D}_2(G_n, G_*)$ both converge to zero.

Step 1. In this step, we use Taylor expansions to decompose the quantity $T_n(s) := \left[\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*) \right] \cdot [g_{G_n}(Y = s|X) - g_{G_*}(Y = s|X)]$ into a linear combination of linearly independent terms as

$$T_n(s) = \sum_{j=1}^{k_*} \omega_j E_j(Y = s|X) + R(X, Y),$$

where $R(X, Y)$ is a Taylor remainder such that $R(X, Y)/\mathcal{D}_2(G_n, G_*) \to 0$.

Step 2. In this step, we show by contradiction that not all the ratios $\omega_j/\mathcal{D}_2(G_n, G_*)$ tend to zero as $n \to \infty$. Assume that all of them converge to zero, then we deduce that $1 = \mathcal{D}_2(G_n, G_*)/\mathcal{D}_2(G_n, G_*) \to 0$, which is a contradiction. Therefore, at least one among the ratios $w_j/\mathcal{D}_2(G_n, G_*)$ does not vanish.

Step 3. Since $\mathbb{E}_X[V(g_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_2(G_n, G_*)$ converges to zero, then by applying the Fatou's lemma, we deduce that $T_n(s)/\mathcal{D}_2(G_n, G_*) \to 0$ as $n \to \infty$. Note that $T_n(s)$ is written as a linear combination of linearly independent terms, thus, all the associated coefficients in that combination go to zero, which contradicts the result in Step 2. Hence, the proof of the local structure is completed.

Global Structure. In this part, we demonstrate that

$$\inf_{\substack{G \in \mathcal{O}_k(\Theta), \\ \mathcal{D}_2(G,G_*) > \varepsilon}} \frac{\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_2(G, G_*)} > 0.$$

Assume that this claim is not true. Then, we can find a mixing measure $G' \in \mathcal{O}_k(\Theta)$ such that $\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))] = 0$. By the Fatou's lemma, we obtain that $g_{G'}(Y = s|X) = g_{G_*}(Y = s|X)$ for any $s \in [K]$ for almost surely X. It follows from Proposition 2.2 that $G' \equiv G$. This result implies that $\mathcal{D}_2(G', G_*) = 0$, which contradicts the constraint $\mathcal{D}_2(G', G_*) > \varepsilon > 0$. Hence, the proof sketch is completed.

3.3.2. PROOF SKETCH OF THEOREM 3.3

Let M_1 be some fixed positive constant. Following from the result of Proposition 3.2, for any suffciently small $\varepsilon > 0$, we can seek a mixing measure $G'_* \in \mathcal{O}_k(\Theta)$ such that $\mathcal{D}_r(G'_*, G_*) = 2\varepsilon$ and $\mathbb{E}_X[V(g_{G'_*}(\cdot|X), g_{G_*}(\cdot|X))] \leq M_1\varepsilon$. Note that for any sequence of mixing measures $\overline{G}_n \in \mathcal{O}_k(\Theta)$, we have

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

Moreover, since $\mathcal{D}_r$ satisfies the weak triangle inequality, there exists some constant $M_2 > 0$ such that

$$\mathcal{D}_r(\overline{G}_n, G_*) + \mathcal{D}_r(\overline{G}_n, G'_*) \geq M_2 \mathcal{D}_r(G_*, G'_*) = 2M_2\varepsilon.$$

Then, by leveraging the Le Cam's minimax lower bound approach (Yu, 1997), we get that

$$\max_{G \in \{G_*, G'_*\}} \mathbb{E}_{g_G}[\mathcal{D}_r(\overline{G}_n, G)]$$
$$\geq M_2\varepsilon(1 - \mathbb{E}_X[V(g^n_{G_*}(\cdot|X), g^n_{G'_*}(\cdot|X))])$$
$$\geq M_2\varepsilon\left[1 - \sqrt{1 - (1 - M_1^2\varepsilon^2)^n}\right].$$

By setting $\varepsilon = n^{-1/2}/M_1$, we obtain that

$$\max_{G \in \{G_*, G'_*\}} \mathbb{E}_{g_G}[\mathcal{D}_r(\overline{G}_n, G)] \gtrsim n^{-1/2},$$

for any sequence $\overline{G}_n \in \mathcal{O}_k(\Theta)$. Furthermore, since $\{G_*, G'_*\}$ is a subset of $\mathcal{O}_k(\Theta) \setminus \mathcal{O}_{k_* - 1}(\Theta)$, we reach the conclusion of Theorem 3.3.

4. Modified Softmax Gating Multinomial Logistic MoE

In this section, we propose a novel class of modified softmax gating functions to resolve the interaction between gating parameters and expert parameters via the PDE (7) which leads to significantly slow parameter estimation rates. Then, we also capture the convergence rates of parameter estimation under the multinomial logistic MoE model with that those gating functions.

First of all, let us introduce the formulation of the modified softmax gating multinomial logistic MoE model.

Problem Setup. Suppose that the i.i.d samples $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \in \mathcal{X} \times [K]$ are drawn from the multinomial logistic MoE model of order k_* with modified softmax gating function, whose conditional density function $\widetilde{g}_{G_*}(Y = s|X)$ is given by

$$\sum_{i=1}^{k_*} \mathrm{Softmax}((\beta_{1i}^*)^\top M(X) + \beta_{0i}^*) \times f(Y = s|X; a_i^*, b_i^*)$$
$$:= \sum_{i=1}^{k_*} \frac{\exp((\beta_{1i}^*)^\top M(X) + \beta_{0i}^*)}{\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top M(X) + \beta_{0j}^*)}$$
$$\times \frac{\exp(a_{is}^* + (b_{is}^*)^\top X)}{\sum_{\ell=1}^K \exp(a_{i\ell}^* + (b_{i\ell}^*)^\top X)}, \tag{8}$$

for any $s \in [K]$, where M is a function defined in Definition 4.1. Here, we reuse all the assumptions imposed on the model in equation (1) unless stating otherwise. In the above model, we would transform the covariate X using the function M first rather than routing it directly to the softmax gating function as in the model (1). This transformation step allows us to overcome the interaction between gating parameters and expert parameters emerging from Section 3.2, which we will discuss below.

Definition 4.1 (Modified Function). Let $M : \mathcal{X} \to \mathbb{R}^d$ be a bounded function such that the following set is linearly independent for almost surely X:

$$\left\{ X^p[M(X)]^q : p, q \in \mathbb{N}^d, \ 0 \leq |p| + |q| \leq 2 \right\}. \tag{9}$$

To understand the above condition better, we provide below both intuitive and technical explanations for it.

Intuitively, Under the multinomial logistic MoE model with the standard softmax gate, we observe an interaction among the gating parameters β_1 and the expert parameters b when a fraction of expert parameters vanish (see equation (3). This interaction mainly accounts for the slow parameter estimation rates, which could be as slow as $\mathcal{O}(1/\log(n))$. We realize that the previous interaction occurs as parameters β_1 and b are both associated with the input X in the condition density in equation (1). To address this issue, we propose transforming the input X in the softmax gate by the function M. Then, in the modified conditional density in equation (8), β_1 is associated with $M(X)$, while b is still with X. However, to eliminate the parameter interaction completely, we need to impose an assumption of linear independence between the input X and its transformation $M(X)$ in Definition 4.1.

Technically, the set in equation (9) is assumed to be linearly independent for almost surely X to guarantee there does not exist any constant $C_{a_i^*}$ depending only a_i^* such that the

PDEs

$$\frac{\partial u}{\partial \beta_{1j}}(Y = s|X; \beta_{1j}^*, a_j^*, b_j^*)$$
$$= C_{a_j^*} \cdot \frac{\partial u}{\partial b_{js}}(Y = s|X; \beta_{1j}^*, a_j^*, b_j^*),$$

for $s \in [K]$, do not hold under any setting of the expert parameters b_i^*. As a result, the interaction between gating parameters and expert parameters mentioned in the Regime 2 does not appear in the modified softmax gating multinomial logistic MoE model. Next, we provide below a few valid instances of the function M.

Example. It can verified that all the following element-wise functions satisfy the condition in Definition 4.1: $M(X) = \tanh(X)$, $M(X) = \cos(X)$ and $M(X) = X^m$ for $m \geq 3$. Additionally, M could also be a normalization function, i.e. $M(X) = \frac{X}{||X||}$. The practical problem of selecting important experts can benefit from this choice of function M. In particular, in many applications of MoE models, there are some experts which do not play an essential role in learning, and even become redundant. In such scenario, we would like the gate to put more weights on the important experts. However, if the input magnitude is huge, then the weight distribution will become uniform, which is undesirable. Therefore, the input normalization function M, which helps remain the input magnitude of one, is beneficial in this case. Regarding the parameter estimation problem, since the function M helps remove the parameter interaction, the parameter estimation rates are ensured to be of polynomial orders under any parameter settings (see Table 1). Another potential option is $M(X) = sigmoid(X)$, which allows the magnitude of the input to vary between 0 and 1.

Maximum Likelihood Estimation. Due to the modification of the conditional density in equation (8), the formulation of MLE of the true mixing measure G_* under the modified softmax gating multinomial logistic MoE model also changes as follows:

$$\widetilde{G}_n \in \arg\max_{G \in \mathcal{O}_k(\Theta)} \sum_{i=1}^n \log(\widetilde{g}_G(Y_i|X_i)). \quad (10)$$

Next, we validate the identifiability of the modified softmax gating multinomial logistic MoE model in the following proposition.

Proposition 4.2 (Identifiability). *Assume that G and G' are two mixing measures in $\mathcal{O}_k(\Theta)$. If $\widetilde{g}_G(Y|X) = \widetilde{g}_{G'}(Y|X)$ holds true for almost surely (X, Y), then we obtain that $G \equiv G'$.*

Proof of Proposition 4.2 is in Appendix B.1.2. This result points out that the modified softmax gating multinomial logistic MoE model is still identifiable. In other words, if

the density estimation $\widetilde{g}_{\widetilde{G}_n}$ converges to the true density $\widetilde{g}_{G_*}$, then the MLE $\widetilde{G}_n$ also approaches its true counterpart G_*.

Now, we are ready to derive the convergence rate of density estimation $\widetilde{g}_{\widetilde{G}_n}$ to the true density $\widetilde{g}_{G_*}$ under the Hellinger distance.

Proposition 4.3 (Density Estimation Rate). *With the MLE $\widetilde{G}_n$ defined in equation (10), the density estimation $\widetilde{g}_{\widetilde{G}_n}$ converges to the true density $\widetilde{g}_{G_*}$ at the following rate:*

$$\mathbb{P}\Big(\mathbb{E}_X[h(\widetilde{g}_{\widetilde{G}_n}(\cdot|X), \widetilde{g}_{G_*}(\cdot|X))] > C_2\sqrt{\log(n)/n}\Big)$$
$$\lesssim \exp(-c_2 \log(n)), \quad (11)$$

where C_2 and c_2 are universal positive constants that depend only on Θ.

Proof of Proposition 4.3 is in Appendix B.2. It follows from the above proposition that the density estimation rate under the modified softmax gating multinomial logistic MoE model is parametric of order $\widetilde{\mathcal{O}}(n^{-1/2})$. This result matches the rate for estimating the true density under the multinomial logistic MoE model with standard softmax gating function in Proposition 2.2. Based on this observation, we then capture the convergence behavior of the MLE $\widetilde{G}_n$ in the following theorem.

Theorem 4.4 (Parameter Estimation Rate). *The following Hellinger lower bound holds true for any mixing measure $G \in \mathcal{O}_k(\Theta)$:*

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

This lower bound together with Proposition 4.3 imply that there exists a constant $C_3 > 0$ depending on Θ and G_* such that

$$\mathbb{P}(\mathcal{D}_2(\widetilde{G}_n, G_*) > C_3\sqrt{\log(n)/n}) \lesssim \exp(-c_3 \log(n)),$$

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

Proof of Theorem 4.4 is deferred to Appendix A.3. This theorem reveals that the MLE $\widetilde{G}_n$ converges to the true mixing measure G_* under the Voronoi loss function $\mathcal{D}_2$ at a rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$. From the definition of $\mathcal{D}_2$ in equation (5), we deduce that the exact-specified parameters $\beta_{1j}^*, a_j^*, b_j^*$, which are fitted by exactly one component, enjoy the same estimation rate of $\widetilde{\mathcal{O}}(n^{-1/2})$, whereas those for their over-specified counterparts, i.e. those fitted by at least two components, are of order $\widetilde{\mathcal{O}}(n^{-1/4})$. It is worth emphasizing that these rates remain stable regardless of any values of the expert parameters b_i^*. This highlights the benefits of modified softmax gating functions over the standard softmax gating in the multinomial logistic MoE model.

In summary, replacing the standard softmax gating function with its modified versions in the multinomial logistic MoE

model remains the identifiability of that model and the parametric density estimation rate as well as ensures the stability of parameter estimation rates under any settings of the expert parameters (see Table 1). As a consequence, we can conclude that the modified softmax gating functions outperform their standard counterpart in the parameter estimation problem under the multinomial logistic MoE model.

5. Discussion

In this paper, we investigate the convergence behavior of maximum likelihood estimation in the softmax gating multinomial logistic mixture of experts under the over-specified settings. For that purpose, we design a novel generic Voronoi loss function, and then discover that the rates for estimating the true density and exact-specified true parameters are all parametric on the sample size, while those for over-specified true parameters are slightly slower. However, when part of the expert parameters vanishes, the estimation rates for the over-specified true parameters experience a surprisingly slow rates due to an intrinsic interaction between the parameters of gating and expert functions. To tackle this issue, we propose a novel class of modified softmax gating functions that not only keep the identifiability of the multinomial logistic mixture of experts and the parametric density estimation rate unchanged, but also stabilize the parameter estimation rates irrespective of the collapse of expert parameters. This highlights the advantages of using modified softmax gating functions over the standard softmax gating function in the parameter estimation problem of the multinomial logistic MoE model.

Technical novelty. Compared to previous works, our paper is technically novel in terms of the following aspects:

(1) Covariate-dependent gate: We consider softmax gate whose value depends on the covariate X, while (Ho et al., 2022) use covariate-free weights, which are significantly simpler. Thus, in Step 1 of our proofs of Theorems 3.1 and 4.4, if we apply Taylor expansions directly to the density discrepancy $g_{G_n}(Y|X) - g_{G_*}(Y|X)$ as in (Ho et al., 2022), we cannot represent that discrepancy as a combination of elements from some linearly independent set, which is a key step. Therefore, we have to take the product of the softmax's denominator and the density discrepancy, denoted by $T_n(s) := \left[\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*) \right] \cdot [g_{G_n}(Y = s|X) - g_{G_*}(Y = s|X)]$. Then, we decompose $T_n(s)$ such that it includes two functions $\exp((\beta_{1i}^n)^\top X) f(Y = s|X; a_i^n, b_i^n)$ and $\exp((\beta_{1i}^n)^\top X) g_{G_n}(Y = s|X)$ (see equation (14)). Then, we need to apply two Taylor expansions to those functions rather than only one as in (Ho et al., 2022).

(2) Minimax lower bound: A key step to derive polynomial rates for estimating parameters (even in extreme cases) in

(Nguyen et al., 2023; 2024b) is to establish Total Variation lower bounds $\mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))] \gtrsim \mathcal{D}(G, G_*)$, where $\mathcal{D}$ is a loss function. However, we show in Proposition 3.2 that such bound does not hold true under the Regime 2 due to the PDEs in equation (3). Thus, we provide a minimax lower bound in Theorem 3.3 to show that the parameter estimation rates are slower than any polynomial rates, and therefore, could be as slow as $\mathcal{O}(1/\log(n))$. As far as we are concerned, this phenomenon has never been observed in previous works.

(3) Non-trivial solution for rate improvement: (Ho et al., 2022; Nguyen et al., 2023; 2024b) characterized the slowest rates of the same form $\mathcal{O}(n^{-1/2\bar{r}})$, for some $\bar{r} \geq 4$, without proposing any solutions to improve them. By contrast, although the slowest rate in our work is even slower than any polynomial rates, we provide a non-trivial modified gating function in Section 4 to alleviate that issue. More importantly, since that modified gate generally helps address interactions among gating and expert parameters, it can be applied to (Ho et al., 2022; Nguyen et al., 2023; 2024b) (up to some changes of the conditions of function $M(\cdot)$ in Definition 4.1) for rate improvement. As shown in Table 1, the modified softmax gate enhances the estimation rates from $\mathcal{O}(n^{-1/2\bar{r}})$ to at least $\mathcal{O}(n^{-1/4})$ under any parameter settings. To the best of our knowledge, such flexible and effective solution had never been proposed in the literature.

Limitation and future directions. In this paper, we consider a well-specified setting when the data are assumed to be sampled from a softmax gating multinomial logistic MoE. However, in practice, the data are not necessarily generated from that model, which we refer to as the misspecified setting. Under the misspecified setting, the MLE converges to a mixing measure $\widetilde{G} \in \arg\min_{G \in \mathcal{O}_k(\Theta)} \mathrm{KL}(P(Y|X)||g_G(Y|X))$, where $P(Y|X)$ is the true conditional distribution of Y given X and KL stands for the Kullback-Leibler divergence. As the space $\mathcal{O}_k(\Theta)$ is non-convex, the existence of $\widetilde{G}$ is not unique. Furthermore, the current analysis of the MLE under the misspecified setting of statistical models is mostly conducted when the function space is convex (van de Geer, 2000). Thus, it is necessary to develop new technical tools to establish the convergence rate of the MLE under the non-convex misspecified setting. Since this is beyond the scope of our work, we leave it for future development. Another potential direction is to comprehend the effects of the temperature parameter (Nie et al., 2022; Nguyen et al., 2024a), which controls the softmax weight distribution and the sparsity of the MoE, on the convergence of parameter estimation under the softmax gating multinomial logistic MoE. This direction has remained unexplored in the literature.

Acknowledgements

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

Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O.-K., Aggarwal, K., Som, S., Piao, S., and Wei, F. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems*, 2022. (Cited on page 1.)

Chen, K., Xu, L., and Chi, H. Improved learning algorithms for mixture of experts in multiclass classification. *Neural Networks*, 12(9):1229–1252, 1999. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(99)00043-X. (Cited on page 2.)

Chen, Z., Wang, P., Ma, L., Wong, K. K., and Wu, Q. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *CVPR*, 2023. (Cited on page 1.)

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. (Cited on page 1.)

Deleforge, A., Forbes, F., and Horaud, R. High-dimensional regression with Gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25(5): 893–911, 2015. ISSN 1573-1375. doi: 10.1007/s11222-014-9461-5. URL https://doi.org/10.1007/s11222-014-9461-5. (Cited on page 1.)

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, September 1977. ISSN 0035-9246. (Cited on page 29.)

Do, T., Le, H., Nguyen, T., Pham, Q., Nguyen, B., Doan, T., Liu, C., Ramasamy, S., Li, X., and HOI, S. HyperRouter: Towards efficient training and inference of sparse mixture of experts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, December 2023. Association for Computational Linguistics. (Cited on page 1.)

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy. (Cited on page 1.)

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. (Cited on page 1.)

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

Grün, B. and Leisch, F. Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification*, 25(2):225–247, 2008. (Cited on pages 28 and 29.)

Gulati, A., Qin, J., Chiu, C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. Interspeech 2020*, pp. 5036–5040, 2020. doi: 10.21437/Interspeech.2020-3015. (Cited on page 1.)

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. (Cited on page 1.)

Hazimeh, H., Zhao, Z., Chowdhery, A., Sathiamoorthy, M., Chen, Y., Mazumder, R., Hong, L., and Chi, E. H. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. In *NeurIPS*, 2021. (Cited on page 1.)

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. (Cited on pages 1 and 9.)

Huynh, B. and Chamroukhi, F. Estimation and feature selection in mixtures of generalized linear experts models. *arXiv preprint arXiv:1907.06994*, 2019. (Cited on page 2.)

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural Computation*, 3, 1991. (Cited on pages 1 and 2.)

Jordan, M. I. and Jacobs, R. A. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6: 181–214, 1994. (Cited on pages 1 and 2.)

Kwon, J. and Caramanis, C. EM converges for a mixture of many linear regressions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1727–1736. PMLR, 26–28 Aug 2020. (Cited on page 30.)

Kwon, J., Qian, W., Caramanis, C., Chen, Y., and Davis, D. Global convergence of the em algorithm for mixtures of two component linear regression. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2055–2110. PMLR, 25–28 Jun 2019. (Cited on page 30.)

Kwon, J., Ho, N., and Caramanis, C. On the minimax optimality of the em algorithm for learning two-component mixed linear regression. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1405–1413. PMLR, 13–15 Apr 2021. (Cited on page 30.)

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., and Chen, Z. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. (Cited on page 1.)

Liang, H., Fan, Z., Sarkar, R., Jiang, Z., Chen, T., Zou, K., Cheng, Y., Hao, C., and Wang, Z. M^{3}ViT: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. In *NeurIPS*, 2022. (Cited on page 1.)

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. (Cited on page 4.)

Nguyen, H., Nguyen, T., and Ho, N. Demystifying softmax gating function in Gaussian mixture of experts. In *Advances in Neural Information Processing Systems*, 2023. (Cited on pages 1, 2, 6, and 9.)

Nguyen, H., Akbarian, P., and Ho, N. Is temperature sample efficient for softmax Gaussian mixture of experts? In *Proceedings of the ICML*, 2024a. (Cited on page 9.)

Nguyen, H., Akbarian, P., Yan, F., and Ho, N. Statistical perspective of top-k sparse softmax gating mixture of experts. In *International Conference on Learning Representations*, 2024b. (Cited on pages 1 and 9.)

Nguyen, H., Nguyen, T., Nguyen, K., and Ho, N. Towards convergence rates for parameter estimation in Gaussian-gated mixture of experts. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, 2024c. (Cited on page 1.)

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. (Cited on page 9.)

Peng, F., Jacobs, R., and Tanner, M. Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association*, 91(435):953–960, 1996. ISSN 01621459. (Cited on page 1.)

Pham, N. and Chamroukhi, F. Functional mixture-of-experts for classification. In *53èmes Journées de Statistique de la Société Française de Statistique (SFdS)*, 2022. (Cited on page 2.)

Ren, J., Li, Y., Ding, Z., Pan, W., and Dong, H. Probabilistic mixture-of-experts for efficient deep reinforcement learning. *arxiv preprint arxiv 2104.09122*, 2021. (Cited on page 1.)

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. (Cited on page 1.)

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 *International Conference on Learning Representations*, 2017. (Cited on page 1.)

van de Geer, S. *Empirical Processes in M-estimation*. Cambridge University Press, 2000. (Cited on pages 9, 25, and 27.)

Villani, C. *Topics in Optimal Transportation*. American Mathematical Society, 2003. (Cited on page 1.)

Villani, C. *Optimal transport: Old and New*. Springer, 2008. (Cited on page 1.)

You, Z., Feng, S., Su, D., and Yu, D. Speechmoe2: Mixture-of-experts model with improved routing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7217–7221, 2022. (Cited on page 1.)

Yu, B. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*, pp. 423–435, 1997. (Cited on page 7.)

Yuksel, S. and Gader, P. Variational mixture of experts for classification with applications to landmine detection. In *2010 20th International Conference on Pattern Recognition*, pp. 2981–2984, 2010. doi: 10.1109/ICPR.2010.730. (Cited on page 2.)

Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A. M., Chen, Z., Le, Q., and Laudon, J. Mixture-of-experts with expert choice routing. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7103–7114. Curran Associates, Inc., 2022. (Cited on page 1.)

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. (Cited on page 1.)

Supplementary Materials for "A General Theory for Softmax Gating Multinomial Logistic Mixture of Experts"

In this supplementary material, we first present rigorous proofs for results regarding the convergence rates of parameter estimation under the (modified) softmax gating multinomial logistic mixture of experts in Appendix A, while those for density estimation are then provided in Appendix B. Next, we theoretically verify the identifiability of the proposed models in Appendix C. Lastly, we conduct several experiments in Appendix D to empirically validate our theoretical results.

A. Proof for Convergence Rates of Parameter Estimation

In this appendix, we provide proofs for Theorem 3.1 and Theorem 3.3 in Appendix A.1 and Appendix A.2, respectively. Meanwhile, the proof of Theorem 4.4 can be found in Appendix A.3.

A.1. Proof of Theorem 3.1

To reach the conclusion in Theorem 3.1, we need to show that

$$
\inf_{G \in \mathcal{O}_k(\Theta)} \mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_2(G, G_*) > 0. \tag{12}
$$

Local Structure: Firstly, we prove by contradiction the local structure of inequality (12), which is given by

$$
\lim_{\varepsilon \to 0} \inf_{\substack{G \in \mathcal{O}_k(\Theta), \\ \mathcal{D}_2(G, G_*) \le \varepsilon}} \mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_2(G, G_*) > 0. \tag{13}
$$

Assume that this claim does not hold true, then there exists a sequence of mixing measures $G_n := \sum_{i=1}^{k_n} \exp(\beta_{0i}^n) \delta_{(\beta_{1i}^n, a_i^n, b_i^n)} \in \mathcal{O}_k(\Theta)$ such that both $\mathbb{E}_X[V(g_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_2(G_n, G_*)$ and $\mathcal{D}_2(G_n, G_*)$ approach zero as n tends to infinity. In addition, we define the set of Voronoi cells generated by the support of G_* for this sequence as follows:

$$
\mathcal{C}_j^n := \{i \in [k_n] : \|\theta_i^n - \theta_j^*\| \le \|\theta_i^n - \theta_\ell^*\|, \quad \forall \ell \ne j\},
$$

where $\theta_i^n := (\beta_{1i}^n, a_i^n, b_i^n)$ and $\theta_j^* := (\beta_{1j}^*, a_i^*, b_i^*)$ for any $j \in [k_*]$. Recall that $\beta_{1k_*}^* = \mathbf{0}_d$ is known, thus we set $\beta_{1i}^n = \mathbf{0}_d$ for any $i \in \mathcal{C}_{k_*}$. Similarly, note that $(a_{jK}^*, b_{jK}^*) = (0, \mathbf{0}_d)$ are known for any $j \in [k_*]$, we also let $(a_{iK}^n, b_{iK}^n) = (0, \mathbf{0}_d)$ for any $i \in k_n$. As our proof argument is asymptotic, we can assume without loss of generality (WLOG) that the above Voronoi cells are independent of n, that is, $\mathcal{C}_j = \mathcal{C}_j^n$ for any $j \in [k_*]$. Next, since $\mathcal{D}_2(G_n, G_*) \to 0$ as $n \to \infty$, it follows from the formulation of $\mathcal{D}_2$ in equation (5) that $\beta_{1i}^n \to \beta_{1j}^*$, $(a_{i\ell}^n, b_{i\ell}^n) \to (a_{j\ell}^*, b_{j\ell}^*)$ and $\sum_{i \in \mathcal{C}_j^n} \exp(\beta_{0i}^n) \to \exp(\beta_{0j}^*)$ for any $\ell \in [K], i \in \mathcal{C}_j$ and $j \in [k_*]$ when $n \to \infty$. Subsequently, we divide our remaining arguments into three steps as below.

Step 1. In this step, we use Taylor expansions to decompose the following quantity:

$$
T_n(s) := \Big[\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)\Big] \cdot \Big[g_{G_n}(Y = s|X) - g_{G_*}(Y = s|X)\Big].
$$

Denote $u(Y = s|X; \beta_{1i}, a_i, b_i) := \exp(\beta_{1i}^\top X) \cdot f(Y|X; a_{is}, b_{is})$ and $v(Y = s|X; \beta_{1i}) = \exp(\beta_{1i}^\top X) g_{G_n}(Y = s|X)$ for any $s \in [K]$, we have

$$
\begin{aligned}
T_n(s) = & \sum_{j=1}^{k_*} \sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n)\Big[u(Y = s|X; \beta_{1i}^n, a_i^n, b_i^n) - u(Y = s|X; \beta_{1j}^*, a_j^*, b_j^*)\Big] \\
& - \sum_{j=1}^{k_*} \sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n)\Big[v(Y = s|X; \beta_{1i}^n) - v(Y = s|X; \beta_{1j}^*)\Big] \\
& + \sum_{j=1}^{k_*} \Big(\sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^*)\Big)\Big[u(Y = s|X; \beta_{1j}^*, a_j^*, b_j^*) - v(Y = s|X; \beta_{1j}^*)\Big] \\
& := A_n - B_n + E_n. \tag{14}
\end{aligned}
$$

Given the above formulations of A_n and B_n, we continue to decompose them into two smaller terms based on the cardinality of Voronoi cells $\mathcal{C}_j$. In particular,

$$
A_n = \sum_{j:|\mathcal{C}_j|=1} \sum_{i\in\mathcal{C}_j} \exp(\beta_{0i}^n) \Big[u(Y=s|X;\beta_{1i}^n, a_i^n, b_i^n) - u(Y=s|X;\beta_{1j}^*, a_j^*, b_j^*) \Big]
$$

$$
+ \sum_{j:|\mathcal{C}_j|>1} \sum_{i\in\mathcal{C}_j} \exp(\beta_{0i}^n) \Big[u(Y=s|X;\beta_{1i}^n, a_i^n, b_i^n) - u(Y=s|X;\beta_{1j}^*, a_j^*, b_j^*) \Big]
$$

$$
= A_{n,1} + A_{n,2}.
$$

Next, let us denote $h_\ell(X, a_\ell, b_\ell) := a_\ell + b_\ell^\top X$ for any $\ell \in [K]$, then

$$
f(Y=s|X;a_i, b_i) = \frac{\exp(a_{is} + b_{is}^\top X)}{\sum_{\ell=1}^K \exp(a_{i\ell} + b_{i\ell}^\top X)} = \frac{\exp(h_s(X, a_{is}, b_{is}))}{\sum_{\ell=1}^K \exp(h_\ell(X, a_{i\ell}, b_{i\ell}))}.
$$

By means of the first-order Taylor expansion, $A_{n,1}$ can be represented as

$$
A_{n,1} = \sum_{j:|\mathcal{C}_j|=1} \sum_{i\in\mathcal{C}_j} \exp(\beta_{0i}^n) \sum_{|\alpha|=1} \frac{1}{\alpha!}(\Delta\beta_{1ij}^n) \prod_{\ell=1}^{K-1} (\Delta a_{ij\ell}^n)^{\alpha_{2\ell}}(\Delta b_{ij\ell}^n)^{\alpha_{3\ell}} \frac{\partial^{|\alpha|}u(Y=s|X;\beta_{1j}^*, a_j^*, b_j^*)}{\partial\beta_{1j}^{\alpha_1} \prod_{\ell=1}^{K-1} \partial a_{j\ell}^{\alpha_{2\ell}} \partial b_{j\ell}^{\alpha_{3\ell}}} + R_1(X,Y).
$$

Here, $\alpha := (\alpha_1, \alpha_{21}, \ldots, \alpha_{2(K-1)}, \alpha_{31}, \ldots, \alpha_{3(K-1)})$, where $\alpha_1, \alpha_{3\ell} \in \mathbb{N}^d$ and $\alpha_{2\ell} \in \mathbb{N}$ for any $\ell \in [K-1]$. Additionally, $R_1(X,Y)$ is a Taylor remainder such that $R_1(X,Y)/\mathcal{D}_2(G_n, G_*) \to 0$ as $n \to \infty$. From the formulation of u, we have

$$
A_{n,1} = \sum_{j:|\mathcal{C}_j|=1} \sum_{i\in\mathcal{C}_j} \exp(\beta_{0i}^n) \sum_{|\alpha|=1} \frac{1}{\alpha!}(\Delta\beta_{1ij}^n) \prod_{\ell=1}^{K-1} (\Delta a_{ij\ell}^n)^{\alpha_{2\ell}}(\Delta b_{ij\ell}^n)^{\alpha_{3\ell}}
$$

$$
\times X^{\alpha_1 + \sum_{\ell=1}^{K-1}\alpha_{3\ell}} \exp((\beta_{1j}^*)^\top X) \frac{\partial^{\sum_{\ell=1}^{K-1}(\alpha_{2\ell}+|\alpha_{3\ell}|)}f}{\partial h_1^{\alpha_{21}+|\alpha_{31}|} \ldots \partial h_{K-1}^{\alpha_{2(K-1)}+|\alpha_{3(K-1)}|}}(Y=s|X;a_j^*, b_j^*) + R_1(X,Y).
$$

Let $q_1 = \alpha_1 + \sum_{\ell=1}^{K-1}\alpha_{3\ell} \in \mathbb{N}^d$, $q_2 = (q_{2\ell})_{\ell=1}^{K-1} := (\alpha_{2\ell} + |\alpha_{3\ell}|)_{\ell=1}^{K-1} \in \mathbb{N}^{K-1}$ and

$$
\mathcal{I}_{q_1,q_2} := \left\{ \alpha = (\alpha_1, \alpha_{21}, \ldots, \alpha_{2(K-1)}, \alpha_{31}, \ldots, \alpha_{3(K-1)}) : \alpha_1 + \sum_{\ell=1}^{K-1}\alpha_{3\ell} = q_1, \ (\alpha_{2\ell}+|\alpha_{3\ell}|)_{\ell=1}^{K-1} = q_2 \right\}, \quad (15)
$$

we can rewrite $A_{n,1}$ as

$$
A_{n,1} = \sum_{j:|\mathcal{C}_j|=1} \sum_{|q_1|+|q_2|=1} \sum_{i\in\mathcal{C}_j} \sum_{\alpha\in\mathcal{I}_{q_1,q_2}} \frac{1}{\alpha!}(\Delta\beta_{1ij}^n)^{\alpha_1} \prod_{\ell=1}^{K-1} (\Delta a_{ij\ell}^n)^{\alpha_{2\ell}}(\Delta b_{ij\ell}^n)^{\alpha_{3\ell}}
$$

$$
\times X^{q_1} \exp((\beta_{1j}^*)^\top X) \frac{\partial^{|q_2|}f}{\partial h_1^{q_{21}} \ldots \partial h_{K-1}^{q_{2(K-1)}}}(Y=s|X;a_j^*, b_j^*) + R_1(X,Y).
$$

Similarly, by means of second order Taylor expansion, $A_{n,2}$ is expressed as follows:

$$
A_{n,2} = \sum_{j:|\mathcal{C}_j|>1} \sum_{|q_1|+|q_2|=1}^{4} \sum_{i\in\mathcal{C}_j} \sum_{\alpha\in\mathcal{I}_{q_1,q_2}} \frac{1}{\alpha!}(\Delta\beta_{1ij}^n)^{\alpha_1} \prod_{\ell=1}^{K-1} (\Delta a_{ij\ell}^n)^{\alpha_{2\ell}}(\Delta b_{ij\ell}^n)^{\alpha_{3\ell}}
$$

$$
\times X^{q_1} \exp((\beta_{1j}^*)^\top X) \frac{\partial^{|q_2|}f}{\partial h_1^{q_{21}} \ldots \partial h_{K-1}^{q_{2(K-1)}}}(Y=s|X;a_j^*, b_j^*) + R_2(X,Y),
$$

where $R_2(X,Y)$ is also a Taylor remainder such that $R_2(X,Y)/\mathcal{D}_2(G_n, G_*) \to 0$ as $n \to \infty$. By employing the same arguments for decomposing A_n to B_n, we get

$$
B_n = \sum_{j:|\mathcal{C}_j|=1} \sum_{|\gamma|=1} \sum_{i\in\mathcal{C}_j} \frac{\exp(\beta_{0i}^n)}{\gamma!}(\Delta\beta_{1ij}^n)^\gamma \times X^\gamma \exp((\beta_{1j}^*)^\top X)g_{G_n}(Y=s|X) + R_3(X,Y)
$$

$$
+ \sum_{j:|\mathcal{C}_j|>1} \sum_{|\gamma|=1}^{2} \sum_{i\in\mathcal{C}_j} \frac{\exp(\beta_{0i}^n)}{\gamma!}(\Delta\beta_{1ij}^n)^\gamma \times X^\gamma \exp((\beta_{1j}^*)^\top X)g_{G_n}(Y=s|X) + R_4(X,Y).
$$

Putting the above results together, we obtain that

$$
T_n(s) = \sum_{j=1}^{k_*} \sum_{|q_1|+|q_2|=0}^{2+2\cdot\mathbf{1}_{\{|\mathcal{C}_j|>1\}}} U_{q_1,q_2}^n(j) \times X^{q_1} \exp((\beta_{1j}^*)^\top X) \frac{\partial^{|q_2|} f}{\partial h_1^{q_{21}} \dots \partial h_{K-1}^{q_{2(K-1)}}} (Y=s|X; a_j^*, b_j^*)
$$
$$
+ \sum_{j=1}^{k_*} \sum_{|\gamma|=0}^{1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}}} W_\gamma^n(j) \times X^\gamma \exp((\beta_{1j}^*)^\top X) g_{G_n}(Y=s|X) + R(X,Y),
$$

where $R(X,Y)$ is the sum of Taylor remainders such that $R(X,Y)/\mathcal{D}_2(X,Y) \to 0$ as $n \to \infty$ and

$$
U_{q_1,q_2}^n(j) = \begin{cases} \sum_{i\in\mathcal{C}_j} \sum_{\alpha\in\mathcal{I}_{q_1,q_2}} \frac{\exp(\beta_{0i}^n)}{\alpha!} (\Delta\beta_{1ij}^n)^{\alpha_1} \prod_{\ell=1}^{K-1} (\Delta a_{ij\ell}^n)^{\alpha_{2\ell}} (\Delta b_{ij\ell}^n)^{\alpha_{3\ell}}, & (q_1,q_2) \neq (\mathbf{0}_d, \mathbf{0}_{K-1}), \\ \sum_{i\in\mathcal{C}_j} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^*), & (q_1,q_2) = (\mathbf{0}_d, \mathbf{0}_{K-1}), \end{cases}
$$

and

$$
W_\gamma^n(j) = \begin{cases} -\sum_{i\in\mathcal{C}_j} \frac{\exp(\beta_{0i}^n)}{\gamma!} (\Delta\beta_{1ij}^n)^\gamma, & |\gamma| \neq \mathbf{0}_d, \\ -\sum_{i\in\mathcal{C}_j} \exp(\beta_{0i}^n) + \exp(\beta_{0j}^*), & |\gamma| = \mathbf{0}_d, \end{cases}
$$

for any $j \in [k_*]$.

Step 2. In this step, we will prove by contradiction that at least one among $U_{q_1,q_2}^n(j)/\mathcal{D}_2(G_n, G_*)$ and $W_\gamma^n(j)/\mathcal{D}_2(G_n, G_*)$ does not vanish as n tends to infinity. Assume that all of them approach zero, then by taking the summation of $|U_{q_1,q_2}^n(j)|/\mathcal{D}_2(G_n, G_*)$ for $j \in [k_*] : |\mathcal{C}_j| = 1$, $q_1 \in \{e_1, e_2, \dots, e_d\}$ and $q_2 = \mathbf{0}_{K-1}$, where $e_i := (0, \dots, 0, \underbrace{1}_{i\text{-}th}, 0, \dots, 0) \in \mathbb{R}^d$, we achieve that

$$
\frac{1}{\mathcal{D}_2(G_n, G_*)} \cdot \sum_{j:|\mathcal{C}_j|=1} \sum_{i\in\mathcal{C}_j} \exp(\beta_{0i}^n) \cdot \|\Delta\beta_{1ij}^n\|_1 \to 0. \tag{16}
$$

Similarly, for $q_1 = \mathbf{0}_d$ and $q_2 \in \{e_1', e_2', \dots, e_{K-1}'\}$, where $e_\ell' := (0, \dots, 0, \underbrace{1}_{\ell\text{-}th}, 0, \dots, 0) \in \mathbb{R}^{K-1}$, we have

$$
\frac{1}{\mathcal{D}_2(G_n, G_*)} \cdot \sum_{j:|\mathcal{C}_j|=1} \sum_{i\in\mathcal{C}_j} \sum_{\ell=1}^{K-1} \exp(\beta_{0i}^n) \cdot |\Delta a_{ij\ell}^n| \to 0. \tag{17}
$$

On the other hand, for $q_1 \in \{e_1, e_2, \dots, e_d\}$ and $q_2 \in \{e_1', e_2', \dots, e_{K-1}'\}$, we obtain that

$$
\frac{1}{\mathcal{D}_2(G_n, G_*)} \cdot \sum_{j:|\mathcal{C}_j|=1} \sum_{i\in\mathcal{C}_j} \sum_{\ell=1}^{K-1} \exp(\beta_{0i}^n) \cdot \|\Delta b_{ij\ell}^n\|_1 \to 0. \tag{18}
$$

Combine the limits in equations (16), (17) and (18), we get

$$
\frac{1}{\mathcal{D}_2(G_n, G_*)} \cdot \sum_{j:|\mathcal{C}_j|=1} \sum_{i\in\mathcal{C}_j} \exp(\beta_{0i}^n) \cdot \left[\|\Delta\beta_{1ij}^n\|_1 + \sum_{\ell=1}^{K-1} (|\Delta a_{ij\ell}^n| + \|\Delta b_{ij\ell}^n\|_1) \right] \to 0.
$$

Due to the topological equivalence between 1-norm and 2-norm, the above limit is equivalent to

$$
\frac{1}{\mathcal{D}_2(G_n, G_*)} \cdot \sum_{j:|\mathcal{C}_j|=1} \sum_{i\in\mathcal{C}_j} \exp(\beta_{0i}^n) \cdot \left[\|\Delta\beta_{1ij}^n\| + \sum_{\ell=1}^{K-1} (|\Delta a_{ij\ell}^n| + \|\Delta b_{ij\ell}^n\|) \right] \to 0. \tag{19}
$$

Next, we consider the summation of $|U_{q_1,q_2}^n(j)|/\mathcal{D}_2(G_n, G_*)$ for $j \in [k_*] : |\mathcal{C}_j| > 1$, $q_1 \in \{2e_1, 2e_2, \dots, 2e_d\}$ and $q_2 = \mathbf{0}_{K-1}$, which leads to

$$
\frac{1}{\mathcal{D}_2(G_n, G_*)} \cdot \sum_{j:|\mathcal{C}_j|>1} \sum_{i\in\mathcal{C}_j} \exp(\beta_{0i}^n) \cdot \|\Delta\beta_{1ij}^n\|^2 \to 0. \tag{20}
$$

For $q_1 = \mathbf{0}_d$ and $q_2 \in \{2e'_1, 2e'_2, \ldots, 2e'_{K-1}\}$, we have

$$\frac{1}{\mathcal{D}_2(G_n, G_*)} \cdot \sum_{j:|\mathcal{C}_j|>1} \sum_{i \in \mathcal{C}_j} \exp(\beta^n_{0i}) \cdot \sum_{\ell=1}^{K-1} |\Delta a^n_{ij\ell}|^2 \to 0. \tag{21}$$

Meanwhile, for $q_1 \in \{2e_1, 2e_2, \ldots, 2e_d\}$ and $q_2 \in \{2e'_1, 2e'_2, \ldots, 2e'_{K-1}\}$, we get

$$\frac{1}{\mathcal{D}_2(G_n, G_*)} \cdot \sum_{j:|\mathcal{C}_j|>1} \sum_{i \in \mathcal{C}_j} \exp(\beta^n_{0i}) \cdot \sum_{\ell=1}^{K-1} \|\Delta b^n_{ij\ell}\|^2 \to 0. \tag{22}$$

It follows from equations (20), (21) and (22) that

$$\frac{1}{\mathcal{D}_2(G_n, G_*)} \cdot \sum_{j:|\mathcal{C}_j|>1} \sum_{i \in \mathcal{C}_j} \exp(\beta^n_{0i}) \cdot \left[\|\Delta\beta^n_{1ij}\|^2 + \sum_{\ell=1}^{K-1} (|\Delta a^n_{ij\ell}|^2 + \|\Delta b^n_{ij\ell}\|^2) \right] \to 0. \tag{23}$$

Note that

$$\sum_{j=1}^{k_*} \frac{|U^n_{\mathbf{0}_d, \mathbf{0}_{K-1}}(j)|}{\mathcal{D}_2(G_n, G_*)} = \frac{1}{\mathcal{D}_2(G_n, G_*)} \cdot \sum_{j=1}^{k_*} \left| \sum_{i \in \mathcal{C}_j} \exp(\beta^n_{0i}) - \exp(\beta^*_{0j}) \right| \to 0. \tag{24}$$

By taking the sum of limits in equations (19), (23) and (24), we deduce that $1 = \mathcal{D}_2(G_n, G_*)/\mathcal{D}_2(G_n, G_*) \to 0$ as $n \to \infty$, which is a contradiction. Thus, at least one among the limits of $U^n_{q_1,q_2}(j)/\mathcal{D}_2(G_n, G_*)$ and $W^n_\gamma(j)/\mathcal{D}_2(G_n, G_*)$ is non-zero.

Step 3. Finally, we will leverage Fatou's lemma to point out a contradiction to the result in Step 2.

Let us denote by m_n the maximum of the absolute values of $U^n_{q_1,q_2}(j)/\mathcal{D}_2(G_n, G_*)$ and $W^n_\gamma(j)/\mathcal{D}_2(G_n, G_*)$ for $j \in [k_*]$, $0 \le |q_1| + |q_2| \le 2 + 2 \cdot \mathbf{1}_{\{|\mathcal{C}_j|>1\}}$ and $0 \le |\gamma| \le 1 + \mathbf{1}_{\{|\mathcal{C}_j|>1\}}$. Then, it follows from the Fatou's lemma that

$$0 = \lim_{n \to \infty} \frac{\mathbb{E}_X[2V(g_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]}{m_n \mathcal{D}_2(G_n, G_*)} \ge \int \sum_{s=1}^{K} \liminf_{n \to \infty} \frac{|g_{G_n}(Y = s|X) - g_{G_*}(Y = s|X)|}{m_n \mathcal{D}_2(G_n, G_*)} \, \mathrm{d}X \ge 0.$$

As a result, we get that $|g_{G_n}(Y = s|X) - g_{G_*}(Y = s|X)|/[m_n \mathcal{D}_2(G_n, G_*)]$ converges to zero, which implies that $T_n(s)/[m_n \mathcal{D}_2(G_n, G_*)] \to 0$ as $n \to \infty$ for any $s \in [K]$ and almost surely X. Let $U^n_{q_1,q_2}(j)/[m_n \mathcal{D}_2(G_n, G_*)] \to \tau_{q_1,q_2}(j)$ and $W^n_\gamma(j) \to \eta_\gamma(j)$ as n approaches infinity, then the previous result indicates that

$$\sum_{j=1}^{k_*} \sum_{|q_1|+|q_2|=0}^{2+2\cdot\mathbf{1}_{\{|\mathcal{C}_j|>1\}}} \tau_{q_1,q_2}(j) \times X^{q_1} \exp((\beta^*_{1j})^\top X) \frac{\partial^{|q_2|} f}{\partial h_1^{q_{21}} \ldots \partial h_{K-1}^{q_{2(K-1)}}} (Y = s|X; a^*_j, b^*_j)$$

$$+ \sum_{j=1}^{k_*} \sum_{|\gamma|=0}^{1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}}} \eta_\gamma(j) \times X^\gamma \exp((\beta^*_{1j})^\top X) g_{G_*}(Y = s|X) = 0. \tag{25}$$

for any $s \in [K]$ and almost surely X. Here, at least one among $\tau_{q_1,q_2}(j)$ and $\eta_\gamma(j)$ is different from zero. Assume the set

$$\mathcal{F} := \left\{ X^{q_1} \exp((\beta^*_{1j})^\top X) \frac{\partial^{|q_2|} f}{\partial h_1^{q_{21}} \ldots \partial h_{K-1}^{q_{2(K-1)}}} (Y = s|X; a^*_j, b^*_j) : j \in [k_*], 0 \le |q_1| + |q_2| \le 2 + 2 \cdot \mathbf{1}_{\{|\mathcal{C}_j|>1\}} \right\}$$

$$\cup \left\{ X^\gamma \exp((\beta^*_{1j})^\top X) g_{G_*}(Y = s|X) : j \in [k_*], 0 \le |\gamma| \le 1 + \mathbf{1}_{\{|\mathcal{C}_j|>1\}} \right\} \tag{26}$$

is linearly independent, we deduce that $\tau_{q_1,q_2}(j) = \eta_\gamma(j) = 0$ for any $j \in [k_*]$, $0 \le |q_1| + |q_2| \le 2 + 2 \cdot \mathbf{1}_{\{|\mathcal{C}_j|>1\}}$ and $0 \le |\gamma| \le 1 + \mathbf{1}_{\{|\mathcal{C}_j|>1\}}$, which is a contradiction.

Thus, it suffices to show that $\mathcal{F}$ is a linearly independent set to attain the local inequality in equation (13). In particular, assume that equation (25) holds true for any $s \in [K]$ and almost surely X, we will show that $\tau_{q_1,q_2}(j) = \eta_\gamma(j) = 0$ for any $j \in [k_*]$, $0 \leq |q_1| + |q_2| \leq 2 + 2 \cdot \mathbf{1}_{\{|\mathcal{C}_j|>1\}}$ and $0 \leq |\gamma| \leq 1 + \mathbf{1}_{\{|\mathcal{C}_j|>1\}}$. Firstly, we rewrite this equation as follows:

$$
\sum_{j=1}^{k_*} \sum_{|\omega|=0}^{1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}}} \left[\sum_{|q_2|=0}^{2+2\cdot\mathbf{1}_{\{|\mathcal{C}_j|>1\}}-|\omega|} \tau_{q_1,q_2}(j) \frac{\partial^{|q_2|} f}{\partial h_1^{q_{21}} \ldots \partial h_{K-1}^{q_{2(K-1)}}} (Y = s|X; a_j^*, b_j^*) + \eta_\omega(j) g_{G_*}(Y = s|X) \right]
$$
$$
\times X^\omega \exp((\beta_{1j}^*)^\top X) = 0,
$$

for any $s \in [K]$ and almost surely X. Note that $\beta_{11}^*, \beta_{12}^*, \ldots, \beta_{1k_*}^*$ are k_* different values, therefore, it can be seen that $\{X^\omega \exp((\beta_{1j}^*)^\top X) : j \in [k_*], \ 0 \leq |\omega| \leq 1 + \mathbf{1}_{\{|\mathcal{C}_j|>1\}}\}$ is a linearly independent set. As a result, we obtain that

$$
\sum_{|q_2|=0}^{2+2\cdot\mathbf{1}_{\{|\mathcal{C}_j|>1\}}-|\omega|} \tau_{q_1,q_2}(j) \frac{\partial^{|q_2|} f}{\partial h_1^{q_{21}} \ldots \partial h_{K-1}^{q_{2(K-1)}}} (Y = s|X; a_j^*, b_j^*) + \eta_\omega(j) g_{G_*}(Y = s|X) = 0,
$$

for any $j \in [k_*]$, $0 \leq |\omega| \leq 1 + \mathbf{1}_{\{|\mathcal{C}_j|>1\}}$, $s \in [K]$ and almost surely X. Similarly, the following set is linearly independent:

$$
\left\{ \frac{\partial^{|q_2|} f}{\partial h_1^{q_{21}} \ldots \partial h_{K-1}^{q_{2(K-1)}}} (Y = s|X; a_j^*, b_j^*), \ g_{G_*}(Y = s|X) : 0 \leq |q_2| \leq 2 + 2 \cdot \mathbf{1}_{\{|\mathcal{C}_j|>1\}} - |\omega| \right\},
$$

we achieve that $\tau_{q_1,q_2}(j) = \eta_\gamma(j) = 0$ for any $j \in [k_*]$, $0 \leq |q_1| + |q_2| \leq 2 + 2 \cdot \mathbf{1}_{\{|\mathcal{C}_j|>1\}}$ and $0 \leq |\gamma| \leq 1 + \mathbf{1}_{\{|\mathcal{C}_j|>1\}}$, which completes the proof of local structure.

Global Structure: As the local inequality in equation (13) holds true, there exists a positive constant ε' that satisfies

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

Therefore, it is sufficient to demonstrate that

$$
\inf_{\substack{G \in \mathcal{O}_k(\Theta), \\ \mathcal{D}_2(G,G_*) > \varepsilon'}} \mathbb{E}_X[V(g_G(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_2(G, G_*) > 0. \tag{27}
$$

Assume by contrary that this inequality does not hold, then we can find a sequence $G_n' \in \mathcal{O}_k(\Theta)$ such that $\mathcal{D}_2(G_n', G_*) > \varepsilon'$ and $\mathbb{E}_X[V(g_{G_n'}(\cdot|X), g_{G_*}(\cdot|X))] \to 0$ as n tends to infinity. It is worth noting that Θ is a compact set, therefore, we can replace G_n' by its subsequence which converges to some mixing measure $G' \in \mathcal{O}_k(\Theta)$. Thus, we get that $\mathcal{D}_2(G', G_*) > \varepsilon'$. On the other hand, according to Fatou's lemma,

$$
0 = \lim_{n\to\infty} \frac{\mathbb{E}_X[2V(g_{G_n'}(\cdot|X), g_{G_*}(\cdot|X))]}{\mathcal{D}_2(G_n', G_*)} \geq \int \sum_{s=1}^{K} \liminf_{n\to\infty} |g_{G_n'}(Y = s|X) - g_{G_*}(Y = s|X)| \, dX.
$$

Consequently, it follows that

$$
\int \sum_{s=1}^{K} |g_{G'}(Y = s|X) - g_{G_*}(Y = s|X)| \, dX = 0,
$$

which means that $g_{G'}(Y = s|X) = g_{G_*}(Y = s|X)$ for any $s \in [K]$ for almost surely X. Recall that the softmax gating multinomial logistic mixture of experts is identifiable, we deduce that $G' \equiv G$, which contradicts the results that $\mathcal{D}_2(G', G_*) > \varepsilon'$. Hence, we obtain the global inequality in equation (27) and complete the proof.

A.2. Proof of Theorem 3.3

In Appendix A.2.1, we present the proof of Theorem 3.3 given the result of Proposition 3.2. Then, we provide the proof of Proposition 3.2 in Appendix A.2.2.

A.2.1. MAIN PROOF

Given a fixed constant $M_1 > 0$ and a sufficiently small $\varepsilon > 0$ that we will choose later, it follows from the result of Proposition 3.2 that there exists a mixing measure $G'_* \in \mathcal{O}_k(\Theta)$ that satisfies $\mathcal{D}_r(G'_*, G_*) = 2\varepsilon$ and $\mathbb{E}_X[V(g_{G'_*}(\cdot|X), g_{G_*}(\cdot|X))] \leq M_1 \varepsilon$. Note that for any sequence $\overline{G}_n \in \mathcal{O}_k(\Theta)$, we have

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

where $\mathbb{E}_{g_G}$ denotes the expectation taken with respect to the product measure with density g_G^n. Moreover, since the loss $\mathcal{D}_r$ satisfies the weak triangle inequality, we can find a constant $M_2 > 0$ such that

$$\mathcal{D}_r(\overline{G}_n, G_*) + \mathcal{D}_r(\overline{G}_n, G'_*) \geq M_2 \mathcal{D}_r(G_*, G'_*) = 2M_2 \varepsilon.$$

As a result, we observe that

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

$$\geq M_2 \varepsilon \cdot \inf_{f_1, f_2} \Big(\mathbb{E}_{g_{G_*}}[f_1] + \mathbb{E}_{g_{G'_*}}[f_2] \Big).$$

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

$$\max_{G \in \{G_*, G'_*\}} \mathbb{E}_{p_G}[\mathcal{D}_r(\overline{G}_n, G)] \geq M_2 \varepsilon \Big(1 - \mathbb{E}_X[V(g_{G_*}^n(\cdot|X), g_{G'_*}^n(\cdot|X))] \Big)$$

$$\geq M_2 \varepsilon \Big[1 - \sqrt{1 - (1 - M_1^2 \varepsilon^2)^n} \Big].$$

By choosing $\varepsilon = n^{-1/2}/M_1$, we have $M_1^2 \varepsilon^2 = \frac{1}{n}$, which implies that

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

for any mixing measure $\overline{G}_n \in \mathcal{O}_k(\Theta)$. Hence, we reach the conclusion of Theorem 3.3, which says that

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

for any $r \geq 1$.

A.2.2. PROOF OF PROPOSITION 3.2

From the conditions of Regime 2, we may assume without loss of generality that $b_{1\ell}^* = \mathbf{0}_d$ for any $\ell \in [K - 1]$. To reach the conclusion of Proposition 3.2, we need to find a sequence $G_n \in \mathcal{O}_k(\Theta)$ such that $\mathcal{D}_r(G_n, G_*)$ and $V(p_{G_n}, p_{G_*})/\mathcal{D}_r(G_n, G_*)$ both tend to zero when n approaches infinity. For that purpose, we choose $G_n = \sum_{i=1}^{k_*+1} \exp(\beta_{0i}^n) \delta_{(\beta_{1i}^n, a_{11}^n, \ldots, a_{1(K-1)}^n, b_{11}^n, \ldots, b_{1(K-1)}^n)}$ where

- $\exp(\beta_{01}^n) = \exp(\beta_{02}^n) = \frac{1}{2} \exp(\beta_{01}^*) - \frac{t_n}{2}$, $\exp(\beta_{0i}^n) = \exp(\beta_{0(i-1)}^*)$ for any $3 \leq i \leq k_* + 1$;

- $\beta_{11}^n = \beta_{12}^n = \beta_{11}^* + c_n \mathbf{1}_d$, $\beta_{1i}^n = \beta_{1(i-1)}^*$ for any $3 \leq i \leq k_* + 1$;

- $a_{1\ell}^n = a_{2\ell}^n = a_{1\ell}^* + c_n$, $a_{i\ell}^n = a_{(i-1)\ell}^*$ for any $\ell \in [K - 1]$ and $3 \leq i \leq k_* + 1$;

- $b_{1\ell}^n = b_{2\ell}^n = b_{1\ell}^*$, $b_{i\ell}^n = b_{(i-1)\ell}^*$ for any $\ell \in [K - 1]$ and $3 \leq i \leq k_* + 1$,

where $t_n, c_n > 0$ will be chosen later such that $t_n \to 0$ and $c_n = \mathcal{O}(t_n)$ as $n \to \infty$. Then, it can be verified that

$$\mathcal{D}_r(G_n, G_*) = (K - 1) \Big[\exp(\beta_{01}^*) - t_n \Big] c_n^r (1 + d^{r/2}) + t_n. \tag{28}$$

Now, we will show that $\mathbb{E}_X[V(g_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_r(G_n, G_*)$ vanishes as $n \to \infty$. Let us reconsider the quantity $T_n(s) = \left[\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)\right] \cdot \left[g_{G_n}(Y = s|X) - g_{G_*}(Y = s|X)\right]$ in equation (14) under the above setting of G_n as follows:

$$T_n(s) = \sum_{i=1}^{2} \exp(\beta_{0i}^n)[u(Y = s|X; \beta_{1i}^n, a_i^n, b_i^n) - u(Y = s|X; \beta_{11}^*, a_1^*, b_1^*)]$$

$$- \sum_{i=1}^{2} \exp(\beta_{0i}^n)[v(Y = s|X; \beta_{1i}^n) - v(Y = s|X; \beta_{11}^*)]$$

$$+ \left(\sum_{i=1}^{2} \exp(\beta_{0i}^n) - \exp(\beta_{01}^*)\right)[u(Y = s|X; \beta_{11}^*, a_1^*, b_1^*) - v(Y = s|X; \beta_{11}^*)]$$

$$:= A_n - B_n + E_n,$$

where $u(Y = s|X; \beta_1, a, b) := \exp(\beta_1^\top X)f(Y = s|X; a, b)$ and $v(Y = s|X; \beta_1) = \exp(\beta_1^\top X)g_{G_n}(Y = s|X)$ for any $s \in [K]$. Recall that we denote $h_\ell(X, a_\ell, b_\ell) := a_\ell + b_\ell^\top X$ for any $\ell \in [K]$ and

$$f(Y = s|X; a_i, b_i) = \frac{\exp(a_{is} + b_{is}^\top X)}{\sum_{\ell=1}^{K} \exp(a_{i\ell} + b_{i\ell}^\top X)} = \frac{\exp(h_s(X, a_{is}, b_{is}))}{\sum_{\ell=1}^{K} \exp(h_\ell(X, a_{i\ell}, b_{i\ell}))}.$$

Then, by means of Taylor expansion up to order r, we have

$$A_n = \sum_{i=1}^{2} \exp(\beta_{0i}^n) \sum_{|\alpha|=1}^{r} \frac{1}{\alpha!}(\beta_{1i}^n - \beta_{11}^*)^{\alpha_1} \prod_{\ell=1}^{K-1} (a_{i\ell}^n - a_{1\ell}^*)^{\alpha_{2\ell}}(b_{i\ell}^n - b_{1\ell}^*)^{\alpha_{3\ell}} \cdot X^{\alpha_1 + \sum_{\ell=1}^{K-1} \alpha_{3\ell}} \exp((\beta_{11}^*)^\top X)$$

$$\times \frac{\partial^{\sum_{\ell=1}^{K-1}(\alpha_{2\ell} + |\alpha_{3\ell}|)} f}{\partial h_1^{\alpha_{21} + |\alpha_{31}|} \dots \partial h_{K-1}^{\alpha_{2(K-1)} + |\alpha_{3(K-1)}|}}(Y = s|h_{11}^*, \dots, h_{1K}^*) + R_5(X, Y),$$

Here, $R_5(X, Y)$ is a Taylor remainder that satisfies $R_5(X, Y)/\mathcal{D}_r(G_n, G_*) \to 0$ as $n \to \infty$. Since $b_{1\ell}^* = \mathbf{0}_d$ for any $\ell \in [K-1]$, the derivatives of f with respect to $h_1, \dots, h_{K-1}$ in the above representation of A_n are constants depending on $(\alpha_{2\ell})_{\ell=1}^{K-1}$ and $(\alpha_{3\ell})_{\ell=1}^{K-1}$. Therefore, we can denote them as C_{α_2, α_3}.

Additionally, since $\beta_{1i}^n - \beta_{11}^* = \mathbf{0}_d$ and $b_{i\ell}^n - b_{1\ell}^* = \mathbf{0}_d$ for any $\ell \in [K-1]$ and $i \in \{1, 2\}$, we can let $\alpha_1 = \mathbf{0}_d$, $\alpha_{3\ell} = \mathbf{0}_d$, and rewrite A_n as follows:

$$A_n = \sum_{|\alpha_1|+|\alpha_2|=1}^{r} \sum_{i=1}^{2} \frac{\exp(\beta_{0i}^n)}{\alpha_1! \alpha_2!}(\beta_{1i}^n - \beta_{11}^*)^{\alpha_1} \prod_{\ell=1}^{K-1} (a_{i\ell}^n - a_{1\ell}^*)^{\alpha_{2\ell}} X^{\alpha_1} \exp((\beta_{11}^*)^\top X)C_{\alpha_2, \mathbf{0}} + R_5(X, Y)$$

$$= \sum_{|\alpha_1|+|\alpha_2|=1}^{r} \frac{\exp(\beta_{01}^*) - t_n}{\alpha_1! \alpha_2!} \cdot c_n^{|\alpha_1|+|\alpha_2|} \cdot X^{\alpha_1} \exp((\beta_{11}^*)^\top X)C_{\alpha_2, \mathbf{0}} + R_5(X, Y).$$

Similarly, by applying the Taylor expansion up to order r to B_n, we have

$$B_n = \sum_{|\gamma|=1}^{r} \sum_{i=1}^{2} \frac{1}{\gamma!} \exp(\beta_{0i}^n)(\beta_{1i}^n - \beta_{11}^*)^\gamma \cdot X^\gamma \exp((\beta_{11}^*)^\top X)g_{G_n}(Y = s|X) + R_6(X, Y)$$

$$= \sum_{|\gamma|=1}^{r} \frac{\exp(\beta_{01}^*) - t_n}{\gamma!} \cdot c_n^{|\gamma|} \cdot X^\gamma \exp((\beta_{11}^*)^\top X)g_{G_n}(Y = s|X) + R_6(X, Y),$$

where $R_6(X, Y)$ is a Taylor remainder such that $R_6(X, Y)/\mathcal{D}_r(G_n, G_*) \to 0$ as $n \to \infty$.

Now, we aim to demonstrate that $(A_n + E_{n,1})/\mathcal{D}_r(G_n, G_*) \to 0$ and $(B_n + E_{n,2})/\mathcal{D}_r(G_n, G_*) \to 0$ as $n \to \infty$, where

we define

$$E_{n,1} := \left(\sum_{i=1}^{2} \exp(\beta_{0i}^n) - \exp(\beta_{01}^*) \right) u(Y = s|X; \beta_{11}^*, a_1^*, b_1^*),$$

$$E_{n,2} := \left(\sum_{i=1}^{2} \exp(\beta_{0i}^n) - \exp(\beta_{01}^*) \right) v(Y = s|X; \beta_{11}^*),$$

in which $u(Y = s|X; \beta_{11}^*, a_1^*, b_1^*) = \exp((\beta_{11}^*)^\top X) f(Y = s|X; a_1^*, b_1^*) = \exp((\beta_{11}^*)^\top X) C_{\mathbf{0},\mathbf{0}}$, and $0 < C_{\mathbf{0},\mathbf{0}} < 1$.

Part 1. Prove that $(A_n + E_{n,1})/\mathcal{D}_r(G_n, G_*) \to 0$ as $n \to \infty$.

Since $\mathcal{X}$ is a bounded set, we assume that $\|X\| \leq B$ for some positive constant M. Then, we can verify that

$$0 \leq A_n + E_{n,1} - R_5(X, Y) \leq \sum_{|\alpha_1|+|\alpha_2|=0}^{r} L_{\alpha_1,\alpha_2}^n \exp((\beta_{11}^*)^\top X),$$

where we denote

$$L_{\alpha_1,\alpha_2}^n := \begin{cases} \dfrac{\exp(\beta_{01}^*) - t_n}{\alpha_1! \alpha_2!} \cdot c_n^{|\alpha_1|+|\alpha_2|} \cdot B^{|\alpha_1|} C_{\alpha_2,\mathbf{0}}, & |\alpha_1| + |\alpha_2| > 0, \\ -t_n C_{\mathbf{0},\mathbf{0}}, & |\alpha_1| + |\alpha_2| = 0. \end{cases}$$

Note that for α_1, α_2 such that $2 \leq |\alpha_1| + |\alpha_2| \leq r$, as $c_n = \mathcal{O}(t_n)$, we have $L_{\alpha_1,\alpha_2}^n/\mathcal{D}_r(G_n, G_*) \to 0$ as $n \to \infty$. Now, we show that $\sum_{|\alpha_1|+|\alpha_2|=0}^{1} L_{\alpha_1,\alpha_2}^n = 0$. Indeed, we have

$$\sum_{|\alpha_1|+|\alpha_2|=0}^{1} L_{\alpha_1,\alpha_2}^n = [\exp(\beta_{01}^*) - t_n] \cdot (C_{\mathbf{0},\mathbf{0}} B + N) c_n - C_{\mathbf{0},\mathbf{0}} t_n,$$

where $N := \sum_{|\alpha_2|=1} C_{\alpha_2,\mathbf{0}}$. By setting $t_n = \frac{B}{nN}$ and $c_n = \frac{1}{nN \exp(\beta_{01}^*) - B}$, the above sum reduces to zero. Thus, we get that

$$[A_n + E_{n,1} - R_5(X, Y)]/\mathcal{D}_r(G_n, G_*) \to 0,$$

as $n \to \infty$. Recall that $R_5(X, Y)/\mathcal{D}_r(G_n, G_*) \to 0$, we deduce that $(A_n + E_{n,1})/\mathcal{D}_r(G_n, G_*) \to 0$ as $n \to \infty$.

Part 2. Prove that $(B_n + E_{n,2})/\mathcal{D}_r(G_n, G_*) \to 0$ as $n \to \infty$.

It is worth noting that

$$0 \leq B_n + E_{n,2} - R_6(X, Y) \leq \sum_{|\gamma|=0}^{r} J_\gamma^n \exp((\beta_{11}^*)^\top X) g_{G_n}(Y = s|X),$$

where we denote

$$J_\gamma^n := \begin{cases} \dfrac{\exp(\beta_{01}^*) - t_n}{\gamma!} \cdot c_n^{|\gamma|} B^{|\gamma|}, & |\gamma| > 0, \\ -t_n, & |\gamma| = 0. \end{cases}$$

For $2 \leq |\gamma| \leq r$, we have that $J_\gamma^n/\mathcal{D}_r(G_n, G_*) \to 0$ as $n \to \infty$. Additionally, we have

$$\sum_{|\gamma|=0}^{1} J_\gamma^n = [\exp(\beta_{01}^*) - t_n] \cdot c_n \cdot B - t_n = 0.$$

As a result, we get that

$$[B_n + E_{n,2} - R_6(X, Y)]/\mathcal{D}_r(G_n, G_*) \to 0,$$

as $n \to \infty$. Since $R_6(X, Y)/\mathcal{D}_r(G_n, G_*) \to 0$, we obtain that $(B_n + E_{n,2})/\mathcal{D}_r(G_n, G_*) \to 0$ as $n \to \infty$.

Putting the results of Part 1 and Part 2 together, we obtain that

$$\frac{T_n(s)}{\mathcal{D}_r(G_n, G_*)} = \frac{A_n - B_n + E_n}{\mathcal{D}_r(G_n, G_*)} \to 0,$$

which indicates that $\mathbb{E}_X[V(g_{G_n}(\cdot|X), g_{G_*}(\cdot|X))]/\mathcal{D}_r(G_n, G_*) \to 0$ as $n \to \infty$. Furthermore, it follows from equation (28) that $\mathcal{D}_r(G_n, G_*) \to 0$ as $n \to \infty$. Hence, we reach the conclusion of Proposition 3.2.

A.3. Proof of Theorem 4.4

To reach the desired conclusion in Theorem 4.4, we need to show the following key inequality:

$$\inf_{G \in \mathcal{O}_k(\Theta)} \mathbb{E}_X[V(\widetilde{g}_G(\cdot|X), \widetilde{g}_{G_*}(\cdot|X))]/\mathcal{D}_2(G, G_*) > 0, \tag{29}$$

which is then divided into two parts named local structure and global structure. Since the global structure can be argued similarly to that in Appendix A.1 with a note that the modified softmax gating multinomial logistic MoE model is identifiable (see Proposition 4.2), the proof for it is omitted in this appendix.

Local Structure: For this part, we use the proof by contradiction method to show that

$$\lim_{\varepsilon \to 0} \inf_{\substack{G \in \mathcal{O}_k(\Theta), \\ \mathcal{D}_2(G, G_*) \leq \varepsilon}} \mathbb{E}_X[V(\widetilde{g}_G(\cdot|X), \widetilde{g}_{G_*}(\cdot|X))]/\mathcal{D}_2(G, G_*) > 0. \tag{30}$$

Assume that this local inequality does not hold, then by utilizing some derivations in Appendix A.1, we proceed the three-step framework as follows:

Step 1. First of all, by deriving in the same fashion as in equation (14), we get the following decomposition of $\widetilde{T}_n(s) := \left[\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)\right] \cdot \left[\widetilde{g}_{G_n}(Y = s|X) - \widetilde{g}_{G_*}(Y = s|X)\right]$:

$$\widetilde{T}_n(s) = \sum_{j=1}^{k_*} \sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n) \left[\widetilde{u}(Y = s|X; \beta_{1i}^n, a_i^n, b_i^n) - \widetilde{u}(Y = s|X; \beta_{1j}^*, a_j^*, b_j^*)\right]$$

$$- \sum_{j=1}^{k_*} \sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n) \left[\widetilde{v}(Y = s|X; \beta_{1i}^n) - \widetilde{v}(Y = s|X; \beta_{1j}^*)\right]$$

$$+ \sum_{j=1}^{k_*} \left(\sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^*)\right) \left[\widetilde{u}(Y = s|X; \beta_{1j}^*, a_j^*, b_j^*) - \widetilde{v}(Y = s|X; \beta_{1j}^*)\right]$$

$$:= A_n - B_n + E_n,$$

where we define

$$\widetilde{u}(Y = s|X; \beta_{1i}, a_i, b_i) := \exp(\beta_{1i}^\top M(X)) \cdot f(Y = s|X; a_i, b_i),$$

$$\widetilde{v}(Y = s|X; \beta_{1i}) := \exp(\beta_{1i}^\top M(X)) \cdot g_{G_n}(Y = s|X).$$

for any $s \in [K]$. Next, we will apply first order and second order Taylor expansions to two terms in the following sum:

$$A_n = \sum_{j:|\mathcal{C}_j|=1} \sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n) \left[\widetilde{u}(Y = s|X; \beta_{1i}^n, a_i^n, b_i^n) - \widetilde{u}(Y = s|X; \beta_{1j}^*, a_j^*, b_j^*)\right]$$

$$+ \sum_{j:|\mathcal{C}_j|>1} \sum_{i \in \mathcal{C}_j} \exp(\beta_{0i}^n) \left[\widetilde{u}(Y = s|X; \beta_{1i}^n, a_i^n, b_i^n) - \widetilde{u}(Y = s|X; \beta_{1j}^*, a_j^*, b_j^*)\right]$$

$$:= A_{n,1} + A_{n,2}.$$

For the first term, we get that

$$
A_{n,1} = \sum_{j:|\mathcal{C}_j|>1} \sum_{i\in\mathcal{C}_j} \exp(\beta_{0i}^n) \sum_{|\alpha|=1} \frac{1}{\alpha!} (\Delta\beta_{1ij}^n)^{\alpha_1} \prod_{\ell=1}^{K-1} (\Delta a_{ij\ell}^n)^{\alpha_{2\ell}} (\Delta b_{ij\ell}^n)^{\alpha_{3\ell}} \cdot [M(X)]^{\alpha_1} X^{\sum_{\ell=1}^{K-1}\alpha_{3\ell}}
$$
$$
\times \exp((\beta_{1j}^*)^\top M(X)) \cdot \frac{\partial^{\sum_{\ell=1}^{K-1}(\alpha_{2\ell}+|\alpha_{3\ell}|)} f}{\partial h_1^{\alpha_{21}+|\alpha_{31}|}\ldots\partial h_{K-1}^{\alpha_{2(K-1)}+|\alpha_{3(K-1)}|}} (Y=s|X;a_j^*,b_j^*) + \widetilde{R}_1(X,Y),
$$

where $\widetilde{R}_1(X,Y)$ is a Taylor remainder such that $\widetilde{R}_1(X,Y)/\mathcal{D}_2(G_n,G_*) \to 0$ as $n\to\infty$. Let $q_2 = (\alpha_{2\ell}+|\alpha_{3\ell}|)_{\ell=1}^{K-1} \in \mathbb{N}^{K-1}$, $q_3 = \sum_{\ell=1}^{K-1}\alpha_{3\ell} \in \mathbb{N}^d$ and $q_4 = \alpha_1 \in \mathbb{N}^d$, we rewrite $A_{n,1}$ as

$$
A_{n,1} = \sum_{j:|\mathcal{C}_j|=1} \sum_{|q_4|+|q_2|=1} \sum_{|q_3|=0}^{|q_2|} \sum_{i\in\mathcal{C}_j} \sum_{\alpha\in\mathcal{I}_{q_2,q_3,q_4}} \frac{\exp(\beta_{0i}^n)}{\alpha!} (\Delta\beta_{1ij}^n)^{\alpha_1} \prod_{\ell=1}^{K-1} (\Delta a_{ij\ell}^n)^{\alpha_{2\ell}} (\Delta b_{ij\ell}^n)^{\alpha_{3\ell}}
$$
$$
\times [M(X)]^{q_4} X^{q_3} \exp((\beta_{1j}^*)^\top M(X)) \cdot \frac{\partial^{|q_2|} f}{\partial h_1^{q_{21}}\ldots\partial h_{K-1}^{q_{2(K-1)}}} (Y=s|X;a_j^*,b_j^*) + \widetilde{R}_1(X,Y),
$$

where we define

$$
\mathcal{I}_{q_2,q_3,q_4} := \Big\{ \alpha = (\alpha_1,\alpha_{21},\ldots,\alpha_{2(K-1)},\alpha_{31},\ldots,\alpha_{3(K-1)}) : \alpha_1 = q_4, \sum_{\ell=1}^{K-1}\alpha_{3\ell} = q_3, (\alpha_{2\ell}+|\alpha_{3\ell}|)_{\ell=1}^{K-1} = q_2 \Big\}.
$$

For the second term $A_{n,2}$, we have

$$
A_{n,2} = \sum_{j:|\mathcal{C}_j|>1} \sum_{|q_4|+|q_2|=1}^{2} \sum_{|q_3|=0}^{|q_2|} \sum_{i\in\mathcal{C}_j} \sum_{\alpha\in\mathcal{I}_{q_2,q_3,q_4}} \frac{\exp(\beta_{0i}^n)}{\alpha!} (\Delta\beta_{1ij}^n)^{\alpha_1} \prod_{\ell=1}^{K-1} (\Delta a_{ij\ell}^n)^{\alpha_{2\ell}} (\Delta b_{ij\ell}^n)^{\alpha_{3\ell}}
$$
$$
\times [M(X)]^{q_4} X^{q_3} \exp((\beta_{1j}^*)^\top M(X)) \cdot \frac{\partial^{|q_2|} f}{\partial h_1^{q_{21}}\ldots\partial h_{K-1}^{q_{2(K-1)}}} (Y=s|X;a_j^*,b_j^*) + \widetilde{R}_2(X,Y),
$$

where $\widetilde{R}_2(X,Y)$ is a Taylor remainder such that $\widetilde{R}_2(X,Y)/\mathcal{D}_2(G_n,G_*) \to 0$ when $n\to\infty$.

Meanwhile, by arguing similarly, we can decompose B_n as

$$
B_n = \sum_{j:|\mathcal{C}_j|=1} \sum_{|\gamma|=1} \sum_{i\in\mathcal{C}_j} \frac{\exp(\beta_{0i}^n)}{\gamma!} (\Delta\beta_{1ij}^n)^\gamma \times [M(X)]^\gamma \exp((\beta_{1j}^*)^\top M(X)) g_{G_n}(Y=s|X) + \widetilde{R}_3(X,Y)
$$
$$
+ \sum_{j:|\mathcal{C}_j|>1} \sum_{|\gamma|=1}^{2} \sum_{i\in\mathcal{C}_j} \frac{\exp(\beta_{0i}^n)}{\gamma!} (\Delta\beta_{1ij}^n)^\gamma \times [M(X)]^\gamma \exp((\beta_{1j}^*)^\top M(X)) g_{G_n}(Y=s|X) + \widetilde{R}_4(X,Y),
$$

where $\widetilde{R}_3(X,Y)$ and $\widetilde{R}_4(X,Y)$ are Taylor remainders such that their ratios to $\mathcal{D}_2(G_n,G_*)$ vanish as n approaches infinity.

Combine the above results, we deduce that $\widetilde{T}_n(s)$ can be represented as follows:

$$
\widetilde{T}_n(s) = \sum_{j=1}^{k_*} \sum_{|q_4|+|q_2|=0}^{1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}}} \sum_{|q_3|=0}^{|q_2|} Z_{q_2,q_3,q_4}^n(j) \times [M(X)]^{q_4} X^{q_3} \exp((\beta_{1j}^*)^\top M(X)) \frac{\partial^{|q_2|} f}{\partial h_1^{q_{21}}\ldots\partial h_{K-1}^{q_{2(K-1)}}} (Y=s|X;a_j^*,b_j^*)
$$
$$
+ \sum_{j=1}^{k_*} \sum_{|\gamma|=0}^{1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}}} W_\gamma^n(j) \times [M(X)]^\gamma \exp((\beta_{1j}^*)^\top M(X)) g_{G_n}(Y=s|X) + \widetilde{R}(X,Y), \tag{31}
$$

where $\widetilde{R}(X,Y)$ is the sum of Taylor remainders such that $\widetilde{R}(X,Y)/\mathcal{D}_2(G_n,G_*) \to 0$ as $n\to\infty$ and

$$
Z_{q_2,q_3,q_4}^n(j) = \begin{cases} \sum_{i\in\mathcal{C}_j} \sum_{\alpha\in\mathcal{I}_{q_2,q_3,q_4}} \frac{\exp(\beta_{0i}^n)}{\alpha!} (\Delta\beta_{1ij}^n)^{\alpha_1} \prod_{\ell=1}^{K-1} (\Delta a_{ij\ell}^n)^{\alpha_{2\ell}} (\Delta b_{ij\ell}^n)^{\alpha_{3\ell}}, & (q_2,q_3,q_4)\neq(\mathbf{0}_{K-1},\mathbf{0}_d,\mathbf{0}_d), \\ \sum_{i\in\mathcal{C}_j} \exp(\beta_{0i}^n) - \exp(\beta_{0j}^*), & (q_2,q_3,q_4)=(\mathbf{0}_{K-1},\mathbf{0}_d,\mathbf{0}_d), \end{cases}
$$

and

$$W_\gamma^n(j) = \begin{cases} -\sum_{i\in\mathcal{C}_j} \frac{\exp(\beta_{0i}^n)}{\gamma!}(\Delta\beta_{1ij}^n)^\gamma, & |\gamma| \neq \mathbf{0}_d, \\ -\sum_{i\in\mathcal{C}_j}\exp(\beta_{0i}^n) + \exp(\beta_{0j}^*), & |\gamma| = \mathbf{0}_d, \end{cases}$$

for any $j \in [k_*]$.

Step 2. Subsequently, we will show that at least one among $Z_{q_2,q_3,q_4}^n(j)/\mathcal{D}_2(G_n,G_*)$ does not approach zero as n tends to infinity. Assume by contrary that all of them vanish as $n \to \infty$, then we consider some typical tuples (q_2,q_3,q_4). Firstly, by taking the sum of $Z_{q_2,q_3,q_4}^n(j)$ for $j \in [k_*] : |\mathcal{C}_j| > 1$ (resp. $j \in [k_*] : |\mathcal{C}_j| = 1$) and $(q_2,q_3,q_4) \in \{(\mathbf{0}_{K-1}, \mathbf{0}_d, 2e_i) : i \in [d]\}$ (resp. $(q_2,q_3,q_4) \in \{(\mathbf{0}_{K-1},\mathbf{0}_d,e_i) : i \in [d]\}$) where $e_i := (0,\ldots,0,\underbrace{1}_{i\text{-}th},0,\ldots,0) \in \mathbb{R}^d$, we get

$$\frac{1}{\mathcal{D}_2(G_n,G_*)} \cdot \sum_{j:|\mathcal{C}_j|>1}\sum_{i\in\mathcal{C}_j}\exp(\beta_{0i}^n)\|\Delta\beta_{1ij}^n\|^2 \to 0,$$
$$\frac{1}{\mathcal{D}_2(G_n,G_*)} \cdot \sum_{j:|\mathcal{C}_j|=1}\sum_{i\in\mathcal{C}_j}\exp(\beta_{0i}^n)\|\Delta\beta_{1ij}^n\| \to 0. \qquad (32)$$

For $(q_2,q_3,q_4) \in \{(\mathbf{0}_{K-1},2e_i,\mathbf{0}_d) : i \in [d]\}$ (resp. $(q_2,q_3,q_4) \in \{(\mathbf{0}_{K-1},e_i,\mathbf{0}_d) : i \in [d]\}$) where , we obtain that

$$\frac{1}{\mathcal{D}_2(G_n,G_*)} \cdot \sum_{j:|\mathcal{C}_j|>1}\sum_{i\in\mathcal{C}_j}\exp(\beta_{0i}^n)\sum_{\ell=1}^{K-1}\|\Delta b_{ij\ell}^n\|^2 \to 0,$$
$$\frac{1}{\mathcal{D}_2(G_n,G_*)} \cdot \sum_{j:|\mathcal{C}_j|=1}\sum_{i\in\mathcal{C}_j}\exp(\beta_{0i}^n)\sum_{\ell=1}^{K-1}\|\Delta b_{ij\ell}^n\| \to 0. \qquad (33)$$

On the other hand, for $(q_2,q_3,q_4) \in \{(e_\ell',\mathbf{0}_d,\mathbf{0}_d) : \ell \in [K-1]\}$ (resp. $(q_2,q_3,q_4) \in \{(e_\ell',\mathbf{0}_d,\mathbf{0}_d) : \ell \in [K-1]\}$) where $e_\ell' := (0,\ldots,0,\underbrace{1}_{\ell\text{-}th},0,\ldots,0) \in \mathbb{R}^{K-1}$, we have

$$\frac{1}{\mathcal{D}_2(G_n,G_*)} \cdot \sum_{j:|\mathcal{C}_j|>1}\sum_{i\in\mathcal{C}_j}\exp(\beta_{0i}^n)\sum_{\ell=1}^{K-1}\|\Delta a_{ij\ell}^n\|^2 \to 0, \qquad (34)$$
$$\frac{1}{\mathcal{D}_2(G_n,G_*)} \cdot \sum_{j:|\mathcal{C}_j|=1}\sum_{i\in\mathcal{C}_j}\exp(\beta_{0i}^n)\sum_{\ell=1}^{K-1}\|\Delta a_{ij\ell}^n\| \to 0. \qquad (35)$$

Additionally, when $(q_2,q_3,q_4) = (\mathbf{0}_{K-1},\mathbf{0}_d,\mathbf{0}_d)$, it follows that

$$\sum_{j=1}^{k_*}\frac{|Z_{\mathbf{0}_{K-1},\mathbf{0}_d,\mathbf{0}_d}^n(j)|}{\mathcal{D}_2(G_n,G_*)} = \frac{1}{\mathcal{D}_2(G_n,G_*)} \cdot \sum_{j=1}^{k_*}\left|\sum_{i\in\mathcal{C}_j}\exp(\beta_{0i}^n) - \exp(\beta_{0j}^*)\right| \to 0. \qquad (36)$$

It is induced from the limits in equations (32), (33), (34) and (36) that $1 = \mathcal{D}_2(G_n,G_*)/\mathcal{D}_2(G_n,G_*) \to 0$ when $n \to \infty$, which is a contradiction. Thus, at least one among $Z_{q_2,q_3,q_4}^n(j)/\mathcal{D}_2(G_n,G_*)$ does not go to zero as $n \to \infty$.

Step 3. Now, we denote $\widetilde{m}_n$ as the maximum of the absolute values of $Z_{q_2,q_3,q_4}^n(j)/\mathcal{D}_2(G_n,G_*)$ and $W_\gamma^n(j)/\mathcal{D}_2(G_n,G_*)$ for any $j \in [k_*], 0 \leq |q_2| + |q_4| \leq 1 + \mathbf{1}_{\{|\mathcal{C}_j|>1\}}, 0 \leq |q_3| \leq |q_2|$ and $0 \leq |\gamma| \leq 1 + \mathbf{1}_{\{|\mathcal{C}_j|>1\}}$. Since at least one among $Z_{q_2,q_3,q_4}^n(j)/\mathcal{D}_2(G_n,G_*)$ does not go to zero as $n \to \infty$, we deduce that $\widetilde{m}_n \not\to 0$, and therefore, $1/\widetilde{m}_n \not\to \infty$. Then, we denote

$$Z_{q_2,q_3,q_4}^n(j)/[m_n\mathcal{D}_2(G_n,G_*)] \to \widetilde{\tau}_{q_2,q_3,q_4}(j)$$
$$W_\gamma^n(j)/\mathcal{D}_2(G_n,G_*) \to \widetilde{\eta}_\gamma(j)$$

as $n \to \infty$. Here, at least one among $\widetilde{\tau}_{q_2,q_3,q_4}(j)$ for $j \in [k_*]$, $0 \leq |q_2| + |q_4| \leq 1 + \mathbf{1}_{\{|\mathcal{C}_j|>1\}}$ and $0 \leq |q_3| \leq |q_2|$ is non-zero. By invoking the Fatou's lemma, we have that

$$0 = \lim_{n \to \infty} \frac{\mathbb{E}_X[2V(\widetilde{g}_{G_n}(\cdot|X), \widetilde{g}_{G_*}(\cdot|X))]}{m_n \mathcal{D}_2(G_n, G_*)} \geq \int \sum_{s=1}^{K} \liminf_{n \to \infty} \frac{|\widetilde{g}_{G_n}(Y = s|X) - \widetilde{g}_{G_*}(Y = s|X)|}{m_n \mathcal{D}_2(G_n, G_*)} \, dX \geq 0,$$

which indicates that $[\widetilde{g}_{G_n}(Y = s|X) - \widetilde{g}_{G_*}(Y = s|X)]/[\widetilde{m}_n \mathcal{D}_2(G_n, G_*)]$ tends to zero as n goes to infinity for any $s \in [K]$ and almost surely X. This result is equivalent to

$$\widetilde{T}_n(s)/[\widetilde{m}_n \mathcal{D}_2(G_n, G_*)] \to 0, \tag{37}$$

as $n \to \infty$, for any $s \in [K]$. Putting the results in equations (31) and (37) together, we have

$$\sum_{j=1}^{k_*} \sum_{|q_4|+|q_2|=0}^{1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}}} \sum_{|q_3|=0}^{|q_2|} \widetilde{\tau}_{q_2,q_3,q_4}(j) \times [M(X)]^{q_4} X^{q_3} \exp((\beta_{1j}^*)^\top M(X)) \frac{\partial^{|q_2|} f}{\partial h_1^{q_{21}} \dots \partial h_{K-1}^{q_{2(K-1)}}} (Y = s|X; a_j^*, b_j^*)$$

$$+ \sum_{j=1}^{k_*} \sum_{|\gamma|=0}^{1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}}} \widetilde{\eta}_\gamma(j) \times [M(X)]^\gamma \exp((\beta_{1j}^*)^\top M(X)) g_{G_*}(Y = s|X) = 0. \tag{38}$$

Regime 1: For any $j \in [k_*]$, there exists an index $\ell \in [K - 1]$ such that $b_{j\ell}^* \neq \mathbf{0}_d$.

By using the same arguments for proving the set $\mathcal{F}$ in equation (26) is linearly independent for almost surely X, we get that the following set also admits that property:

$$\left\{ [M(X)]^{q_4} X^{q_3} \exp((\beta_{1j}^*)^\top M(X)) \frac{\partial^{|q_2|} f}{\partial h_1^{q_{21}} \dots \partial h_{K-1}^{q_{2(K-1)}}} (Y = s|X; a_j^*, b_j^*), \right.$$

$$\left. [M(X)]^\gamma \exp((\beta_{1j}^*)^\top M(X)) g_{G_*}(Y = s|X) : j \in [k_*], \ 0 \leq |q_2| + |q_4|, |\gamma| \leq 1 + \mathbf{1}_{\{|\mathcal{C}_j|>1\}}, \ 0 \leq |q_3| \leq |q_2| \right\}.$$

As a result, it follows that $\widetilde{\tau}_{q_2,q_3,q_4}(j) = \widetilde{\eta}_\gamma(j) = 0$ for any $j \in [k_*]$, $0 \leq |q_2| + |q_4|, |\gamma| \leq 1 + \mathbf{1}_{\{|\mathcal{C}_j|>1\}}$ and $0 \leq |q_3| \leq |q_2|$, which contradicts the fact that at least one among $\widetilde{\tau}_{q_2,q_3,q_4}(j)$ is different from zero.

Regime 2: There exists an index $j \in [k_*]$ such that $b_{j\ell}^* = \mathbf{0}_d$ for any $\ell \in [K - 1]$.

Note that equation (38) can be rewritten as

$$\sum_{j=1}^{k_*} P^{(j)}(X) \exp((\beta_{1j}^*)^\top M(X)) + \sum_{j=1}^{k_*} Q^{(j)}(X) \exp((\beta_{1j}^*)^\top M(X)) g_{G_*}(Y = s|X) = 0,$$

where we define

$$P^{(j)}(X) := \sum_{|q_4|+|q_2|=0}^{1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}}} \sum_{|q_3|=0}^{|q_2|} \widetilde{\tau}_{q_2,q_3,q_4}(j) X^{q_3} [M(X)]^{q_4} \cdot \frac{\partial^{|q_2|} f}{\partial h_1^{q_{21}} \dots \partial h_{K-1}^{q_{2(K-1)}}} (Y = s|X; a_j^*, b_j^*),$$

$$Q^{(j)}(X) := \sum_{|\gamma|=0}^{1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}}} \widetilde{\eta}_\gamma(j) [M(X)]^\gamma.$$

Since the following set is linearly independent for almost surely X:

$$\left\{ \exp((\beta_{1j}^*)^\top M(X)), \ \exp((\beta_{1j}^*)^\top M(X)) g_{G_*}(Y = s|X) : j \in [k_*] \right\},$$

we achieve that $P^{(j)}(X) = Q^{(j)}(X) = 0$ for any $j \in [k_*]$ for almost surely X. Then, it follows from the formulation of $P^{(j)}(X)$ that

$$\sum_{|q_4|+|q_2|=0}^{1+\mathbf{1}_{\{|\mathcal{C}_j|>1\}}} \sum_{|q_3|=0}^{|q_2|} \widetilde{\tau}_{q_2,q_3,q_4}(j) X^{q_3} [M(X)]^{q_4} = 0,$$

24

for any $j \in [k_*]$ for almost surely X. It can be seen from the above equation that $0 \leq |q_3| + |q_4| \leq |q_2| + |q_4| \leq 1 + \mathbf{1}_{\{|\mathcal{C}_j|>1\}} \leq 2$. Moreover, by definition of function M (see Definition 4.1), the set $\left\{ X^p[M(X)]^q : p, q \in \mathbb{N}^d, \, 0 \leq |p| + |q| \leq 2 \right\}$ is linearly independent for almost surely X. As a consequence, we achieve that $\widetilde{\tau}_{q_2,q_3,q_4}(j) = 0$ for any $j \in [k_*]$, $0 \leq |q_2| + |q_4|, |\gamma| \leq 1 + \mathbf{1}_{\{|\mathcal{C}_j|>1\}}$ and $0 \leq |q_3| \leq |q_2|$, contradicting the fact that at least one among $\widetilde{\tau}_{q_2,q_3,q_4}(j)$ is non-zero.

Combine the results of the above two regimes, we reach the bound in equation (30).

B. Proofs for Convergence Rates of Density Estimation

In this appendix, we present the proof of Propostion 2.2 in Appendix B.1, while that for Proposition 4.3 is given in Appendix B.2.

B.1. Proof of Proposition 2.2

In this appendix, we will firstly introduce key results on density estimation with MLE which are mainly based on (van de Geer, 2000), and then provide the proof of Proposition 2.2 at the end.

B.1.1. KEY RESULTS

To begin with, it is necessary to define some notations that will be used in our presentation. First, we define $\mathcal{P}_k(\Theta) := \{g_G(Y|X) : G \in \mathcal{O}_k(\Theta)\}$ as the set of conditional density functions of all mixing measures belonging to $\mathcal{O}_k(\Theta)$. In addition, let $N(\varepsilon, \mathcal{P}_k(\Theta), \|\cdot\|_1)$ be the covering number (van de Geer, 2000) of metric space $(\mathcal{P}_k(\Theta), \|\cdot\|_1)$ while $H_B(\varepsilon, \mathcal{P}_k(\Theta), h)$ be the bracketing entropy (van de Geer, 2000) of $\mathcal{P}_k(\Theta)$ under the Hellinger distance. Then, the following result gives us the upper bound of these quantities:

Lemma B.1. *For any bounded set Θ and $\varepsilon \in (0, 1/2)$, we have*

(i) $\log N(\varepsilon, \mathcal{P}_k(\Theta), \|\cdot\|_1) \lesssim \log(1/\varepsilon);$

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

Proof of Lemma B.1. **Part (i)** Firstly, we define $\Omega := \{(a, b) \in \mathbb{R}^K \times \mathbb{R}^{d \times K} : (\beta_0, \beta_1, a, b) \in \Theta\}$. Since Θ is a compact set, then Ω is also compact. Therefore, Ω admits an ε-cover of size T_2 denoted by Ω_ε. In addition, we also define $\Delta := \{(\beta_0, \beta_1) \in \mathbb{R} \times \mathbb{R}^d : (\beta_0, \beta_1, a, b) \in \Theta\}$, and Δ_ε as an ε-cover of Δ. It can be validated that

$$|\Omega_\varepsilon| \lesssim \mathcal{O}(\varepsilon^{-K(d+1)k}), \quad |\Delta_\varepsilon| \lesssim \mathcal{O}(\varepsilon^{-(d+1)k}).$$

Next, given a mixing measure $G = \sum_{i=1}^{k'} \exp(\beta_{0i})\delta_{(\beta_{1i}, a_i, b_i)} \in \mathcal{O}_k(\Theta)$, where $k' \in [k]$, we define $\widetilde{G} := \sum_{i=1}^{k'} \exp(\beta_{0i})\delta_{(\beta_{1i}, \bar{a}_i, \bar{b}_i)}$ in which $(\bar{a}_i, \bar{b}_i) \in \Omega_\varepsilon$ such that it is the closet point to (a_i, b_i) for any $i \in [k']$. Additionally, we also consider the mixing measure $\overline{G} := \sum_{i=1}^{k'} \exp(\overline{\beta}_{0i})\delta_{(\overline{\beta}_{1i}, \bar{a}_i, \bar{b}_i)}$, where $(\overline{\beta}_{0i}, \overline{\beta}_{1i}) \in \Delta_\varepsilon$ is the closest point to (β_{0i}, β_{1i}). By this construction, it can be justified that $g_{\overline{G}} \in \mathcal{H}$, where we define

$$\mathcal{H} := \left\{ g_G \in \mathcal{P}_k(\Theta) : (\overline{\beta}_{0i}, \overline{\beta}_{1i}) \in \Delta_\varepsilon, \, (\bar{a}_i, \bar{b}_i) \in \Omega_\varepsilon, \, \forall i \in [k] \right\}.$$

Now, we show that $\mathcal{H}$ is an ε-cover of the metric space $(\mathcal{P}_k(\Omega), \|\cdot\|_1)$ but not necessarily the smallest one. For that purpose, we aim to find a bound for the term $\|g_G - g_{\overline{G}}\|_1$. According to the triangle inequality, we have

$$\|g_G - g_{\overline{G}}\|_1 \leq \|g_G - g_{\widetilde{G}}\|_1 + \|g_{\widetilde{G}} - g_{\overline{G}}\|_1.$$

Regarding the first term in the above right hand side,

$$
\begin{aligned}
\|g_{\widetilde{G}} - g_{\overline{G}}\|_1 &= \sum_{s=1}^{K} \int_{\mathcal{X}} \Big| g_G(Y|X) - g_{\widetilde{G}}(Y|X) \Big| \mathrm{d}X \\
&= \sum_{s=1}^{K} \int_{\mathcal{X}} \sum_{i=1}^{k'} \mathrm{Softmax}(\beta_{1i}^\top X + \beta_{0i}) \cdot \Big| f(Y=s|X; a_i, b_i) - f(Y=s|X; \overline{a}_i, \overline{b}_i) \Big| \mathrm{d}X \\
&\leq \sum_{s=1}^{K} \int_{\mathcal{X}} \sum_{i=1}^{k'} \Big| \mathrm{Softmax}(a_{is} + b_{is}^\top X) - \mathrm{Softmax}(\overline{a}_{is} + \overline{b}_{is}^\top X) \Big| \mathrm{d}X.
\end{aligned}
$$

Since $\mathrm{Softmax}$ is a differentiable function, it is also a L-Lipschitz function where $L > 0$. Additionally, as $\mathcal{X}$ is a bounded function, we may assume that $\|X\| \leq B$ for some constant $B > 0$. As a result, we have

$$
\begin{aligned}
\|g_{\widetilde{G}} - g_{\overline{G}}\|_1 &\leq \sum_{s=1}^{K} \sum_{i=1}^{k'} L \cdot \Big(|a_{is} - \overline{a}_{is}| + \|X\| \cdot \|b_{is} - \overline{b}_{is}\| \Big) \\
&\leq Kk'L \cdot (\varepsilon + B\varepsilon) \lesssim \varepsilon.
\end{aligned}
$$

Similarly, the second term is bounded as

$$
\begin{aligned}
\|g_{\widetilde{G}} - g_{\overline{G}}\|_1 &= \sum_{s=1}^{K} \int_{\mathcal{X}} \sum_{i=1}^{k'} \Big| \mathrm{Softmax}(\beta_{1i}^\top X + \beta_{0i}) - \mathrm{Softmax}(\overline{\beta}_{1i}^\top X + \overline{\beta}_{0i}) \Big| \cdot f(Y|X; \overline{a}_{is}, \overline{b}_{is}) \mathrm{d}X \\
&\leq \sum_{s=1}^{K} \sum_{i=1}^{k'} L \cdot \Big(\|\beta_{1i} - \overline{\beta}_{1i}\| \cdot \|X\| + |\beta_{0i} - \overline{\beta}_{0i}| \Big) \\
&\leq Kk'L(B\varepsilon + \varepsilon) \lesssim \varepsilon.
\end{aligned}
$$

Consequently, we get $\|g_G - g_{\overline{G}}\|_1 \lesssim \varepsilon$. This result implies that $\mathcal{H}$ is an ε-cover of the metric space $(\mathcal{P}_k(\Theta), \|\cdot\|_1)$. Then, it follows from the definition of the covering number that

$$
N(\varepsilon, \mathcal{P}_k(\Theta), \|\cdot\|_1) \leq |\mathcal{H}| = |\Theta_\varepsilon| \times |\Delta_\varepsilon| = \mathcal{O}(\varepsilon^{-K(d+1)k}) \times \mathcal{O}(\varepsilon^{-(d+1)k}) = \mathcal{O}(\varepsilon^{-(K+1)(d+1)k}),
$$

which is equivalent to $\log N(\varepsilon, \mathcal{P}_k(\Theta), \|\cdot\|_1) \leq \log(1/\varepsilon)$.

Part (ii) Given $\varepsilon > 0$ and let $\eta \leq \varepsilon$ that we will choose later. Assume that $\mathcal{P}_k(\Theta)$ has an η-cover denoted by $\{p_1, p_2, \ldots, p_N\}$ where $N := N(\eta, \mathcal{P}_k(\Theta), \|\cdot\|_1)$. Next, we start to construct brackets of the form $[L_i(Y|X), U_i(Y|X)]$ for all $i \in [N]$ as below:

$$
\begin{aligned}
L_i(Y|X) &:= \max\{p_i(Y|X) - \eta, 0\}, \\
U_i(Y|X) &:= \max\{p_i(Y|X) + \eta, 1\}.
\end{aligned}
$$

By this construction, we can verify that $\mathcal{P}_k(\Theta) \subset \bigcup_{i=1}^{N} [L_i(Y|X), U_i(Y|X)]$ and $U_i(Y|X) - L_i(Y|X) \leq \min\{2\eta, 1\}$. Additionally, we also have

$$
\|U_i(\cdot|X) - L_i(\cdot|X)\|_1 = \sum_{\ell=1}^{K} [U_i(Y=\ell|X) - L_i(Y=\ell|X)] \leq 2K\eta.
$$

By definition, since $H_B(2K\eta, \mathcal{P}_k(\Theta), \|\cdot\|_1)$ is the logarithm of the smallest number of brackets of size $2K\eta$ required for covering $\mathcal{P}_k(\Theta)$, we obtain that

$$
H_B(2K\eta, \mathcal{P}_k(\Theta), \|\cdot\|_1) \leq \log N(\eta, \mathcal{P}_k(\Theta), \|\cdot\|_1) \leq \log(1/\eta),
$$

where the last inequality follows from the result of Part (i). Thus, by choosing $\eta = \varepsilon/(2K)$, we have $H_B(\varepsilon, \mathcal{P}_k(\Theta), \|\cdot\|_1) \lesssim \log(1/\varepsilon)$. Furthermore, as $h \leq \|\cdot\|_1$, we achieve the desired conclusion:

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

Hence, the proof is completed. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Subsequently, we denote

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

For any $\xi > 0$, the Hellinger ball centered around the true conditional density $g_{G_*}(Y|X)$ and intersected with the set $\overline{\mathcal{P}}_k^{1/2}(\Theta)$ is defined as

$$\overline{\mathcal{P}}_k^{1/2}(\Theta, \xi) := \{g^{1/2} \in \overline{\mathcal{P}}_k(\Theta) : \mathbb{E}_X[h(g(\cdot|X), g_{G_*}(\cdot|X))] \leq \xi\}.$$

Moreover, Geer et al. (van de Geer, 2000) proposes the following term to capture the size of the Hellinger ball $\overline{\mathcal{P}}_k^{1/2}(\Theta, \xi)$:

$$\mathcal{J}_B(\xi, \overline{\mathcal{P}}_k^{1/2}(\Theta, \xi)) := \int_{\xi^2/2^{13}}^{\xi} H_B^{1/2}(t, \overline{\mathcal{P}}_k^{1/2}(\Theta, t), \|\cdot\|) \mathrm{d}t \vee \xi, \tag{39}$$

where $t \vee \xi := \max\{t, \xi\}$. Now, let us recall below an important result regarding the density estimation rate in (van de Geer, 2000) with adapted notations of this paper.

Lemma B.2 (Theorem 7.4, (van de Geer, 2000)). *Take $\Psi(\xi) \geq \mathcal{J}_B(\xi, \overline{\mathcal{P}}_k^{1/2}(\Theta, \xi))$ such that $\Psi(\xi)/\xi^2$ is a non-increasing function of ξ. Then, given a universal constant c and a sequence (ξ_n) that satisfies $\sqrt{n}\xi_n^2 \geq c\Psi(\xi_n)$, we get*

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

for any $\xi \geq \xi_n$.

The proof of this Lemma is in (van de Geer, 2000).

B.1.2. MAIN PROOF

It is worth noting that $H_B(t, \overline{\mathcal{P}}_k^{1/2}(\Theta, t), \|\cdot\|) \leq H_B(t, \mathcal{P}_k(\Theta), h)$ for any $t > 0$. Then, equation (39) indicates that

$$\mathcal{J}_B(\xi, \overline{\mathcal{P}}_k^{1/2}(\Theta, \xi)) \leq \int_{\xi^2/2^{13}}^{\xi} H_B^{1/2}(t, \overline{\mathcal{P}}_k^{1/2}(\Theta, t), h) \mathrm{d}t \vee \xi \lesssim \int_{\xi^2/2^{13}}^{\xi} \log(1/t) \mathrm{d}t \vee \xi,$$

where the second inequality follows from part (ii) of Lemma B.1. By setting $\Psi(\xi) = \xi\sqrt{\log(1/\xi)}$ such that $\Psi(\xi) \geq \mathcal{J}_B(\xi, \overline{\mathcal{P}}_k^{1/2}(\Theta, \xi))$ and $\xi_n = \xi\sqrt{\log(1/\xi)}$, Lemma B.2 gives us that

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

where C and c are universal positive constants depending only on Θ. Hence, the proof is completed.

B.2. Proof of Proposition 4.3

From Definition 4.1, since $M(X)$ is a bounded function of X, the arguments presented in Appendix B.1 still hold under the modified softmax gating multinomial logistic mixture of experts.

C. Proofs for the Identifiablity of the (Modified) Softmax Gating Multinomial Logistic MoE

In this appendix, we provide the proofs of Proposition 2.1 and Proposition 4.2 in Appendix C.1 and Appendix C.2, respectively.

C.1. Proof of Proposition 2.1

From the assumption of Proposition 2.1, the following equation holds for any $s \in [K]$ and almost surely $X \in \mathcal{X}$:

$$
\sum_{i=1}^{k} \frac{\exp((\beta_{1i})^\top X + \beta_{0i})}{\sum_{j=1}^{k} \exp((\beta_{1i})^\top X + \beta_{0i})} \cdot \frac{\exp(a_{is} + (b_{is})^\top X)}{\sum_{\ell=1}^{K} \exp(a_{i\ell} + (b_{i\ell})^\top X)}
$$
$$
= \sum_{i=1}^{k'} \frac{\exp((\beta'_{1i})^\top X + \beta'_{0i})}{\sum_{j=1}^{k'} \exp((\beta'_{1i})^\top X + \beta'_{0i})} \cdot \frac{\exp(a'_{is} + (b'_{is})^\top X)}{\sum_{\ell=1}^{K} \exp(a'_{i\ell} + (b'_{i\ell})^\top X)}. \tag{40}
$$

According to (Grün & Leisch, 2008), the multinomial logistic mixtures are identifiable, which implies that two mixing measures G and G' share the number of experts and the gating set of the mixing measure, i.e. $k = k'$ and

$$
\left\{ \frac{\exp((\beta_{1i})^\top X + \beta_{0i})}{\sum_{j=1}^{k} \exp((\beta_{1i})^\top X + \beta_{0i})} : i \in [k] \right\} \equiv \left\{ \frac{\exp((\beta'_{1i})^\top X + \beta'_{0i})}{\sum_{j=1}^{k} \exp((\beta'_{1i})^\top X + \beta'_{0i})} : i \in [k] \right\},
$$

for almost surely $X \in \mathcal{X}$. WLOG, we assume that

$$
\frac{\exp((\beta_{1i})^\top X + \beta_{0i})}{\sum_{j=1}^{k} \exp((\beta_{1i})^\top X + \beta_{0i})} = \frac{\exp((\beta'_{1i})^\top X + \beta'_{0i})}{\sum_{j=1}^{k} \exp((\beta'_{1i})^\top X + \beta'_{0i})},
$$

for any $i \in [k]$. As $\beta_{1k} = \beta'_{1k} = \mathbf{0}_d$ and $\beta_{0k} = \beta'_{0k} = 0$, the above result leads to $\beta_{1i} = \beta'_{1i}$ and $\beta_{0i} = \beta'_{0i}$ for all $i \in [k]$. Thus, the equation (40) becomes

$$
\sum_{i=1}^{k} \exp(\beta_{0i}) u(Y = s|X; \beta_{1i}, a_i, b_i) = \sum_{i=1}^{k} \exp(\beta_{0i}) u(Y = s|X; \beta_{1i}, a'_i, b'_i), \tag{41}
$$

for any $s \in [K]$ and almost surely $X \in \mathcal{X}$, where $u(Y = s|X; \beta_{1i}, a_i, b_i) := \exp(\beta_{1i}^\top X) \cdot \dfrac{\exp(a_{is} + (b_{is})^\top X)}{\sum_{\ell=1}^{K} \exp(a_{i\ell} + (b_{i\ell})^\top X)}$ and $a_i = (a_{i1}, a_{i2}, \ldots, a_{iK})$, $b_i = (b_{i1}, b_{i2}, \ldots, b_{iK})$.

Subsequently, we will consider r subsets of the set $[k]$, denoted by $S_1, S_2, \ldots, S_r$ that satisfy the following property: $\exp(\beta_{0i}) = \exp(\beta_{0i'})$ for any $i, i' \in S_t$ for $t \in [r]$. Therefore, we can rewrite equation (41) as

$$
\sum_{t=1}^{r} \sum_{i \in S_t} \exp(\beta_{0i}) u(Y = s|X; \beta_{1i}, a_i, b_i) = \sum_{t=1}^{r} \sum_{i \in S_t} \exp(\beta_{0i}) u(Y = s|X; \beta_{1i}, a'_i, b'_i),
$$

for any $s \in [K]$ and almost surely $X \in \mathcal{X}$. It follows from the above equation that for each $t \in [r]$, we get $\{(a_{i\ell} + (b_{i\ell})^\top X)_{\ell=1}^{K} : i \in S_t\} \equiv \{(a'_{i\ell} + (b'_{i\ell})^\top X)_{\ell=1}^{K} : i \in S_t\}$ for almost surely $X \in \mathcal{X}$. This leads to

$$
\left\{ (a_{i1}, \ldots, a_{iK}, b_{i1}, \ldots, b_{iK}) : i \in S_t \right\} \equiv \left\{ (a'_{i1}, \ldots, a'_{iK}, b'_{i1}, \ldots, b'_{iK}) : i \in S_t \right\}.
$$

Again, we may assume WLOG that $(a_{i1}, \ldots, a_{iK}, b_{i1}, \ldots, b_{iK}) = (a'_{i1}, \ldots, a'_{iK}, b'_{i1}, \ldots, b'_{iK})$ for any $i \in S_t$. As a result, we obtain that

$$
\sum_{t=1}^{r} \sum_{i \in S_t} \exp(\beta_{0i}) \delta_{(\beta_{1i}, a_{i1}, \ldots, a_{iK}, b_{i1}, \ldots, b_{iK})} = \sum_{t=1}^{r} \sum_{i \in S_t} \exp(\beta'_{0i}) \delta_{(\beta'_{1i}, a'_{i1}, \ldots, a'_{iK}, b'_{i1}, \ldots, b'_{iK})}.
$$

In other words, we achieve that $G \equiv G'$, which completes the proof.

C.2. Proof of Proposition 4.2

According to the assumption of Proposition 4.2, the following equation holds for any $s \in [K]$ and almost surely $X \in \mathcal{X}$:

$$
\sum_{i=1}^{k} \frac{\exp((\beta_{1i})^\top M(X) + \beta_{0i})}{\sum_{j=1}^{k} \exp((\beta_{1i})^\top M(X) + \beta_{0i})} \cdot \frac{\exp(a_{is} + (b_{is})^\top X)}{\sum_{\ell=1}^{K} \exp(a_{i\ell} + (b_{i\ell})^\top X)}
$$
$$
= \sum_{i=1}^{k'} \frac{\exp((\beta'_{1i})^\top M(X) + \beta'_{0i})}{\sum_{j=1}^{k'} \exp((\beta'_{1i})^\top M(X) + \beta'_{0i})} \cdot \frac{\exp(a'_{is} + (b'_{is})^\top X)}{\sum_{\ell=1}^{K} \exp(a'_{i\ell} + (b'_{i\ell})^\top X)}. \tag{42}
$$

Since the multinomial logistic mixtures are identifiable (see (Grün & Leisch, 2008)), two mixing measures G and G' admit the same number of experts and the same gating set, i.e. $k = k'$ and

$$\left\{ \frac{\exp((\beta_{1i})^\top M(X) + \beta_{0i})}{\sum_{j=1}^k \exp((\beta_{1i})^\top M(X) + \beta_{0i})} : i \in [k] \right\} \equiv \left\{ \frac{\exp((\beta'_{1i})^\top M(X) + \beta'_{0i})}{\sum_{j=1}^k \exp((\beta'_{1i})^\top M(X) + \beta'_{0i})} : i \in [k] \right\},$$

for almost surely $X \in \mathcal{X}$. WLOG, we assume that

$$\frac{\exp((\beta_{1i})^\top X + \beta_{0i})}{\sum_{j=1}^k \exp((\beta_{1i})^\top X + \beta_{0i})} = \frac{\exp((\beta'_{1i})^\top X + \beta'_{0i})}{\sum_{j=1}^k \exp((\beta'_{1i})^\top X + \beta'_{0i})},$$

for any $i \in [k]$. From the Definition 4.1, we know that $M(X)$ is a bounded function of X. Moreover, since $\beta_{1k} = \beta'_{1k} = \mathbf{0}_d$ and $\beta_{0k} = \beta'_{0k} = 0$, the above result implies that $\beta_{1i} = \beta'_{1i}$ and $\beta_{0i} = \beta'_{0i}$ for all $i \in [k]$. Therefore, the equation (42) can be reformulated as follows:

$$\sum_{i=1}^k \exp(\beta_{0i}) u(Y = s|X; \beta_{1i}, a_i, b_i) = \sum_{i=1}^k \exp(\beta_{0i}) u(Y = s|X; \beta_{1i}, a'_i, b'_i),$$

for any $s \in [K]$ and almost surely $X \in \mathcal{X}$, where $u(Y = s|X; \beta_{1i}, a_i, b_i) := \exp(\beta_{1i}^\top X) \cdot \dfrac{\exp(a_{is} + (b_{is})^\top X)}{\sum_{\ell=1}^K \exp(a_{i\ell} + (b_{i\ell})^\top X)}$ and $a_i = (a_{i1}, a_{i2}, \ldots, a_{iK})$, $b_i = (b_{i1}, b_{i2}, \ldots, b_{iK})$. Then, we can apply the arguments used in Appendix C.1 to deduce that $G \equiv G'$.

D. Simulation Studies

In this appendix, we carry out several numerical experiments to empirically verify our theoretical results regarding the convergence rates of maximum likelihood estimation in the standard softmax gating multinomial logistic MoE model under the Regime 1 in Appendix D.1. Meanwhile, under the Regime 2, we aim to empirically demonstrate the benefits of using modified softmax gating functions over the standard softmax gating function in the parameter estimation problem of the multinomial logistic MoE model in Appendix D.2.

D.1. Regime 1

Synthetic Data. We first sample the covariates X from the uniform distribution over $[0, 1]$. Then, we draw the response Y from the following conditional density $g_{G_*}(Y = s|X)$ of a softmax gating binomial logistic mixture of $k_* = 2$ experts:

$$g_{G_*}(Y = s|X) := \sum_{i=1}^2 \frac{\exp((\beta_{1i}^*)^\top X + \beta_{0i}^*)}{\sum_{j=1}^2 \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)} \times \frac{\exp(a_{is}^* + (b_{is}^*)^\top X)}{\sum_{\ell=1}^K \exp(a_{i\ell}^* + (b_{i\ell}^*)^\top X)}, \tag{43}$$

for $s \in [K]$, where $K = 2$. Here, the true mixing measure $G_* = \sum_{i=1}^2 \exp(\beta_{0i}^*) \delta_{(\beta_{1i}^*, a_{i1}^*, a_{i2}^*, b_{i1}^*, b_{i2}^*)}$ consists of $k_* = 2$ components with parameters given in Table 2, which satisfy the assumptions of Regime 1.

	Gating parameters	Expert parameters	
		Class 1	Class 2
$i = 1$	$(\beta_{01}^*, \beta_{11}^*) = (1, 3)$	$(a_{11}^*, b_{11}^*) = (-1, 2)$	$(a_{12}^*, b_{12}^*) = (0, 0)$
$i = 2$	$(\beta_{02}^*, \beta_{12}^*) = (0, 0)$	$(a_{21}^*, b_{21}^*) = (1, -1)$	$(a_{22}^*, b_{22}^*) = (0, 0)$

Table 2. True parameters under the Regime 1.

Initialization. We then compute the MLE $\widehat{G}_n$ w.r.t. with the number of components $k \in \{k_* + 1, k_* + 2\}$ for each sample using the EM algorithm (Dempster et al., 1977) with convergence criterion $\varepsilon = 10^{-6}$ and 2000 maximum EM iterations.

For each $k \in \{k_* + 1, k_* + 2\}$, we randomly assign elements of the set $\{1, 2, ..., k\}$ into k_* distinct sets $S_1, S_2, \ldots, S_{k_*}$, ensuring that each set contains at least one element. Moreover, we repeat this process for each replication. Following this, for each $i \in [k_*]$, we initialize the parameters by sampling from a Gaussian distribution with a mean centered around its true counterpart and a small variance. For each of 200 different choices of sample size n between $n_{\min} = 10^4$ and $n_{\max} = 10^5$, we generate 40 samples of size n. All the code for our simulation study was written in Python 3.9 on a standard Unix machine.

Empirical Convergence Rates. Subsequently, we report the empirical means of the discrepancy $\mathcal{D}_2$ between $\widehat{G}_n$ and G_*, and the choices of k under the Regime 1 in Figure 1. It can be observed from Figures 1a and 1b that the empirical vanishing rates of the average discrepancy $\mathcal{D}_2(\widehat{G}_n, G_*)$ are of orders $\widetilde{\mathcal{O}}(n^{-0.48})$ and $\widetilde{\mathcal{O}}(n^{-0.43})$ when $k = 3$ and $k = 4$, respectively. These rates are slightly slower than the theoretical rate of order $\widetilde{\mathcal{O}}(n^{-0.5})$ in Theorem 3.1. The main reason is that there has been only theoretical guarantee of global convergence for the parameter estimation under the mixture of experts with covariate-free gating function (see (Kwon et al., 2021; Kwon & Caramanis, 2020; Kwon et al., 2019)), while that for the softmax gating mixture of experts has remained missing in the literature. In order for the empirical vanishing rate to match the theoretical one, the sample size n must be large enough to compensate for the global convergence problem.
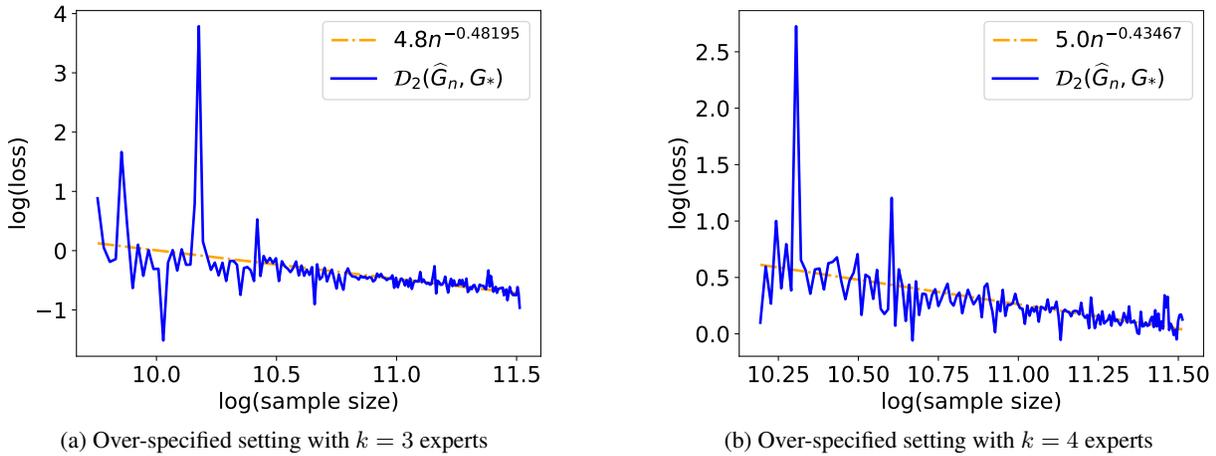

(a) Over-specified setting with $k = 3$ experts   (b) Over-specified setting with $k = 4$ experts

Figure 1. Two log-log scaled plots for the empirical convergence rates of the MLE $\widehat{G}_n$ when the true model in equation (43) is over-specified by a softmax gating binomial logistic mixture of $k = 3$ and $k = 4$ experts, respectively. In these figures, the empirical means of the discrepancy $\mathcal{D}_2(\widehat{G}_n, G_*)$ are illustrated by the blue curves, while the oranges dash-dotted lines represent for the least-squares fitted linear regression lines.

D.2. Regime 2

Synthetic Data. We also generate the covariates X from the uniform distribution over $[0, 1]$. For the multinomial logistic MoE model with standard softmax gating function, we draw the response Y from the conditional density $g_{G_*}(Y = s|X)$ given in equation (43), while for the modified softmax gating function, we sample Y from the following conditional density:

$$\widetilde{g}_{G_*}(Y = s|X) := \sum_{i=1}^{2} \frac{\exp((\beta_{1i}^*)^\top M(X) + \beta_{0i}^*)}{\sum_{j=1}^{2} \exp((\beta_{1j}^*)^\top M(X) + \beta_{0j}^*)} \times \frac{\exp(a_{is}^* + (b_{is}^*)^\top X)}{\sum_{\ell=1}^{K} \exp(a_{i\ell}^* + (b_{i\ell}^*)^\top X)}, \tag{44}$$

where we consider the standard softmax gating function $M(X) = X$ and $M(X) = \mathrm{sigmoid}(X)$. Next, while we keep the formulation of the true mixing measure $G_* = \sum_{i=1}^{2} \exp(\beta_{0i}^*) \delta_{(\beta_{1i}^*, a_{i1}^*, a_{i2}^*, b_{i1}^*, b_{i2}^*)}$, the parameter values are slightly changed to satisfy the assumptions of Regime 2. In particular, we set $b_{21}^* = b_{22}^* = 0$, while other parameters remains the same. More details can be found in Table 3.

D.2.1. EMPIRICAL CONVERGENCE RATES OF THE VORONOI-BASED LOSS

Initialization. We first initialize fitted parameters in the same fashion as those in Appendix D.1.

Standard Gating Function. For $k = 3$ we choose 35 different values of sample size n between $n_{\min} \approx 37 \times 10^3$ and $n_{\max} = 10^5$, while for $k = 4$ we select 28 different choices of sample size n between $n_{\min} \approx 50 \times 10^3$ and $n_{\max} = 10^5$.

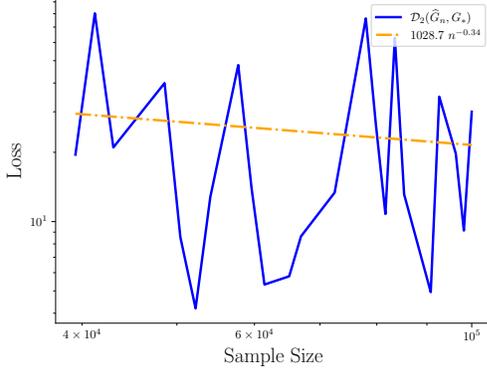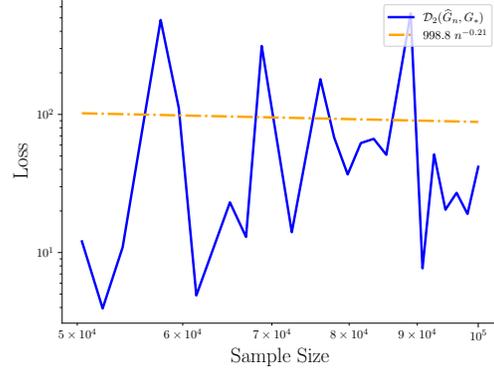|	Gating parameters	Expert parameters	
		Class 1	Class 2
$i = 1$	$(\beta_{01}^*, \beta_{11}^*) = (1, 3)$	$(a_{11}^*, b_{11}^*) = (-1, 2)$	$(a_{12}^*, b_{12}^*) = (0, 0)$
$i = 2$	$(\beta_{02}^*, \beta_{12}^*) = (0, 0)$	$(a_{21}^*, b_{21}^*) = (1, 0)$	$(a_{22}^*, b_{22}^*) = (0, 0)$

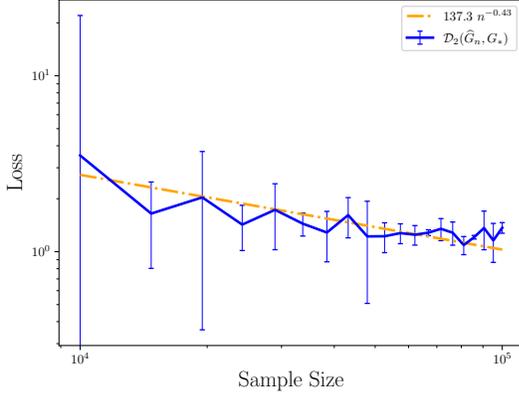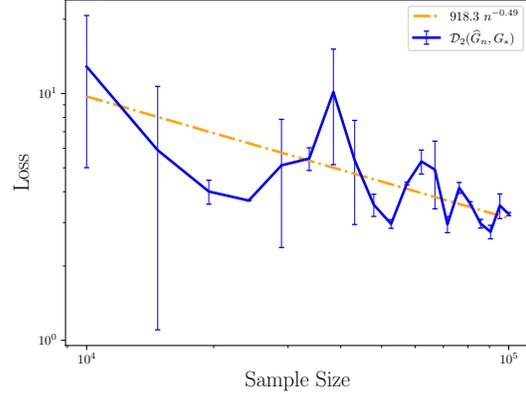*Table 3.* True parameters under the Regime 2.

(a) Over-specified setting with $k = 3$ experts, $M(X) = X$

(b) Over-specified setting with $k = 4$ experts, $M(X) = X$

(c) Over-specified setting with $k = 3$ experts, $M(X) = \mathrm{sigmoid}(X)$.

(d) Over-specified setting with $k = 4$ experts, $M(X) = \mathrm{sigmoid}(X)$.

Figure 2. log-log scaled plots for the empirical convergence rates of the MLE $\widehat{G}_n$ when the true model in equation (44) is over-specified by a softmax gating binomial logistic mixture with $M(X) = X$ and $M(X) = \mathrm{sigmoid}(X)$ of $k = 3$ and $k = 4$ experts, respectively. In these figures, the empirical means of the discrepancy $\mathcal{D}_2(\widehat{G}_n, G_*)$ are illustrated by the blue curves, while the oranges dash-dotted lines represent for the least-squares fitted linear regression lines.

In both cases we generate the corresponding samples of size n.

Modified Gating Function. We employ $M(X) = \mathrm{sigmoid}(X)$ and conduct 40 experiments for each sample size, covering a spectrum of 20 different sample sizes ranging from 10^4 to 10^5.

Empirical Convergence Rates. Subsequently, we report the empirical means of the discrepancy $\mathcal{D}_2$ between $\widehat{G}_n$ and G_*, and the choices of k under the Regime 2 for standard gating function and modified gating function in Figure 2. It can be observed from Figures 2a and 2b that the empirical vanishing rates of the average discrepancy $\mathcal{D}_2(\widehat{G}_n, G_*)$ are of orders $\widetilde{\mathcal{O}}(n^{-0.34})$ and $\widetilde{\mathcal{O}}(n^{-0.21})$ when $k = 3$ and $k = 4$, respectively. These rates are slightly slower than the theoretical rate of order $\widetilde{\mathcal{O}}(n^{-1/2r})$ for some $r \geq 1$ in Theorem 3.3. Moreover, for $M(X) = \mathrm{sigmoid}(X)$ as illustrated in Figures 2c and

2d the utilization of $M(X) = \text{sigmoid}(X)$ in the gating function resulted in an enhanced convergence rate of $\widetilde{\mathcal{O}}(n^{-1/2})$ for both $k = 3$ and $k = 4$. The main reasons are that the literature lacks a theoretical guarantee of global convergence for parameter estimation under the softmax mixture of experts, and that the MLE $\widehat{G}_n$ (which is the standard softmax mixture function) takes a very long time to converge to the true mixture measure G_* even when we use 2000 maximum number of EM iterations, see more in Appendices D.1 and D.2.2.

D.2.2. EMPIRICAL CONVERGENCE RATES OF THE EM ALGORITHM

Initialization. We assume that the MLEs $\widehat{G}_n$ (w.r.t. the standard softmax gating function) and $\widetilde{G}_n$ (w.r.t. the modified softmax gating functions) have $k = 3$ components. Next, we initialize fitted parameters in the same fashion as those in Appendix D.1. With the sample size $n = 10^4$, we run the EM algorithm for $N = 200$ iterations, and compute the negative log-likelihood value at each iteration.
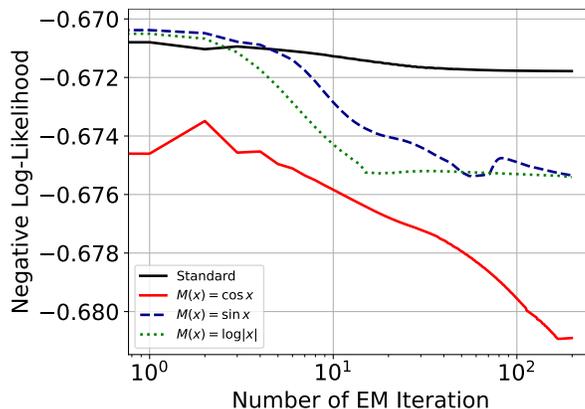

Figure 3. Empirical convergence rates of the EM algorithm with the standard softmax gating function and three different modified softmax gating functions with $M(X) \in \{\sin(X), \cos(X), \log(|X|)\}$. The y-axis indicates the negative log-likelihood, while the x-axis illustrates the number of EM iterations.

Negative Log-likelihood. It can be seen from Figure 3 that the negative log-likelihood corresponding to the modified softmax gating function with $M(X) = \cos(X)$ experiences a sharp drop after 200 iterations. Meanwhile, those for $M(X) = \sin(X)$ and $M(X) = \log(|X|)$ decline at a nearly same rate but slower than that for $M(X) = \cos(X)$. On the other hand, the negative log-likelihood corresponding to the standard softmax gating function remains almost unchanged. Those observations suggest that it would take a very long time for the MLE $\widehat{G}_n$ (w.r.t. the standard softmax gating function) to converge to the true mixing measure G_*. By contrast, if we use the modified softmax gating functions with $M(X) \in \{\sin(X), \cos(X), \log(|X|)\}$, the convergence rates of the corresponding MLE $\widetilde{G}_n$ to G_* would be substantially faster. As a consequence, this figure highlights the advantages of using the modified softmax gating functions over the standard softmax gating function in the parameter estimation of the multinomial logistic MoE model.