A Statistical Theory of Gated Attention through the Lens of Hierarchical Mixture of Experts

Viet Nguyen[1,⋆] **Tuan Minh Pham**[2,⋆] **Thinh Cao**[1,⋆] **Tan Dinh**[3]
Huy Nguyen[4] **Nhat Ho**[4,⋆⋆] **Alessandro Rinaldo**[4,⋆⋆]

[1]Hanoi University of Science and Technology [2]Purdue University
[3]Trivita AI [4]The University of Texas at Austin

Abstract

Self-attention has greatly contributed to the success of the widely used Transformer architecture by enabling learning from data with long-range dependencies. In an effort to improve performance, a gated attention model that leverages a gating mechanism within the multi-head self-attention has recently been proposed as a promising alternative. Gated attention has been empirically demonstrated to increase the expressiveness of low-rank mapping in standard attention and even to eliminate the attention sink phenomenon. Despite its efficacy, a clear theoretical understanding of gated attention's benefits remains lacking in the literature. To close this gap, we rigorously show that each entry in a gated attention matrix or a multi-head self-attention matrix can be written as a hierarchical mixture of experts. By recasting learning as an expert estimation problem, we demonstrate that gated attention is more sample-efficient than multi-head self-attention. In particular, while the former needs only a polynomial number of data points to recover the model parameters, the latter requires exponentially many data points to achieve the same estimation error. Furthermore, our analysis also provides a theoretical justification for why gated attention is more effective at mitigating the attention sink phenomenon and the low-rank bottleneck when a gate is placed at the output of the scaled dot product attention or the value map rather than at other positions in the multi-head self-attention architecture.

1 Introduction

The self-attention mechanism, introduced as a central component of the Transformer architecture by Vaswani et al. [34], has fundamentally transformed sequence modeling tasks in deep learning. Unlike recurrent and convolutional neural networks that process tokens sequentially or within fixed local windows, self-attention allows every position in a sequence to directly attend to all other positions in a single parallel operation. In particular, for each token in the input sequence, it produces a context vector as a weighted sum of all tokens, allocating higher weights to those that are more relevant to the current token's context. Therefore, the self-attention is capable of capturing long-range dependencies with unprecedented flexibility and efficiency. As a consequence, Transformers have served as the backbone of virtually all state-of-the-art large language models (GPT series [28], Gemini [7], LLaMA [8], DeepSeek [5], Qwen [30]), vision language models (ViT [6], Swin [19]), and multimodal models (CLIP [31], LLaVA [18]).

Despite the remarkable success of the self-attention architecture, there is still room for improvement. Firstly, since the value and dense projections (i.e., the matrices W_V and W_O; see Section 3 below for details) are two adjacent linear layers in the multi-head self-attention, they can be rewritten as one low-

⋆Equal contribution, ⋆⋆Co-last authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

rank linear projection [29]. This low-rank mapping turns out to limit the expressiveness of attention head linear combinations. Secondly, the attention sink phenomenon – whereby a disproportionate amount of attention mass is allocated to a small number of initial tokens even when those tokens carry little or no semantic relevance to the current prediction – is a widely observed behavior in autoregressive large language models [35, 9, 32]. To mitigate these problems, Qiu et al. [29] proposed using gated attention where a gating mechanism is applied at one of five positions in the multi-head self-attention: following the scaled dot product attention (SDPA) outputs (G1); after the value (G2), key (G3), and query projections (G4); and after the final dense output layer (G5). Through extensive experiments and comparisons, the authors noted that adopting gating mechanisms at positions G1 and G2 appears to yield the highest performance gains. This enhancement is attributed to two key properties of gating mechanisms: non-linearity and sparsity. In particular, the non-linearity augments the expressiveness of the aforementioned low-rank linear projection, while the sparsity helps circumvent the attention sink phenomenon. Although there is now compelling empirical evidence to corroborate the benefits of gated attention, a theoretical understanding of gated attention is still lacking in the literature.

Contributions. The main goal of this work is to theoretically understand the benefits of the gated attention mechanism from a statistical perspective, complementing the extensive empirical evidence provided in [29]. To this end, we establish a novel relation between multi-head attention/gated attention and hierarchical mixture of experts (HMoE) models [14]. Leveraging this connection, we analyze the sample complexity of gated attention by recasting it as a problem of expert specialization [27]. Here, the sample complexity refers to the number of samples required to recover the parameters of the attention block, rather than the number needed to achieve a given prediction error. Our contributions are threefold and can be summarized as follows.

1. Gated attention meets HMoE. In Section 3, we prove that each entry of a gated attention (resp. multi-head self-attention) matrix can be represented as a three-level HMoE with non-linear (resp. linear) experts.

2. Sample complexity of multi-head self-attention. Due to the simple linear structure of experts in the representation of multi-head self-attention, we show that the sample complexity of multi-head self-attention is exponential. Specifically, we prove that in order to estimate an expert within a given accuracy $\epsilon > 0$ (see also Table 1), one needs exponentially many data points, on the order of $\mathcal{O}(\exp(\epsilon^{-1/\tau}))$ for some positive constant τ. This result should discourage the use of vanilla self-attention.

3. Sample complexity of gated attention. We investigate the sample complexity of the gated attention variants G1 and G2 that were experimentally found in [29] to be most effective. Our theoretical results provide an explanation for why strongly identifiable activation functions, such as sigmoid and tanh with small bias, are more effective than the SiLU activation function. In our main result, we show that gating significantly reduces the sample complexity, from an exponential order to a polynomial order $\mathcal{O}(\epsilon^{-4})$ (see also Table 1). Hence, we claim that gated attention is more sample-efficient than multi-head self-attention.

In addition, we perform several numerical experiments to corroborate our theories in Section 6. Finally, in Section 7, we highlight some practical implications of our findings, and discuss a few limitations of our work as well as some potential future directions.

Notation. For any $n \in \mathbb{N}$, we set $[n] = \{1, 2, \ldots, n\}$. For a set S, $|S|$ denotes its cardinality. For a vector $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$, $\|v\|$ is its Euclidean (ℓ_2) norm. For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$, we write $|\alpha| = \sum_{i=1}^d \alpha_i$, $\alpha! = \alpha_1! \cdots \alpha_d!$, and for $v \in \mathbb{R}^d$, $v^\alpha = v_1^{\alpha_1} \cdots v_d^{\alpha_d}$. For positive sequences (a_n) and (b_n), we write $a_n = \mathcal{O}(b_n)$ or $a_n \lesssim b_n$ if $a_n \leq C b_n$ for all n, with a universal constant $C > 0$. For random positive sequences (t_n) and (s_n), $t_n = \mathcal{O}_P(s_n)$ means t_n/s_n is stochastically bounded: for any $\epsilon > 0$ there exists $M > 0$ such that $\mathbb{P}(t_n/s_n > M) < \epsilon$ for all sufficiently large n.

2 Related Work

Self-Attention and Transformers. The self-attention mechanism, introduced by Vaswani et al. [34], has become the de facto backbone of modern deep learning, enabling the modeling of long-range dependencies in foundation models. Despite its success, standard dot-product attention exhibits several limitations. First, the *attention sink* phenomenon causes autoregressive large language models

Table 1: Summary of convergence rates for attention mechanisms under different gating configurations. The parameter estimation rate refers to the convergence of the mixing measure under the Voronoi losses $\mathcal{L}_1$ (Theorem 1) and $\mathcal{L}_2$ (Theorems 2 and 3) defined in Sections 4 and 5, respectively. The sample complexity describes the number of data points required for expert estimation with error ϵ.

Mechanisms	**Expert estimation rates**	**Sample complexity**	**Theorems**
Multi-head self-attention	Slower than any polynomial order	$\mathcal{O}(\exp(\epsilon^{-1/\tau}))$	Theorem 1
Gated attention (Setting I)	$\mathcal{O}_P([\log(n)/n]^{1/2})$ to $\mathcal{O}_P([\log(n)/n]^{1/4})$	$\mathcal{O}(\epsilon^{-4})$	Theorem 2
Gated attention (Setting II)	$\mathcal{O}_P([\log(n)/n]^{1/2})$ to $\mathcal{O}_P([\log(n)/n]^{1/4})$	$\mathcal{O}(\epsilon^{-4})$	Theorem 3

(LLMs) to disproportionately attend to initial tokens regardless of their semantic relevance [35, 9]. Second, the low-rank bottleneck of multi-head attention limits its expressivity [2]. Finally, the quadratic time and memory complexity of standard attention poses challenges for long-context applications, motivating linear-time alternatives [15] and kernel-based approximations [4]. Together, these limitations highlight the structural and computational challenges of standard attention, particularly in efficient streaming applications.

Mixture of Experts (MoE) and Hierarchical MoE. Our theoretical framework builds on the literature of Mixture of Experts (MoE), pioneered by Jacobs et al. [11] and Jordan et al. [14]. MoE models enhance model capacity through conditional expert activation, while HMoE extends this approach to multi-level architectures for modeling complex data structures [12]. Recent developments include hierarchical representation learning for improved generalization [16] and hierarchical mixtures of low-rank experts for parameter-efficient adaptation of large language models [17]. From a statistical perspective, research has evolved from likelihood-based density estimation [13] toward understanding parameter estimation and expert specialization. Chen et al. [3] investigated the conditions underlying expert specialization in deep networks, whereas Ho et al. [10] and Nguyen et al. [25] derived convergence rates for Gaussian MoE and softmax-gated Gaussian MoE, respectively. Subsequent studies extended these analyzes to sparse [21], sigmoid [23], dense-to-sparse [24], and hierarchical MoE [22], revealing how different gating mechanisms affect expert parameter estimation rates.

Gated Attention Mechanisms. To address the limitations of standard attention, incorporating gating mechanisms into attention blocks has emerged as a promising approach. Ramapuram et al. [32] and Yan et al. [36] investigated the application of the sigmoid function to self-attention, provided comprehensive analyses, and demonstrated that this modification can improve performance. Subsequently, Qiu et al. [29] introduced gated attention, empirically demonstrating that non-linear gating enhances the expressiveness of low-rank mappings and mitigates attention sinks through induced sparsity. However, existing studies on gated attention remain largely empirical, lacking a rigorous explanation for the benefits of non-linear gating. We address this gap by establishing a connection between gated attention and HMoE, leveraging statistical tools from the MoE literature to characterize its improved sample efficiency. We note that although recent works [1, 36] have drawn connections between single-head attention and standard MoE architectures, their analyses do not fully capture multi-head attention, particularly the linear interaction between the value projection and output projection matrices, a key limitation addressed by gated attention [29].

3 Preliminaries

In this section, we present gated attention as an extension of multi-head attention, and then introduce hierarchical MoE. Later on, we will relate these two models and present a unified analysis.

Multi-head Self-attention (MHA). Denote the input sequence by $\mathbb{X} := [x_1, \ldots, x_N]^\top \in \mathbb{R}^{N\times d}$, where N is the sequence length and d is the embedding dimension. A single-head/Scaled dot-product attention (SDPA) is computed by

$$\mathrm{SDPA}(\mathbb{X}; W_Q, W_K, W_V) = \mathrm{softmax}\left(\frac{\mathbb{X}W_Q W_K^\top \mathbb{X}^\top}{\sqrt{d_v}}\right)\mathbb{X}W_V \in \mathbb{R}^{N\times d_v},$$

with learnable projection matrices $W_Q, W_K, W_V \in \mathbb{R}^{d\times d_v}$. Here and throughout, the dimensions N and d_v are assumed as given. Multi-head self-attention (MHA) extends SDPA by computing multiple attention heads in parallel, each of which processes the input sequence through a different learned

projection subspace. Formally, the output of an MHA layer is calculated as

$$\mathrm{MHA}(\mathbb{X}) := \mathrm{Concat}\big(\mathrm{head}_1, \ldots, \mathrm{head}_H\big) W_O = \sum_{h=1}^{H} \mathrm{softmax}\left(\frac{\mathbb{X} W_{Q,h} W_{K,h}^\top \mathbb{X}^\top}{\sqrt{d_v}}\right) \mathbb{X} W_{V,h} W_{O,h},$$

where $\mathrm{head}_h := \mathrm{SDPA}(\mathbb{X}; W_{Q,h}, W_{K,h}, W_{V,h})$ and $W_O := [W_{O,1}^\top, \ldots, W_{O,H}^\top]^\top \in \mathbb{R}^{Hd_v \times d}$.

Gated Attention. In conventional MHA, the composition of the two adjacent projections $W_{V,h}$ and $W_{O,h}$ results in an implicit low-rank linear transformation. To enhance the expressiveness of this low-rank structure, Qiu et al. [29] empirically demonstrated that introducing non-linearity either following the SDPA output or directly after the value output enhances the expressive power of the model. When a non-linear activation φ is applied to the value output, the output of the gated attention layer can be expressed as

$$\mathrm{GatedAttention}_{\mathrm{Value}}(\mathbb{X}) = \sum_{h=1}^{H} \mathrm{softmax}\left(\frac{\mathbb{X} W_{Q,h} W_{K,h}^\top \mathbb{X}^\top}{\sqrt{d_v}}\right) \varphi(\mathbb{X} W_{V,h}) W_{O,h}.$$

Alternatively, if the non-linearity is applied to the SDPA output, the gated attention layer can be written as

$$\mathrm{GatedAttention}_{\mathrm{SDPA}}(\mathbb{X}) = \sum_{h=1}^{H} \varphi\left(\mathrm{softmax}\left(\frac{\mathbb{X} W_{Q,h} W_{K,h}^\top \mathbb{X}^\top}{\sqrt{d_v}}\right) \mathbb{X} W_{V,h}\right) W_{O,h}.$$

Hierarchical mixture of experts (HMoE). The HMoE architecture [14] extends and adds structure to the standard MoE model [11] by organizing local experts into a tree-structured hierarchy rather than a flat architecture. This structure employs a multi-level probabilistic routing mechanism to partition complex input spaces. For instance, let us consider a two-level HMoE model as follows

$$y = \sum_{u=1}^{M} \phi_u(x) \sum_{v=1}^{K} \psi_{v|u}(x) \cdot \mathcal{E}_{v|u}(x).$$

The first level is governed by a root gating network $\phi(x) := [\phi_1(x), \phi_2(x), \ldots, \phi_M(x)]^\top$, whereas the second level contains conditional gating networks $\psi_{v|u}(x)$ embedded within each branch. The model output is then formed as a hierarchical combination of the leaf experts $\mathcal{E}_{v|u}(x)$.

Gated Attention meets Hierarchical Mixture of Experts. Recent studies [1, 36] have established a connection between single-head attention and standard MoE architectures. However, these analyzes do not fully reflect realistic multi-head attention, as they fail to account for the linear product between the value projection and the output matrix, which is precisely the key issue addressed by gated attention [29]. Therefore, in this work, we first establish a new connection between the outputs of multi-head attention/gated attention and HMoE models. Let $x = \mathrm{Vec}(\mathbb{X}) = (x_1^\top, \ldots, x_N^\top)^\top \in \mathbb{R}^{Nd}$ be the vectorization of $\mathbb{X}$ and set $P_h := \frac{W_{Q,h}(W_{K,h})^\top}{\sqrt{d_v}}$ and $J_i := \boldsymbol{e}_i^\top \otimes I_d \in \mathbb{R}^{d \times Nd}$, where $\otimes$ stands for Kronecker product. Then, J_i extracts the transpose of the i^{th} row of matrix $\mathbb{X}$: $J_i x = x_i$. Let us denote $M_{h,ij} = J_i^\top P_h J_j$ and $A_{h,j} = J_j^\top W_{V,h}$. As a result, the (i,j)-th entry of the softmax output can be expressed as $[\mathbb{X} P_h \mathbb{X}^\top]_{i,j} = x_i^\top P_h x_j = x^\top J_i^\top P_h J_j x = x^\top M_{h,ij} x$. Hence, the transpose of the i^{th} row of the multi-head self-attention matrix is given by

$$[\mathrm{MHA}(\mathbb{X})]_i^\top = \sum_{h=1}^{H} \sum_{j=1}^{N} \frac{\exp(x^\top M_{h,ij} x)}{\sum_{l=1}^{N} \exp(x^\top M_{h,il} x)} \cdot W_{O,h}^\top (A_{h,j}^\top) x.$$

Set $W_{O,h} = (\omega_{h,ii'})_{i \in [d_v], i' \in [d]}$ and let $a_{h,j,k} \in \mathbb{R}^{Nd}$ be the k-th column vector of the matrix $A_{h,j}$ Then one can verify that the (i,i')-th entry of the MHA matrix can be written in a HMoE form:

$$[\mathrm{MHA}(\mathbb{X})]_{i,i'} = \sum_{h=1}^{H} \sum_{k=1}^{d_v} \omega_{h,ki'} \sum_{j=1}^{N} \frac{\exp(x^\top M_{h,ij} x)}{\sum_{l=1}^{N} \exp(x^\top M_{h,il} x)} \cdot a_{h,j,k}^\top x.$$

Two other variants of gated attention are considered later in equations (8) and (11).

4 Sample Complexity of Multi-head Attention

In this section, we analyze the statistical sample complexity of multi-head self-attention in the expert specialization problem [27] by leveraging its HMoE representation. In particular, we analyze how fast an expert learns a specific region of the data. To begin, let us formally present the problem setup.

Problem setup. Assume that $\{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^{\bar{d}} \times \mathbb{R}$ are i.i.d samples from the regression model

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

where the regression function $x \mapsto f_{G_*}(x)$ is the (H^*, K^*, N^*)-HMoE model given by

$$f_{G^*}(x) = \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \omega_{h,k}^* \sum_{i=1}^{N^*} \frac{\exp(x^\top M_{h,i}^* x)}{\sum_{j=1}^{N^*} \exp(x^\top M_{h,j}^* x)} \cdot (a_{h,i,k}^*)^\top x. \tag{2}$$

Above, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent Gaussian noise variables such that $\mathbb{E}[\varepsilon_i | X_i] = 0$ and $\mathrm{Var}[\varepsilon_i | X_i] = \sigma^2$, for all $1 \leq i \leq n$. The covariates X_i's are supported on a subset $\mathcal{X}$ of $\mathbb{R}^{\bar{d}}$. Meanwhile, the ground-truth *mixing measure* $G_* := \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \omega_{h,k}^* \sum_{i=1}^{N^*} \delta_{(M_{h,i}^*, a_{h,i,k}^*)}$ inherits the hierarchical structure of the HMoE model, where $(\omega_{h,k}^*, M_{h,i}^*, a_{h,i,k}^*)_{h \in [H^*], k \in [K^*], i \in [N^*]}$ are true yet unknown parameters in the parameter space $\Theta \subset \mathbb{R} \times \mathbb{R}^{\bar{d} \times \bar{d}} \times \mathbb{R}^{\bar{d}}$.

Least squares estimation. As the convergence analysis would become needlessly complicated if H^* and N^* are not given, we assume that their values are known for ease of presentation, while the value of K^* still remains unknown. Under these assumptions, we over-specify the ground-truth model (2) by taking into account a least squares estimator within a class of (H^*, K, N^*)-HMoE models, where $K > K^*$ is given, as follows: $\widehat{G}_n \in \arg\min_{G \in \mathcal{G}_{H^*, K, N^*}(\Theta)} \sum_{i=1}^n \Big(Y_i - f_G(X_i)\Big)^2$, where $\mathcal{G}_{H^*, K, N^*}(\Theta) = \{G = \sum_{h=1}^{H^*} \sum_{k=1}^{K} \omega_{h,k} \sum_{i=1}^{N^*} \delta_{(M_{h,i}, a_{h,i,k})} : (\omega_{h,k}, M_{h,i}, a_{h,i,k}) \in \Theta\}$ stands for the set of all feasible mixing measures.

Assumptions. In our analysis, we make the following assumptions throughout, unless explicitly stated otherwise. (**A.1**) The parameter space Θ is compact and the input space $\mathcal{X} \subset \mathbb{R}^{\bar{d}}$ is bounded. This guarantees the convergence of least squares estimation. (**A.2**) All mixture weights $\omega_{h,k}^*$ are nonnegative, and at least one among them is strictly positive. This assumption ensures that the MoE model is well-defined. (**A.3**) The matrices $M_{h,i}$ are symmetric and $M_{h,N^*} = \mathbf{0}_{\bar{d} \times \bar{d}}$, for all $h \in [H^*], i \in [N^*]$. This normalization is required for model identifiability because softmax weights are invariant to translations. (**A.4**) For each $h \in [H^*]$, there exists $i \in [N^* - 1]$ such that $M_{h,i}^* \neq \mathbf{0}_{\bar{d} \times \bar{d}}$; this ensures that the softmax weights depend on the input value. (**A.5**) All the expert parameters $a_{h,i,k}^*$, for all $h \in [H^*], i \in [N^*], k \in [K^*]$ are nonzero and pairwise distinct.

We first show that any least-squared estimator of the regression function $f_{\widehat{G}_n}$ converges to the ground-truth regression function f_{G_*} at the parametric rate on the sample size in Proposition 1.

Proposition 1. *For the least squares estimator $\widehat{G}_n$, it holds that*

$$\|f_{\widehat{G}_n} - f_{G_*}\|_{L^2(\mu)} = \mathcal{O}_P([\log(n)/n]^{1/2}). \tag{3}$$

The proof of Proposition 1 is provided in Appendix B.1. This result implies that the regression function can be estimated at a nearly parametric rate. Next, to derive estimation rates for the parameters of the HMoE model of equation (2) from the above convergence result, we decompose the discrepancy $f_{\widehat{G}_n} - f_{G^*}$ into a sum of linearly independent components. This is achieved by applying a Taylor expansion to the product of the second-level gating weights and the expert function $u(x; M, a) := \exp(x^\top M x) \cdot (a^\top x)$. Unfortunately, there is an intrinsic interaction between the expert parameter a and the function $u(x; M, a)$, which is captured by the partial differential equation (PDE) as follows

$$a^\top \cdot \frac{\partial u(x; M, a)}{\partial a} = u(x; M, a). \tag{4}$$

The above PDE indicates that the function u and its partial derivatives from the Taylor expansion are linearly dependent. As a result, even if the regression error tends to zero, the coefficients of those terms, which encode parameter mismatches, need not vanish, so a small regression discrepancy does

not necessarily imply parameter convergence. To account for this coupling among parameters, we build a Voronoi loss function [20], and use it to elucidate how these interactions affect the parameter estimation rate.

Voronoi loss. For each mixing measure $G \in \mathcal{G}_{H^*,K,N^*}(\Theta)$, we consider the Voronoi cells $\{V_{h,k} \equiv V_{h,k}(G) : h \in [H^*], k \in [K^*]\}$ defined as the set of all pairs $(h', k') \in [H^*] \times [K]$ such that

$$\|f_G^{h',k'} - f_{G^*}^{h,k}\| \leq \|f_G^{h',k'} - f_{G^*}^{h_1,k_1}\|, \forall (h_1, k_1) \neq (h, k) \in [H^*] \times [K^*],$$

where we define $f_G^{h,k}(x) := \sum_{i=1}^{N^*} \frac{\exp(x^\top M_{h,i} x)}{\sum_{j=1}^{N} \exp(x^\top M_{h,j} x)} \cdot (a_{h,i,k}^\top x)$. Next, for each pair $(h', k') \in V_{h,k}$, we let $\kappa_G^{h',k'}$ be the permutation of the set $[N^*]$ such that

$$\|\theta_{h',\kappa_G^{h',k'}(i),k'} - \theta_{h,i,k}\| \leq \|\theta_{h',\kappa_G^{h',k'}(i),k'} - \theta_{h,i',k}\|,$$

for any $i' \neq i$, where $\theta_{h,i,k} := (M_{h,i}, a_{h,i,k})$. Without loss of generality, we assume that $\kappa_G^{h',k'}(i) = i$ for all $(h', k') \in [H^*] \times [K^*]$. Then, the Voronoi loss of interest is given by

$$\mathcal{L}_{1,r}(G, G_*) := \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \left| \sum_{(h',k') \in V_{h,k}} \omega_{h',k'} - \omega_{h,k}^* \right| + \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \sum_{(h',k') \in V_{h,k}} \omega_{h',k'} \cdot R_{h',h,k',k}(r),$$

where we denote $R_{h',h,k',k}(r) = \sum_{i=1}^{N^*} \|\Delta M_{h'h,i}\|^r + \|\Delta a_{h'h,i,k'k}\|^r$, where $\Delta M_{h'h,i} := M_{h',i} - M_{h,i}^*$ and $\Delta a_{h'h,i,k'k} := a_{h',i,k'} - a_{h,i,k}^*$. Based on the above Voronoi loss, we obtain the following minimax lower bound for estimating the true mixing measure G^*.

Theorem 1. *For any $r \geq 1$, it holds that*

$$\inf_{G_n \in \mathcal{G}_{H^*,K,N^*}(\Theta)} \sup_{G \in \mathcal{G}_{H^*,K,N^*}(\Theta) \setminus \mathcal{G}_{H^*,K^*-1,N^*}(\Theta)} \mathbb{E}_{f_G}\big[\mathcal{L}_{1,r}(G_n, G)\big] \gtrsim n^{-1/2}. \tag{5}$$

We remark that above, $\mathbb{E}_{f_G}$ denotes the expectation with respect to the joint distribution of the response and features obeying the model (1) with G in place of G^*, and the infimum is taken over all estimators G_n with values in $\mathcal{G}_{H^*,K,N^*}(\Theta)$.

The proof of Theorem 1 is provided in Appendix A.1. Combining the minimax lower bound with the Voronoi loss $\mathcal{L}_{1,r}$ shows that the estimation rates for parameters $\omega_{h,k}^*$, $M_{h,i}^*$, and $a_{h,i,k}^*$ are slower than any polynomial rate $\mathcal{O}_P(n^{-1/(2r)})$, for any $r \geq 1$. This behavior suggests that parameter estimation may achieve a logarithmic-type convergence rate $\mathcal{O}_P(1/\log^\lambda(n))$, for some constant $\lambda > 0$.

The resulting slow parameter convergence also adversely impacts the estimation of the expert functions. Since the input space $\mathcal{X}$ is bounded, the expert map $x \mapsto a^\top x$ is Lipschitz continuous over $\mathcal{X}$. Hence, there exists a constant $L > 0$ such that

$$\sup_{x \in \mathcal{X}} \big|(\widehat{a}_{h,i,k}^n)^\top x - (a_{h,i,k}^*)^\top x\big| \leq L \, \|\widehat{a}_{h,i,k}^n - a_{h,i,k}^*\|. \tag{6}$$

Consequently, expert estimation inherits the same slow rates for estimating parameters. As a consequence, we need exponentially many data points of the order $\mathcal{O}(\exp(\varepsilon^{-1/\lambda}))$ to approximate these experts with a given error ε. As will be shown in the following section 5, this slow rate in standard MHA is effectively addressed by the gated attention mechanism, which applies a non-linear activation function to the output of the value projections or after the SDPA output.

5 Sample Complexity of Gated Attention

In this section, we analyze the sample complexity of gated attention models in two settings where a nonlinear activation function is applied after the value projection (Setting I) and after the SDPA output (Setting II), respectively.

5.1 Setting I

We begin with Setting I, in which a nonlinear activation is applied after the value projection output. From the HMoE perspective in Section 3, this modification is equivalent to transforming a linear expert function $a^\top x$ in equation (2) to a non-linear one $\varphi(a^\top x)$.

Problem setup. We assume an i.i.d. sample $\{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^{\overline{d}} \times \mathbb{R}$ generated from the model

$$Y_i = g_{G_*}(X_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{7}$$

where the regression function $x \mapsto g_{G^*}(x)$ is now

$$g_{G^*}(x) := \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \omega_{h,k}^* \sum_{i=1}^{N^*} \frac{\exp(x^\top M_{h,i}^* x)}{\sum_{j=1}^{N^*} \exp(x^\top M_{h,j}^* x)} \cdot \varphi\left((a_{h,i,k}^*)^\top x\right). \tag{8}$$

Following the same reasoning as in Section 4, we tackle parameter estimation by applying a Taylor expansion to the function $\overline{u}(x; M, a) := \exp(x^\top M x) \cdot \varphi(a^\top x)$, which allows us to decompose the discrepancy $g_G(x) - g_{G^*}(x)$ into a collection of terms. To rule out unwanted interactions, we assume a *type-1 strong identifiability* condition on the activation function $\varphi(\cdot)$, which ensures that $\overline{u}(x; M, a)$ and its first- and second-order derivatives are linearly independent.

Definition 1. *(Type-1 Strong Identifiability). A function $\varphi : \mathbb{R} \to \mathbb{R}$ is said to be* type-1 strong identifiable *if it is injective, twice differentiable, and Lipschitz continuous, and if the collection of functions of x*

$$\bigcup_{i=1}^{N} \bigcup_{k=0}^{2} \left\{ x^\alpha \exp(x^\top M_i x) \varphi^{(k)}(a_i^\top x) : \alpha \in \mathcal{I}_k \right\},$$

where $\mathcal{I}_0 = \{\alpha \in \mathbb{N}^{\overline{d}} : |\alpha| \in \{0, 2, 4\}\}$, $\mathcal{I}_1 = \{\alpha \in \mathbb{N}^{\overline{d}} : |\alpha| \in \{1, 3\}\}$, and $\mathcal{I}_2 = \{\alpha \in \mathbb{N}^{\overline{d}} : |\alpha| = 2\}$, is linearly independent for any pairwise distinct nonzero parameters $a_1, a_2, \ldots, a_N$.

The above condition helps address the detrimental PDE-typed interaction in equation (4). In other words, it ensures the linear independence of terms in the Taylor expansion of the function $\overline{u}(x; M, a)$ in the decomposition of $g_G(x) - g_{G^*}(x)$, which is an important step in our proof techniques.

Examples. It can be verified that the function $\varphi(z) \in \{\mathrm{sigmoid}(z+b), \tanh(z+b), \mathrm{softplus}(z+b)\}$, where $b \neq 0$ is a bias, satisfies the type-1 strong identifiability condition. In contrast, the identity function $\varphi(z) = z$ fails to meet this condition as shown in the PDE in equation (4).

Next, we introduce a Voronoi-based loss to characterize the parameter estimation convergence rate in Setting I, as stated in Theorem 2. For notational convenience, we reuse the Voronoi cells $V_{h,k}$ from Section 4, defined as the collection of index pairs $(h', k') \in [H^*] \times [K]$ such that

$$\|\overline{f}_G^{h',k'} - \overline{f}_{G^*}^{h,k}\| \leq \|\overline{f}_G^{h',k'} - \overline{f}_{G^*}^{h_1,k_1}\|, \forall (h_1, k_1) \neq (h, k) \in [H^*] \times [K^*],$$

where $\overline{f}_G^{h,k}(x) := \sum_{i=1}^{N^*} \frac{\exp(x^\top M_{h,i} x)}{\sum_{j=1}^{N} \exp(x^\top M_{h,j} x)} \cdot \varphi(a_{h,i,k}^\top x)$. Hence, the Voronoi loss is defined as

$$\begin{aligned}
\mathcal{L}_2(G, G_*) := &\sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \left| \sum_{(h',k') \in V_{h,k}} \omega_{h',k'} - \omega_{h,k} \right| \\
&+ \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \sum_{\substack{(h',k') \in V_{h,k} \\ |V_{h,k}| = 1}} \omega_{h',k'} \cdot R_{h',h,k',k}(1) + \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \sum_{\substack{(h',k') \in V_{h,k} \\ |V_{h,k}| > 1}} \omega_{h',k'} \cdot R_{h',h,k',k}(2).
\end{aligned}$$

Theorem 2. *Under Setting I of gated attention, suppose that the activation function φ is type-1 strongly identifiable. Then, the following bound holds for any mixing measure $G \in \mathcal{G}_{H^*,K,N^*}(\Theta)$:*

$$\|g_G - g_{G^*}\|_{L^2(\mu)} \gtrsim \mathcal{L}_2(G, G^*).$$

This bound implies that $\mathcal{L}_2(\widehat{G}_n, G^*) = \mathcal{O}_P([\log(n)/n]^{1/2})$.

The proof of Theorem 2 is provided in Appendix A.2. This result reveals that parameters $(\omega_{h,k}^*, M_{h,i}^*, a_{h,i,k}^*)$ associated with a single fitted component ($|V_{h,k}| = 1$) achieve the standard parametric estimation rate $\mathcal{O}_P([\log(n)/n]^{1/2})$, ignoring logarithmic components. On the other hand, parameters that are approximated by multiple components ($|V_{h,k}| > 1$) converge at a slower rate of

order $\mathcal{O}_P([\log(n)/n]^{1/4})$. Since the function $x \mapsto \varphi(a^\top x)$ is Lipschitz continuous, there exists a constant $L_1 > 0$ such that

$$\sup_{x\in\mathcal{X}} \left|\varphi\left((\widehat{a}^n_{h,i,k})^\top x\right) - \varphi\left((a^*_{h,i,k})^\top x\right)\right| \leq L_1 \|\widehat{a}^n_{h,i,k} - a^*_{h,i,k}\|. \tag{9}$$

Consequently, the expert estimation rate also lies between $\mathcal{O}_P([\log(n)/n]^{1/4})$ and $\mathcal{O}_P([\log(n)/n]^{1/2})$, depending on the number of fitted experts $|V_{h,k}|$. Equivalently, achieving an approximation error ε requires at most a polynomial number of samples on the order of $\mathcal{O}(\varepsilon^{-4})$ to approximate these experts with a given error ε, which is substantially smaller than in the original MHA model in Section 4.

5.2 Setting II

We now turn to Setting II of Gated Attention in which the non-linear activation is applied after the SDPA output.

Problem setup. We assume that $\{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^{\bar{d}} \times \mathbb{R}$ is an i.i.d. sample of size n generated according to the following model

$$Y_i = g_{G_*}(X_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{10}$$

where the regression function $g_{G^*}(\cdot)$ takes the form:

$$g_{G^*}(x) := \sum_{h=1}^{H^*}\sum_{k=1}^{K^*} \omega^*_{h,k} \cdot \varphi\left(\sum_{i=1}^{N^*} \frac{\exp(x^\top M^*_{h,i} x)}{\sum_{j=1}^{N^*}\exp(x^\top M^*_{h,j} x)} \cdot (a^*_{h,i,k})^\top x\right), \tag{11}$$

The key difference between the regression functions in Setting I and Setting II lies in where the non-linear activation $\varphi(\cdot)$ is applied. In particular, this activation introduces non-linearity into the product of the second-level gating weights and the expert functions, rather than acting only on the expert functions as in Setting I. In this setting, the model can be viewed as an MoE model whose experts are themselves mixture models. Therefore, we impose an additional assumption to ensure that the HMoE model is well-defined and identifiable.

Assumption (A.6). For each $h \in [H^*]$, the expert parameters $M^*_{h,i}$ and $a^*_{h,i,k}$ are pairwise distinct across $i \in [N^*]$ and $k \in [K^*]$. Furthermore, the discrepancies between $M^*_{h,i} - M^*_{h,j}$, for $1 \leq i < j \leq N^*$ and $1 \leq h \leq H^*$, are pairwise distinct.

As in Setting I, our main challenge is to establish the required L^2 lower bound $\|g_G - g_{G^*}\|_{L^2(\mu)} \gtrsim \mathcal{L}_2(G, G^*)$, which is achieved by decomposing $g_{G_n}(x) - g_{G^*}(x)$ into linearly independent terms via a Taylor expansion of the function $\varphi(f^{h,k}_G(x))$. Therefore, we again need to establish a *type-2 strong identifiability* condition on φ to ensure the linear independence property.

Definition 2. *(Type-2 Strong Identifiability). A function $\varphi : \mathbb{R} \to \mathbb{R}$ is said to be* type-2 strong identifiable *if it is injective, twice differentiable, and Lipschitz continuous, and if the collection of functions of x*

$$\left\{\varphi\left(f^{h,k}_{G^*}(x)\right)\right\} \bigcup \left\{x^\alpha u(x; M^*_{h,i}, a^*_{h,i,k})\varphi'\left(f^{h,k}_{G^*}(x)\right) : \alpha \in \mathcal{J}\right\}_{(h,i,k)\in[H^*]\times[N^*]\times[K^*]}$$
$$\bigcup \left\{x^\alpha u(x; M^*_{h,i_1}, a^*_{h,i_1,k}) u(x; M^*_{h,i_2}, a^*_{h,i_2,k})\varphi''\left(f^{h,k}_{G^*}(x)\right) : \alpha \in \mathcal{J}\right\}_{(h,i,k)\in[H^*]\times[N^*]\times[K^*]}$$

is linearly independent, where $f^{h,k}_{G^*}(x) = \sum_{i=1}^{N^*} \frac{\exp(x^\top M^*_{h,i} x)}{\sum_{j=1}^{N^*}\exp(x^\top M^*_{h,j} x)} \cdot (a^*_{h,i,k})^\top x$.

Examples. Analogous to Setting I, the map $\varphi(z) \in \{\mathrm{sigmoid}(z+b), \tanh(z+b), \mathrm{softplus}(z+b)\}$ with a nonzero bias $b \neq 0$ satisfies the type-2 strong identifiability condition. By contrast, the identity activation $\varphi(z) = z$ does not meet this requirement and therefore cannot mitigate the interaction captured by the PDE in (4).

Now, we present the convergence rate of parameter estimation under Setting II in Theorem 3.

Theorem 3. *Under Setting II of gated attention, suppose that the activation function φ is type-2 strongly identifiable. Then, the following bound holds for any mixing measure $G \in \mathcal{G}_{H^*,K,N^*}(\Theta)$:*

$$\|g_G - g_{G^*}\|_{L^2(\mu)} \gtrsim \mathcal{L}_2(G, G^*).$$

This bound implies that $\mathcal{L}_2(\widehat{G}_n, G^*) = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}})$.

The proof of Theorem 3 is provided in Appendix A.3. Compared to the results of Setting I in Theorem 2, the convergence rates of parameter estimation and expert estimation in Setting II remain unchanged, which range from $\mathcal{O}_P([\log(n)/n]^{1/4})$ to $\mathcal{O}_P([\log(n)/n]^{1/2})$. Consequently, to approximate experts within a given error ε, it requires $\mathcal{O}(\varepsilon^{-4})$ number of samples.

5.3 Practical implications

Our theoretical results offer two notable insights for the design of attention mechanisms.

(I.1) Gated attention is more sample-efficient than Multi-head self-attention. Our key theoretical insight is that gated attention substantially enhances the sample efficiency compared to multi-head self-attention. As summarized in Table 1, while multi-head self-attention requires exponentially many samples to recover experts, gated attention needs only a polynomial number of samples. This provides a clear statistical justification for adopting gated attention to obtain better performance as suggested in [29].

(I.2) Placing a non-linear activation at appropriate locations fundamentally improves sample efficiency. Within gated attention architectures, not all placements of non-linearity are equally effective. We show that placing the non-linearity after the SDPA output or immediately after the value projection substantially improves sample efficiency. In particular, we rigorously demonstrate that the exponential-order sample complexity of vanilla MHA stems from the PDE-type interaction in equation (4), induced by the linear expert form $a^\top x$. Applying a non-linear activation after the SDPA output or the values makes the experts non-linear, e.g., $\varphi(a^\top x)$. In contrast, placing the activation after the queries or keys keeps the experts linear, according to the connection between gated attention and HMoE in Section 3. Hence, the PDE-type interaction in equation (4) still remains, and the model's sample efficiency is not improved.

Furthermore, our theoretical results shed light on why SDPA Gated Attention can mitigate the attention sink phenomenon.

(I.3) Explanation for why Gated Attention can mitigate the attention sink phenomenon. As demonstrated in [29], Gated Attention introduces sparse, input-dependent gating scores to suppress irrelevant value projections, thereby mitigating attention sinks and improving long-context extrapolation. From the HMoE perspective, this mechanism corresponds to assigning higher weights to informative experts while suppressing uninformative ones. Such selective weighting requires accurate estimation of gating parameters to distinguish between these experts. Since our analysis shows that Gated Attention has lower sample complexity than standard MHA, its gating parameters can be estimated more accurately with a given number of samples. Consequently, more reliable gating scores can facilitate the suppression of irrelevant experts, leading to sparser gating patterns.

6 Numerical Experiments

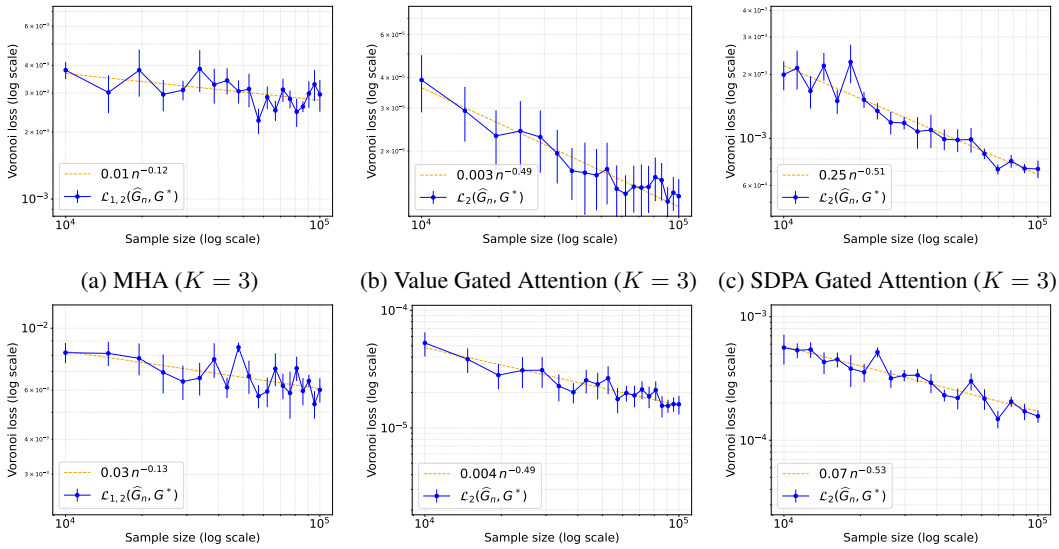
In this section, we conduct numerical experiments to empirically validate our theoretical findings that gated attention mechanisms are more sample-efficient compared to multi-head self-attention.

Overview of experiments: We design three experiments. *Experiment 1* uses a base configuration with $d = 2$, $H^* = N^* = K^* = 2$, and Sigmoid activation. *Experiment 2* repeats the base configuration with tanh activation to test robustness to the choice of nonlinearity. *Experiment 3* scales to $d = 3$ and $H^* = N^* = K^* = 3$ (with $K \in \{4, 5\}$ fitted channels) to evaluate larger configurations. We report Experiment 1 in the main text and defer Experiments 2 and 3 to Appendix C.

Data generation: For each sample size n, we generate i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^n$ by drawing X_i from the uniform distribution over $[-1, 1]^d$ and sampling Y_i from the true regression function specified in each theorem configuration: $f_{G_*}(X)$ for multi-head self-attention (Model (2)), $g_{G^*}(X)$ for gated attention Setting I (Model (8)), and $g_{G^*}(X)$ for gated attention Setting II (Model (11)), with additive Gaussian noise $\varepsilon_i \sim \mathcal{N}(0, \nu^2)$ and $\nu = 0.1$.

Experimental setup: Ground-truth parameters are listed in Appendix C, Table 2 (Experiments 1–2) and Table 3 (Experiment 3).

Training procedure: For each $n \in [10^4, 10^5]$, we conduct 10 runs. We consider $K \in \{3, 4\}$ fitted channels in Experiments 1–2 and $K \in \{4, 5\}$ in Experiment 3 to evaluate different degrees of over-specification. Parameters are initialized near the true values with perturbations of order



(a) MHA ($K = 3$) (b) Value Gated Attention ($K = 3$) (c) SDPA Gated Attention ($K = 3$)

(d) MHA ($K = 4$) (e) Value Gated Attention ($K = 4$) (f) SDPA Gated Attention ($K = 4$)

Figure 1: Log-log plots of empirical Voronoi losses versus sample size n for multi-head self-attention and gated attention mechanisms with Sigmoid activation function (Experiment 1).

$O(n^{-0.083})$, and we fit via gradient descent with backtracking line search for up to 1000 epochs and an initial learning rate $\eta = 0.05$. We report the mean and standard deviation of the Voronoi losses in Figure 1 for Experiment 1, Figure 2 for Experiment 2, and Figure 3 for Experiment 3.

Results of Experiment 1: Figure 1 shows that gated attention with Sigmoid activation achieves fast convergence under both settings. In Setting I, the Voronoi loss decays at empirical rates of about $O(n^{-0.49})$ in Figures 1b and 1e. In Setting II, the rates are similarly fast, around $O(n^{-0.51})$ and $O(n^{-0.53})$ in Figures 1c and 1f. By contrast, vanilla multi-head self-attention converges much more slowly, with rates $O(n^{-0.12})$ and $O(n^{-0.13})$ in Figures 1a and 1d. These results empirically confirm the sample-efficiency advantage of gated attention.

7 Discussion

This paper provides a theoretical justification for the advantage of gated attention over vanilla MHA. First, we establish a novel link between gated attention/MHA and hierarchical mixture of experts. From the perspective of expert specialization in HMoE, we show that MHA suffers from exponential-order sample complexity due to a PDE-type interaction. In contrast, by introducing non-linearity either after the SDPA output or after the value projection, gated attention improves the sample complexity to polynomial order. These results indicate that gated attention is more sample-efficient than multi-head self-attention. More importantly, our theory reveals that the PDE-type interaction in multi-head self-attention is caused by the linear form of experts. Thus, placing the non-linearity after the SDPA output or the values resolves this issue by making the experts non-linear, whereas applying it at other positions leaves the experts linear and the problem unresolved.

There are some potential research directions that arise from our work. First, our theoretical analysis is conducted in a well-specified setting where the data are generated from a regression framework whose regression function is an HMoE model. Future work can extend our results to cases where the ground-truth function is not necessarily an HMoE model. Second, this paper compares gated attention with multi-head self-attention through the statistical problem of estimating experts in their representations. We believe that comparing these two models from another perspective, such as universal approximation capability as in [32], would further strengthen the claim that gated attention can outperform multi-head self-attention.

Impact Statement

This paper aims to contribute to the advancement of machine learning. Due to its primarily theoretical focus, we do not anticipate any direct societal impacts arising from this work.

References

[1] P. Akbarian, H. Nguyen, X. Han, and N. Ho. Quadratic gating mixture of experts: Statistical insights into self-attention. *arXiv preprint arXiv:2410.11222*, 2024.

[2] S. Bhojanapalli, C. Yun, A. S. Rawat, S. Reddi, and S. Kumar. Low-rank bottleneck in multi-head attention models. In *International Conference on Machine Learning*, 2020.

[3] Z. Chen, Y. Deng, Y. Wu, Q. Gu, and Y. Li. Towards understanding the mixture-of-experts layer in deep learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23049–23062. Curran Associates, Inc., 2022.

[4] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, and D. Belanger. Rethinking attention with Performers. In *International Conference on Learning Representations*, 2021.

[5] DeepSeek-AI et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[7] GeminiTeam. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[8] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[9] X. Gu, T. Pang, C. Du, Q. Liu, F. Zhang, C. Du, Y. Wang, and M. Lin. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

[11] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

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

[13] W. Jiang and M. A. Tanner. On the Approximation Rate of Hierarchical Mixtures-of-Experts for Generalized Linear Models. *Neural Computation*, 11(5):1183–1198, July 1999.

[14] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

[15] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5156–5165, 2020.

[16] W. Li, D. Wang, Z. Ding, A. Sohrabizadeh, Z. Qin, J. Cong, and Y. Sun. Hierarchical mixture of experts: Generalizable learning for high-level synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[17] M. Liao, W. Chen, J. Shen, S. Guo, and H. Wan. HMoRA: Making LLMs More Effective with Hierarchical Mixture of LoRA Experts. In *International Conference on Learning Representations*, 2025.

[18] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[19] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

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

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

[22] H. Nguyen, X. Han, C. W. Harris, S. Saria, and N. Ho. On expert estimation in hierarchical mixture of experts: Beyond softmax gating functions. *arXiv preprint arXiv:2410.02935*, 2024.

[23] H. Nguyen, N. Ho, and A. Rinaldo. Sigmoid gating is more sample efficient than softmax gating in mixture of experts. In *Advances in Neural Information Processing Systems*, 2024.

[24] H. Nguyen, N. Ho, and A. Rinaldo. Convergence rates for softmax gating mixture of experts. *IEEE Transactions on Information Theory*, 72(2):1276–1304, 2026.

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

[26] N. V. Nguyen, H. Nguyen, Q. Pham, V. Nguyen, S. Ramasamy, and N. Ho. CompeteS-MoE: Statistically guaranteed mixture of experts training via competition. *arXiv preprint arXiv:2505.13380*, 2025.

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

[28] OpenAI et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.

[29] Z. Qiu, Z. Wang, B. Zheng, Z. Huang, K. Wen, S. Yang, R. Men, L. Yu, F. Huang, S. Huang, D. Liu, J. Zhou, and J. Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[30] Qwen et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[32] J. Ramapuram, F. Danieli, E. G. Dhekane, F. Weers, D. Busbridge, P. Ablin, T. Likhomanenko, J. Digani, Z. Gu, A. Shidani, et al. Theory, analysis, and best practices for sigmoid self-attention. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[35] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024.

[36] F. Yan, H. Nguyen, P. Akbarian, N. Ho, and A. Rinaldo. Sigmoid self-attention has lower sample complexity than softmax self-attention: A mixture-of-experts perspective. *arXiv preprint arXiv:2502.00281*, 2025.

[37] B. Yu. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*, pages 423–435, 1997.

Supplementary Material for "A Statistical Theory of Gated Attention Mechanism through the Lens of Hierarchical Mixture of Experts"

In this supplementary material, we first expand on related works in Section 2, providing additional context and connections to existing literature. We then present rigorous proofs for results regarding the convergence rates of parameter estimation under the hierarchical mixture of experts representations of attention mechanisms in Section A. Specifically, we provide proofs for Theorem 1 on multi-head self-attention in Appendix A.1, Theorem 2 on gated attention (Setting I) in Appendix A.2, and Theorem 3 on gated attention (Setting II) in Appendix A.3. Next, we provide proofs for auxiliary results, including Proposition 1 in Appendix B. Finally, we provide the true parameter configurations used in our numerical experiments in Appendix C.

A Proofs of Main Results

A.1 Proof of Theorem 1

Proof overview. We begin by establishing that

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{G}_k(\Theta):\, \mathcal{L}_{1,r}(G, G^*) \leq \varepsilon} \frac{\|f_G - f_{G^*}\|_{L^2(\mu)}}{\mathcal{L}_{1,r}(G, G^*)} = 0 \tag{12}$$

for any $r \geq 1$. With this result, we then obtain the minimax lower bound stated in Theorem 1:

$$\inf_{G_n \in \mathcal{G}_{H^*, K, N^*}(\Theta)} \sup_{G \in \mathcal{G}_{H^*, K, N^*}(\Theta) \setminus \mathcal{G}_{H^*, K^*-1, N^*}(\Theta)} \mathbb{E}_{f_G}\big[\mathcal{L}_{1,r}(G_n, G)\big] \gtrsim n^{-1/2}. \tag{13}$$

Proof of equation (12)**.** To show this, it is enough to exhibit a sequence of mixing measures (G_n) with

$$\mathcal{L}_{1,r}(G_n, G^*) \to 0 \quad \text{and} \quad \frac{\|f_{G_n} - f_{G^*}\|_{L_2(\mu)}}{\mathcal{L}_{1,r}(G_n, G^*)} \to 0, \text{ as } n \to \infty. \tag{14}$$

Consider the sequence (G_n) given by

$$G_n := \sum_{h=1}^{H^*} \sum_{k=1}^{K} \omega_{h,k}^{(n)} \sum_{i=1}^{N^*} \delta_{(M_{h,i}^{(n)}, a_{h,i,k}^{(n)})}, \tag{15}$$

where $K = K^* + 1$ and for all $h \in [H^*]$ and $i \in [N^*]$, the parameters are specified as follows:

$$\omega_{h,1}^{(n)} = \omega_{h,2}^{(n)} = \frac{1}{2}\omega_{h,1}^* + \frac{1}{2n^{r+1}}, \qquad \omega_{h,k}^{(n)} = \omega_{h,k-1}^*, \quad 3 \leq k \leq K, \tag{16}$$

$$M_{h,i}^{(n)} = M_{h,i}^*, \quad 1 \leq i \leq N^*, \tag{17}$$

and for the expert parameters,

$$a_{h,i,1}^{(n)} = a_{h,i,1}^* + \frac{1}{n}e_i, \qquad a_{h,i,2}^{(n)} = a_{h,i,1}^* - \frac{1}{n}e_i, \qquad a_{h,i,k}^{(n)} = a_{h,i,k-1}^*, \quad 3 \leq k \leq K, \tag{18}$$

with $e_i = J_i^\top u$, where $u \in \mathbb{R}^d$ is any unit vector satisfying $\|u\| = 1$.

Under the above construction, the Voronoi cells $\{V_{h,k}(G_n)\}$ can be chosen as follows. For each $h \in [H^*]$, we have

$$V_{h,1}(G_n) = \{(h,1), (h,2)\}, \qquad V_{h,k}(G_n) = \{(h, k+1)\}, \quad 2 \leq k \leq K^*.$$

Indeed, the two components $(h,1)$ and $(h,2)$ are both constructed to be closest to $f_{G^*}^{h,1}$, while for each $k \geq 2$ the component $(h, k+1)$ is an exact copy of $f_{G^*}^{h,k}$ up to the index shift.

Next, for $(h', k') \in V_{h,1}(G_n)$, we have

$$\Delta M_{h',i,\,k'} = 0, \qquad \Delta a_{h',i,\,k'} = a_{h,i,k'}^{(n)} - a_{h,i,1}^* = \frac{(-1)^{k'-1}}{n}e_i, \qquad k' \in \{1, 2\},$$

while for $2 \le k \le K^*$ and $(h', k') = (h, k+1) \in V_{h,k}(G_n)$, we have

$$\Delta M_{h',i,\,k'} = 0, \qquad \Delta a_{h',i,\,k'} = a^{(n)}_{h,i,k+1} - a^*_{h,i,k} = 0,$$

so that the corresponding $(k \ge 2)$ terms in $\mathcal{L}_{1,r}(G_n, G^*)$ vanish.

Hence, by the definition of $\mathcal{L}_{1,r}$, we obtain

$$\begin{aligned}
\mathcal{L}_{1,r}(G_n, G^*) &= \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\left|\sum_{(h',k')\in V_{h,k}(G_n)} \omega^{(n)}_{h',k'} - \omega^*_{h,k}\right| \\
&\quad + \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\sum_{(h',k')\in V_{h,k}(G_n)} \omega^{(n)}_{h',k'} \sum_{i=1}^{N^*}\Big(\|\Delta M_{h',i,\,k'}\|^r + \|\Delta a_{h',i,\,k'}\|^r\Big) \\
&= \frac{H^*}{n^{r+1}} + \sum_{h=1}^{H^*}\Big(\omega^{(n)}_{h,1} + \omega^{(n)}_{h,2}\Big)\sum_{i=1}^{N^*}\Big\|\frac{1}{n}e_i\Big\|^r \\
&= \frac{H^*}{n^{r+1}} + \sum_{h=1}^{H^*}(\omega^*_{h,1} + \frac{1}{n^{r+1}})\,N^*\,\frac{1}{n^r} = O\Big(\frac{1}{n^r}\Big). \qquad (19)
\end{aligned}$$

The above formulation ensures that $\mathcal{L}_{1,r}(G_n, G^*) \to 0$ as $n \to \infty$. Therefore, it remains to establish (14). For this purpose, for each $(h,k) \in [H^*] \times [K]$, define

$$u^{h,k}_G(x) := \sum_{i=1}^{N^*} \frac{\exp\big(x^\top M_{h,i} x\big)}{\sum_{t=1}^{N^*}\exp\big(x^\top M_{h,t} x\big)}\,(a_{h,i,k})^\top x. \qquad (20)$$

Now, we have $f_G(x) = \sum_{h=1}^{H^*}\sum_{k=1}^{K}\omega_{h,k}\,u^{h,k}_G(x)$. We can partition

$$f_{G_n}(x) - f_{G^*}(x) = \sum_{h=1}^{H^*}\left[\sum_{k'=1}^{K}\omega^{(n)}_{h,k'}\,u^{h,k'}_{G_n}(x) - \sum_{k=1}^{K^*}\omega^*_{h,k}\,u^{h,k}_{G^*}(x)\right]. \qquad (21)$$

Next, for each fixed $h \in [H^*]$, we decompose

$$\begin{aligned}
&\left[\sum_{k'=1}^{K}\omega^{(n)}_{h,k'}\,u^{h,k'}_{G_n}(x) - \sum_{k=1}^{K}\omega^*_{h,k}\,u^{h,k}_{G^*}(x)\right] \\
&= \left[\omega^{(n)}_{h,1}u^{h,1}_{G_n}(x) + \omega^{(n)}_{h,2}u^{h,2}_{G_n}(x) - \big(\omega^{(n)}_{h,1} + \omega^{(n)}_{h,2}\big)u^{h,1}_{G^*}(x)\right] \\
&\quad + \Big(\omega^{(n)}_{h,1} + \omega^{(n)}_{h,2} - \omega^*_{h,1}\Big)\,u^{h,1}_{G^*}(x) \\
&\quad + \sum_{k=2}^{K^*}\Big[\omega^{(n)}_{h,k+1}u^{h,k+1}_{G_n}(x) - \omega^*_{h,k}u^{h,k}_{G^*}(x)\Big] \\
&=: A_{n,h}(x) + B_{n,h}(x) + C_{n,h}(x). \qquad (22)
\end{aligned}$$

We now evaluate the three terms. By the construction of G_n, the head-shift part satisfies $\omega^{(n)}_{h,k+1} = \omega^*_{h,k},\ u^{h,k+1}_{G_n}(x) = u^{h,k}_{G^*}(x),\ 2 \le k \le K$, and hence $C_{n,h}(x) = 0$. Moreover, using $\omega^{(n)}_{h,1} = \omega^{(n)}_{h,2}$ together with the symmetric perturbation $a^{(n)}_{h,i,1} = a^*_{h,i,1} + \frac{1}{n}e_i,\ a^{(n)}_{h,i,2} = a^*_{h,i,1} - \frac{1}{n}e_i$, we obtain exact cancellation in the pivot term:

$$A_{n,h}(x) = \omega^{(n)}_{h,1}\Big(u^{h,1}_{G_n}(x) + u^{h,2}_{G_n}(x) - 2u^{h,1}_{G^*}(x)\Big) = 0. \qquad (23)$$

Finally, since $\omega^{(n)}_{h,1} + \omega^{(n)}_{h,2} - \omega^*_{h,1} = \frac{1}{n^{r+1}}$, the remaining term satisfies

$$B_{n,h}(x) = \frac{1}{n^{r+1}}u^{h,1}_{G^*}(x) = O\Big(\frac{1}{n^{r+1}}\Big). \qquad (24)$$

Combining the above, we conclude that,

$$f_{G_n}(x) - f_{G^*}(x) = \sum_{h=1}^{H^*}(A_{n,h}(x) + B_{n,h}(x) + C_{n,h}(x)) = \sum_{h=1}^{H^*} B_{n,h}(x),$$

where $A_{n,h}(x) = 0$, $C_{n,h}(x) = 0$, and $B_{n,h}(x) = O(\frac{1}{n^{r+1}})$ for almost every x. Since $\mathcal{L}_{1,r}(G_n, G^*) = O(\frac{1}{n^r})$, it follows that

$$\frac{f_{G_n}(x) - f_{G^*}(x)}{\mathcal{L}_{1,r}(G_n, G^*)} = \frac{\sum_{h=1}^{H^*} B_{n,h}(x)}{\mathcal{L}_{1,r}(G_n, G^*)} \to 0 \quad \text{for almost every } x. \tag{25}$$

As a consequence,

$$\frac{\|f_{G_n} - f_{G^*}\|_{L_2(\mu)}}{\mathcal{L}_{1,r}(G_n, G^*)} \to 0 \quad \text{as } n \to \infty.$$

This completes the proof of (12).

Proof of equation (13)**.** Given that the noise variables $\varepsilon_i \mid x_i$ are Gaussian,

$$Y_i \mid x_i \sim \mathcal{N}^*(f_{G^*}(x_i), \sigma^2) \quad \text{for all } i \in [n].$$

From (12), for sufficiently small $\varepsilon > 0$ and some constant $c > 0$ (to be chosen later), we can find $G*' \in \mathcal{G}_k(\Theta)$ such that

$$\mathcal{L}_{1,r}(G'_*, G^*) = 2\varepsilon \quad \text{and} \quad \|f_{G'_*} - f_{G^*}\|_{L_2(\mu)} \le c \cdot \varepsilon.$$

Applying Le Cam's lemma [37] and using the weak triangle inequality for $\mathcal{L}_{1,r}$, we obtain

$$\begin{aligned}
&\inf_{G_n \in \mathcal{G}_{H^*,K,N^*}(\Theta)} \sup_{G \in \mathcal{G}_{H^*,K,N^*}(\Theta) \setminus \mathcal{G}_{H^*,K^*-1,N^*}(\Theta)} \mathbb{E}_{f_G}\big[\mathcal{L}_{1,r}(G_n, G)\big] \\
&\gtrsim \frac{\mathcal{L}_{1,r}(G'_*, G^*)}{8} \exp\Big(- n\, \mathbb{E}_{x\sim\mu}\big[\mathrm{KL}(\mathcal{N}^*(f_{G'_*}(x), \sigma^2), \mathcal{N}^*(f_{G^*}(x), \sigma^2))\big]\Big) \\
&\gtrsim \varepsilon \cdot \exp\big(- n \|f_{G'_*} - f_{G^*}\|^2_{L_2(\mu)}\big) \gtrsim \varepsilon \cdot \exp(-cn\varepsilon^2).
\end{aligned} \tag{26}$$

Here, the second inequality uses $\mathrm{KL}\big(\mathcal{N}^*(f_{G'_*}(x), \sigma^2), \mathcal{N}^*(f_{G^*}(x), \sigma^2)\big) = \frac{\big(f_{G'_*}(x) - f_{G^*}(x)\big)^2}{2\sigma^2}$.

Choosing $\varepsilon = n^{-1/2}$ gives

$$\varepsilon \cdot \exp(-cn\varepsilon^2) = n^{-1/2} \exp(-c).$$

Thus, the minimax lower bound (13) follows, completing the proof.

A.2 Proof of Theorem 2

In this proof, we first show that

$$\inf_{G \in \mathcal{G}_{H^*,K,N^*}(\Theta)} \|g_G - g_{G^*}\|_{L^2(\mu)} / \mathcal{L}_2(G, G^*) > 0 \tag{27}$$

We separate the proof of (27) into a global estimate and a local estimate. The local portion requires showing that (27) holds whenever $\mathcal{L}_2(G, G_*)$ is small enough as follows

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{G}_{H^*,K,N^*}(\Theta): \mathcal{L}_2(G,G^*) \le \varepsilon} \|g_G - g_{G^*}\|_{L^2(\mu)} / \mathcal{L}_2(G, G^*) > 0. \tag{28}$$

In the local regime, the problem is addressed using Taylor expansion as the main analytical tool. In contrast, the global part of the proof examines the behavior when $\mathcal{L}_2(G, G_*)$ is sufficiently large.

$$\inf_{G \in \mathcal{G}_{H^*,K,N^*}(\Theta): \mathcal{L}_2(G,G^*) > \varepsilon} \|g_G - g_{G^*}\|_{L^2(\mu)} / \mathcal{L}_2(G, G^*) > 0. \tag{29}$$

Proof of the local part (Equation (28)**).**

Suppose that this local inequality fails. Then there exists a sequence of mixing measures $(G_n) \in \mathcal{G}_{H^*K^*N}(\Theta)$ such that $\mathcal{L}_2(G_n, G^*) \to 0$ and $\|g_{G_n} - g_{G^*}\|_{L^2(\mu)}/\mathcal{L}_2(G_n, G^*) \to 0$ when $n \to \infty$.
Step 1. First of all, we denote

$$\overline{f}^{h',k'}_{G_n}(x) = \sum_{i=1}^{N^*} \frac{\exp(x^\top M^n_{h',i}x)}{\sum_{j=1}^{N^*}\exp(x^\top M^n_{h',j}x)} \cdot \varphi((a^n_{h',i,k'})^\top x);$$

$$\overline{f}^{h,k}_{G^*}(x) = \sum_{i=1}^{N^*} \frac{\exp(x^\top M^*_{h,i}x)}{\sum_{j=1}^{N^*}\exp(x^\top M^*_{h,j}x)} \cdot \varphi((a^*_{h,i,k})^\top x).$$

Then, we have $g_{G_n}(x) = \sum_{h'=1}^{H^*}\sum_{k'=1}^{K}\omega^n_{h',k'} \cdot \overline{f}^{h',k'}_{G_n}(x)$, $g_{G^*}(x) = \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\omega^*_{h,k} \cdot \overline{f}^{h,k}_{G^*}(x)$. Consequently, the Voronoi loss function between G_n and G^* becomes

$$\begin{aligned}
\mathcal{L}_{2n} := \mathcal{L}_2(G_n, G_*) &= \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\Big|\sum_{(h',k')\in V_{h,k}}\omega^n_{h',k'} - \omega_{h,k}\Big| \\
&+ \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\sum_{\substack{(h',k')\in V_{h,k},\\ |V_{h,k}|=1}}\omega^n_{h',k'}\Big(\sum_{i=1}^{N^*}\|\Delta M_{h'h,i}\| + \|\Delta a_{h'h,i,k'k}\|\Big) \\
&+ \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\sum_{\substack{(h',k')\in V_{h,k},\\ |V_{h,k}|>1}}\omega^n_{h',k'}\Big(\sum_{i=1}^{N^*}\|\Delta M_{h'h,i}\|^2 + \|\Delta a_{h'h,i,k'k}\|^2\Big),
\end{aligned}$$

where $\Delta M_{h'h,i} := M^n_{h',i} - M^*_{h,i}$ and $\Delta a_{h'h,i,k'k} := a^n_{h',i,k'} - a^*_{h,i,k}$ for any $i \in [N^*]$. We now decompose the following difference:

$$\begin{aligned}
\mathcal{D}_n(x) &:= g_{G_n}(x) - g_{G^*}(x) \\
&= \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\sum_{(h',k')\in V_{h,k}}\omega^n_{h',k'}\overline{f}^{h',k'}_{G_n}(x) - \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\omega^*_{h,k}\overline{f}^{h,k}_{G^*}(x) \\
&= \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\sum_{(h',k')\in V_{h,k}}\omega^n_{h',k'}\Big[\overline{f}^{h',k'}_{G_n}(x) - \overline{f}^{h,k}_{G^*}(x)\Big] \\
&\quad + \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\left(\sum_{(h',k')\in V_{h,k}}\omega^n_{h',k'} - \omega^*_{h,k}\right)\overline{f}^{h,k}_{G^*}(x) \\
&= \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\sum_{(h',k')\in V_{h,k}}\omega^n_{h',k'}\mathcal{D}^{h'h,k'k}_n + \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\left(\sum_{(h',k')\in V_{h,k}}\omega^n_{h',k'} - \omega^*_{h,k}\right)\overline{f}^{h,k}_{G^*}(x),
\end{aligned} \tag{30}$$

where $\mathcal{D}^{h'h,k'k}_n(x) := \overline{f}^{h',k'}_{G_n}(x) - \overline{f}^{h,k}_{G^*}(x)$. We define

$$\begin{aligned}
E^{h'}_n(x) &= \sum_{j=1}^{N^*}\exp(x^\top M^n_{h',j}x) \\
E^h_*(x) &= \sum_{j=1}^{N^*}\exp(x^\top M^*_{h,j}x) \\
\overline{u}(x; M, a) &= \exp(x^\top Mx)\cdot\varphi(a^\top x) \\
\overline{v}^{h',k'}_n(x; M) &= \exp(x^\top Mx)\cdot\overline{f}^{h',k'}_{G_n}(x),
\end{aligned}$$

for all $h, h' \in [H^*]$ and $k' \in [K]$. Each term $\mathcal{D}_n^{h'h,k'k}(x)$ can be decomposed as follows:

$$\begin{aligned}
\mathcal{D}_n^{h'h,k'k}(x) &= \sum_{i=1}^{N^*} \frac{1}{E_n^{h'}(x)} \overline{u}(x; M_{h',i}^n, a_{h',i,k'}^n) - \sum_{i=1}^{N^*} \frac{1}{E_*^h(x)} \overline{u}(x; M_{h,i}^*, a_{h,i,k}^*) \\
&= \frac{1}{E_*^h(x)} \sum_{i=1}^{N^*} \overline{u}(x; M_{h',i}^n, a_{h',i,k'}^n) - \frac{1}{E_*^h(x)} \sum_{i=1}^{N^*} \overline{u}(x; M_{h,i}^*, a_{h,i,k}^*) \\
&\quad - \left[\frac{1}{E_*^h(x)} \sum_{i=1}^{N^*} \overline{v}_n^{h',k'}(x; M_{h',i}^n) - \frac{1}{E_*^h(x)} \sum_{i=1}^{N^*} \overline{v}_n^{h',k'}(x; M_{h,i}^*) \right] \\
&:= A_n^{h'h,k'k}(x) - B_n^{h'h,k'k}(x)
\end{aligned}$$

We now expand each term $A_n^{h'h,k'k}(x)$ and $B_n^{h'h,k'k}(x)$ via a second-order Taylor expansion. Specifically,

$$\begin{aligned}
&A_n^{h'h,k'k}(x) \\
&= \frac{1}{E_*^h(x)} \sum_{i=1}^{N^*} \overline{u}(x; M_{h',i}^n, a_{h',i,k'}^n) - \frac{1}{E_*^h(x)} \sum_{i=1}^{N^*} \overline{u}(x; M_{h,i}^*, a_{h,i,k}^*) \\
&= \frac{1}{E_*^h(x)} \sum_{i=1}^{N^*} \left(\overline{u}(x; M_{h',i}^n, a_{h',i,k'}^n) - \overline{u}(x; M_{h,i}^*, a_{h,i,k}^*) \right) \\
&= \frac{1}{E_*^h(x)} \left(\sum_{i=1}^{N^*} \sum_{1 \le |t| \le 2} \frac{1}{t!} \left(\Delta M_{h'h,i}\right)^{t_1} \left(\Delta a_{h'h,i,k'k}\right)^{t_2} \frac{\partial^{|t_1|+|t_2|} \overline{u}}{\partial M^{t_1} \partial a^{t_2}}(x; M_{h,i}^*, a_{h,i,k}^*) \right) + \mathcal{R}_{n,1}^{h'h,k'k}(x),
\end{aligned}$$

where $t = (t_1, t_2) \in \mathbb{N}^{\overline{d} \times \overline{d}} \times \mathbb{N}^{\overline{d}}$ and $\mathcal{R}_{n,1}^{h'h,k'k}(x)$ are Taylor remainders such that $\mathcal{R}_{n,1}^{h'h,k'k}(x)/\mathcal{L}_{2n} \to 0$ when $n \to \infty$. Next, we denote

$$A_n(x) := \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \sum_{(h',k') \in V_{h,k}} \omega_{h',k'}^n A_n^{h'h,k'k}(x)$$

$$B_n(x) := \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \sum_{(h',k') \in V_{h,k}} \omega_{h',k'}^n B_n^{h'h,k'k}(x)$$

Recall that $\mathcal{L}_{2n}$ is given by summing first-order terms $|\Delta M_{h'h,i}|$ and $|\Delta a_{h'h,i,k'k}|$ over singleton Voronoi cells ($|V_{h,k}| = 1$), and second-order terms $|\Delta M_{h'h,i}|^2$ and $|\Delta a_{h'h,i,k'k}|^2$ over cells with $|V_{h,k}| > 1$. Therefore, A_n can be approximated by retaining only the dominant contributions:

$$\begin{aligned}
A_n(x) = \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \frac{1}{E_*^h(x)} &\sum_{\substack{(h',k') \in V_{h,k} \\ |V_{h,k}|=1}} \omega_{h',k'}^n \sum_{i=1}^{N^*} \sum_{|t|=1} \frac{1}{t!} \\
&\times \left(\Delta M_{h'h,i}\right)^{t_1} \left(\Delta a_{h'h,i,k'k}\right)^{t_2} \frac{\partial^{|t_1|+|t_2|} \overline{u}}{\partial M^{t_1} \partial a^{t_2}}(x; M_{h,i}^*, a_{h,i,k}^*) \\
+ \sum_{\substack{(h',k') \in V_{h,k} \\ |V_{h,k}|>1}} &\omega_{h',k'}^n \sum_{i=1}^{N^*} \sum_{1 \le |t| \le 2} \frac{1}{t!} \\
&\times \left(\Delta M_{h'h,i}\right)^{t_1} \left(\Delta a_{h'h,i,k'k}\right)^{t_2} \frac{\partial^{|t_1|+|t_2|} \overline{u}}{\partial M^{t_1} \partial a^{t_2}}(x; M_{h,i}^*, a_{h,i,k}^*) + \mathcal{R}_{n,2}(x)
\end{aligned}$$

Similarly, decomposing $B_n(x)$ using the same reasoning yields

$$B_n(x) = \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \frac{1}{E_*^h(x)} \sum_{\substack{(h',k') \in V_{h,k} \\ |V_{h,k}|=1}} \omega_{h',k'}^n \sum_{i=1}^{N^*} \sum_{|r|=1} \frac{1}{r!} \left(\Delta M_{h'h,i}\right)^r \frac{\partial^{|r|} \overline{v}_n^{h',k'}}{\partial M^r}(x; M_{h,i}^*)$$

$$+\sum_{\substack{(h',k')\in V_{h,k}\\ |V_{h,k}|>1}}\omega_{h',k'}^{n}\sum_{i=1}^{N^*}\sum_{1\le|r|\le 2}\frac{1}{r!}\left(\Delta M_{h'h,i}\right)^{r}\frac{\partial^{|r|}\overline{v}_{n}^{h',k'}}{\partial M^{r}}(x;M_{h,i}^{*})+\mathcal{R}_{n,3}(x)$$

where $r\in\mathbb{N}^{\overline{d}\times\overline{d}}$ and $\mathcal{R}_{n,3}(x)$ is Taylor remainder such that $\mathcal{R}_{n,3}(x)/\mathcal{L}_{2n}\to 0$ as $n\to\infty$.

Now, we define $\mathcal{J}_t:=\{(t_1,t_2)\in\mathbb{N}^{\overline{d}\times\overline{d}}\times\mathbb{N}^{\overline{d}}:|t_1|+|t_2|=|t|\}$. Combining the preceding results, the function $\mathcal{D}_n^{h'h,k'k}(x)$ can be represented as

$$\begin{aligned}
&\mathcal{D}_n(x)\\
&=\sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\frac{1}{E_*^h(x)}\sum_{i=1}^{N^*}\sum_{|t|=1}^{1+\mathbf{1}_{\{|V_{h,k}|>1\}}}\sum_{(t_1,t_2)\in\mathcal{J}_t}\left(\frac{1}{t!}\sum_{(h',k')\in V_{h,k}}\omega_{h',k'}^{n}\left(\Delta M_{h'h,i}\right)^{t_1}\left(\Delta a_{h'h,i,k'k}\right)^{t_2}\right)\\
&\quad\times\frac{\partial^{|t_1|+|t_2|}\overline{u}}{\partial M^{t_1}\partial a^{t_2}}(x;M_{h,i}^{*},a_{h,i,k}^{*})\\
&+\sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\frac{1}{E_*^h(x)}\sum_{i=1}^{N^*}\sum_{|r|=1}^{1+\mathbf{1}_{\{|V_{h,k}|>1\}}}\frac{1}{r!}\sum_{(h',k')\in V_{h,k}}\omega_{h',k'}^{n}\left(\Delta M_{h'h,i}\right)^{r}\frac{\partial^{|r|}\overline{v}_{n}^{h',k'}}{\partial M^{r}}(x;M_{h,i}^{*})\\
&+\sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\left(\sum_{(h',k')\in V_{h,k}}\omega_{h',k'}^{n}-\omega_{h,k}^{*}\right)\overline{f}_{G^*}^{h,k}(x)+\mathcal{R}_{n,4}(x),
\end{aligned}$$

where $\mathcal{R}_{n,4}(x)$ is Taylor remainder such that $\mathcal{R}_{n,4}(x)/\mathcal{L}_{2n}\to 0$ as $n\to\infty$ and we denote

$$\begin{aligned}
\overline{Y}_{t_1,t_2}^{h,k}(i)&=\frac{1}{t!}\sum_{(h',k')\in V_{h,k}}\omega_{h',k'}^{n}\left(\Delta M_{h'h,i}\right)^{t_1}\left(\Delta a_{h'h,i,k'k}\right)^{t_2},\\
\overline{Z}_{r}^{h'h,k'}(i)&=\frac{1}{r!}\omega_{h',k'}^{n}\left(\Delta M_{h'h,i}\right)^{r}
\end{aligned}$$

Step 2 (Non-vanishing coefficients).
In this step, we will show that at least one coefficient in $\{\overline{Y}_{t_1,t_2}^{h,k}(i)/\mathcal{L}_{2n};\left(\sum_{(h',k')\in V_{h,k}}\omega_{h',k'}^{n}-\omega_{h,k}^{*}\right)/\mathcal{L}_{2n}\}$ does not go to 0 as $n\to\infty$. Suppose, for contradiction, that every such coefficient go to 0. Summing over the coefficients of $\overline{f}_{G_*}^{h,k}(x)$ for all $1\le h\le H^*,1\le k\le K^*$, we obtain

$$\frac{1}{\mathcal{L}_{2n}}\sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\Big|\sum_{(h',k')\in V_{h,k}}\omega_{h',k'}^{n}-\omega_{h,k}^{*}\Big|\to 0. \tag{31}$$

For index $(h,k)\in[H^*]\times[K^*]$ such that $|V_{h,k}|=1$, we take $(t_1,t_2)=(e_{uv},\mathbf{0}_{\overline{d}})$ and $(t_1,t_2)=(\mathbf{0}_{\overline{d}\times\overline{d}},e_u)$, where e_{uv} denotes the canonical basis matrix in $\mathbb{R}^{\overline{d}\times\overline{d}}$ with a 1 in the (u,v)-th entry and 0 elsewhere. Summing the limits $\overline{Y}_{t_1,t_2}^{h,k}(i)/\mathcal{L}_2\to 0$ for all $i\in[N^*]$, we obtain

$$\frac{1}{\mathcal{L}_2}\sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\sum_{\substack{(h',k')\in V_{h,k},\\ |V_{h,k}|=1}}\omega_{h',k'}^{n}\Big(\sum_{i=1}^{N^*}\|\Delta M_{h'h,i}\|+\|\Delta a_{h'h,i,k'k}\|\Big)\to 0 \tag{32}$$

For index $(h,k)\in[H^*]\times[K^*]$ such that $|V_{k|h}|>1$, set $(t_1,t_2)=(2e_{uv},\mathbf{0}_{\overline{d}})$ and $(t_1,t_2)=(\mathbf{0}_{\overline{d}\times\overline{d}},2e_u)$. Taking the summation with respect to the limits $\overline{Y}_{t_1,t_2}^{h,k}(i)/\mathcal{L}_2\to 0$ for all $i\in[N^*]$, we have

$$\frac{1}{\mathcal{L}_2}\sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\sum_{\substack{(h',k')\in V_{h,k},\\ |V_{h,k}|>1}}\omega_{h',k'}^{n}\Big(\sum_{i=1}^{N^*}\|\Delta M_{h'h,i}\|^2+\|\Delta a_{h'h,i,k'k}\|^2\Big)\to 0 \tag{33}$$

Combining the results in Equations (31), (32), (33) yields $1 = \dfrac{\mathcal{L}_2}{\mathcal{L}_2} \to 0$, which is a contradiction. Consequently, at least one coefficient of the linearly independent functions $\mathcal{D}_n(x)/\mathcal{L}_2$ does not vanish as $n \to \infty$.

Step 3 (Application of the Fatou's lemma). Let $\overline{m}_n$ be the maximum of the absolute values among the coefficients of the linearly independent functions in $\mathcal{D}_n(x)/\mathcal{L}_{2n}$. Since at least one of these coefficients does not vanish, we have $1/\overline{m}_n \not\to \infty$ as $n \to \infty$. Applying Fatou's lemma, we obtain

$$0 = \lim_{n\to\infty} \frac{\|g_{G_n} - g_{G^*}\|_{L^2(\mu)}}{\overline{m}_n \mathcal{L}_{2n}} \geq \int \liminf_{n\to\infty} \frac{|g_{G_n} - g_{G^*}|}{\overline{m}_n \mathcal{L}_{2n}} d\mu(x) \geq 0.$$

As a result, we achieve that

$$\liminf_{n\to\infty} \frac{|g_{G_n} - g_{G^*}|}{\overline{m}_n \mathcal{L}_{2n}} = 0.$$

When $n \to \infty$, we denote

$$\begin{aligned}
\frac{\sum_{(h',k')\in V_{h,k}} \omega^n_{h',k'} - \omega^*_{h,k}}{\overline{m}_n \mathcal{L}_{2n}} &\to \overline{\lambda}^{h,k}_{\omega}; \\
\frac{\overline{Y}^{h,k}_{t_1,t_2}(i)}{\overline{m}_n \mathcal{L}_{2n}} &\to \overline{\lambda}^{h,k}_{y,t_1,t_2,i}; \\
\frac{\overline{Z}^{h'h,k'}_{r}(i)}{\overline{m}_n \mathcal{L}_{2n}} &\to \overline{\lambda}^{h'h,k'}_{z,r,i}; \\
\sum_{(h',k')\in V_{h,k}} \overline{\lambda}^{h'h,k'}_{z,r,i} &= \overline{\lambda}^{h}_{z,r,i}
\end{aligned}$$

Moreover, since $\sum_{(h',k')\in V_{h,k}} \omega^n_{h',k'} \to \omega^*_{h,k}$ and $\overline{f}^{h',k'}_{G_n}(x) \to \overline{f}^{h,k}_{G^*}(x)$ as $n \to \infty$ for all $(h',k') \in V_{h,k}$, we have

$$\overline{v}^{h',k'}_{n}(x; M) \to \exp(x^\top M x) \cdot \overline{f}^{h,k}_{G^*}(x) := \overline{v}^{h,k}_{*}(x; M)$$

Since $\overline{f}^{h,k}_{G^*}(x) = \dfrac{1}{E^h_*(x)} \sum_{i=1}^{N^*} \overline{u}(x; M^*_{h,i}, a^*_{h,i,k})$, the limit $\liminf_{n\to\infty} \dfrac{|g_{G_n} - g_{G^*}|}{\overline{m}_n \mathcal{L}_2} = 0$ can be expressed as

$$\begin{aligned}
&\frac{1}{E^h_*(x)} \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \sum_{i=1}^{N^*} \sum_{|t|=1}^{1+\mathbf{1}_{\{|V_{h,k}|>1\}}} \sum_{(t_1,t_2)\in\mathcal{J}_t} \overline{\lambda}^{h,k}_{y,t_1,t_2,i} \frac{\partial^{|t_1|+|t_2|}\overline{u}}{\partial M^{t_1} \partial a^{t_2}}(x; M^*_{h,i}, a^*_{h,i,k}) \\
&+ \frac{1}{E^h_*(x)} \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \sum_{i=1}^{N^*} \sum_{|r|=1}^{1+\mathbf{1}_{\{|V_{h,k}|>1\}}} \overline{\lambda}^{h}_{z,r,i} \frac{\partial^{|r|}\overline{v}^{h,k}_{*}}{\partial M^r}(x; M^*_{h,i}) \\
&+ \frac{1}{E^h_*(x)} \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \overline{\lambda}^{h,k}_{\omega} \sum_{i=1}^{N^*} \overline{u}(x; M^*_{h,i}, a^*_{h,i,k}) \\
&= 0
\end{aligned}$$

for almost every x. It is important to note that $\overline{\lambda}^{h,k}_{y,t_1,t_2,N^*} = \overline{\lambda}^{h,k}_{z,r,N^*} = 0$, since $M^n_{h,N^*} = M^*_{h,N^*} = 0$ (due to assumption A.3). Hence, the left-hand side of the above equation can be expressed as a linear combination of the following functions:

$$\left\{ \frac{\partial^{|t_1|+|t_2|}\overline{u}}{\partial M^{t_1} \partial a^{t_2}}(x; M^*_{h,i}, a^*_{h,i,k}), \frac{\partial^{|r|}\overline{v}^{h,k}_{*}}{\partial M^r}(x; M^*_{h,i}) \right\}_{0\leq|t_1|+|t_2|\leq 2, 1\leq|r|\leq 2, (h,k,i)\in[H^*]\times[K^*]\times[N^*-1]}$$

Since the function $\varphi(\cdot)$ is type-1 strongly identifiable, the set above is linearly independent for almost every x. Hence, all coefficients $\overline{\lambda}^{h,k}_{y,t_1,t_2,i}$, $\overline{\lambda}^{h}_{z,r,i}$, and $\overline{\lambda}^{h,k}_{\omega}$ must be zero. This yields a contradiction, establishing the desired result.

Proof of the global part (Equation (29)**).**

Assume, by contradiction, that the equation (29) is not true. Then, there exists a sequence $G'_n \in \mathcal{G}_{H^*,K,N^*}(\Theta)$ such that $\mathcal{L}_2(G'_n, G^*) > \varepsilon$ and $\|g_{G'_n} - g_{G^*}\|_{L^2(\mu)} \to 0$, as $n \to \infty$. Recall that Θ is a compact set, therefore, we can replace the sequence G'_n by one of its subsequences that converges to a mixing measure $G' \in \mathcal{G}_{H^*,K,N^*}(\Theta)$. Since $\mathcal{L}_2(G'_n, G^*) > \varepsilon$, we deduce that $\mathcal{L}_2(G', G^*) > \varepsilon$.

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

$$0 = \lim_{n\to\infty} \|g_{G'_n} - g_{G^*}\|^2_{L^2(\mu)} \geq \int \liminf_{n\to\infty} \left|g_{G'_n}(x) - g_{G^*}(x)\right|^2 \mathrm{d}\mu(x) \tag{34}$$

Consequently, we obtain that $g_{G'}(x) = g_{G^*}(x)$ for almost every x. Then, we have

$$\begin{aligned}
&\sum_{h=1}^{H^*}\sum_{k=1}^{K}\omega'_{h,k}\sum_{i=1}^{N^*}\frac{\exp(x^\top M'_{h,i}x)}{\sum_{j=1}^{N^*}\exp(x^\top M'_{h,j}x)}\cdot\varphi\left((a'_{h,i,k})^\top x\right)\\
&=\sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\omega^*_{h,k}\sum_{i=1}^{N^*}\frac{\exp(x^\top M^*_{h,i}x)}{\sum_{j=1}^{N^*}\exp(x^\top M^*_{h,j}x)}\cdot\varphi\left((a^*_{h,i,k})^\top x\right)
\end{aligned}$$

Since the function $\varphi(\cdot)$ is type-1 strong identifiable, the set of functions $\varphi(a^\top x)$ is linearly independent with distinct parameters a. Therefore, we deduce that for each $i \in [N^*]$ and $(h,k) \in [H^*]\times[K^*]$, there exists a set $V_{i,h,k} = \left\{(i_1, h_1, k_1) \in [N^*]\times[H^*]\times[K] : \varphi((a'_{h_1,i_1,k_1})^\top x) = \varphi((a^*_{h,i,k})^\top x)\right\}$. Without loss of generality, we assume that $\varphi((a'_{h_1,i,k_1})^\top x) = \varphi((a^*_{h,i,k})^\top x)$ for all $i \in [N^*]$ and $(h_1,k_1) \in V_{h,k}$ for all $(h,k) \in [H^*]\times[K^*]$. Consequently, we have $a'_{h_1,i,k_1} = a^*_{h,i,k}$ (since the function $\varphi(\cdot)$ is injective) and

$$\sum_{(h_1,k_1)\in V_{h,k}}\omega'_{h_1,k_1}\frac{\exp(x^\top M'_{h_1,i}x)}{\sum_{j=1}^{N^*}\exp(x^\top M'_{h_1,j}x)} = \omega^*_{h,k}\frac{\exp(x^\top M^*_{h,i}x)}{\sum_{j=1}^{N^*}\exp(x^\top M^*_{h,j}x)},$$

for all $(i,h,k) \in [N^*]\times[H^*]\times[K^*]$. By summing the above equation over $i \in [N^*]$, we deduce that

$$\sum_{(h_1,k_1)\in V_{h,k}}\omega'_{h_1,k_1} = \omega^*_{h,k}, \forall\,(h,k) \in [H^*]\times[K^*]. \tag{35}$$

Consequently, we have $\dfrac{\exp(x^\top M'_{h_1,i}x)}{\sum_{j=1}^{N^*}\exp(x^\top M'_{h_1,j}x)} = \dfrac{\exp(x^\top M^*_{h,i}x)}{\sum_{j=1}^{N^*}\exp(x^\top M^*_{h,j}x)}, \forall\,(h_1,k_1) \in V_{h,k}$. Since the set of matrices $\{M_{h,i}\}_{i=1}^{N^*}$ is invariant to translations, we deduce that

$$M'_{h_1,i} = M^*_{h,i}, \forall\, i \in [N^*], (h_1,k_1) \in V_{h,k}. \tag{36}$$

Combining (35) with (36), we obtain that $\mathcal{L}_2(G', G^*) = 0$, which yields a contradiction. This completes the proof.

A.3 Proof of Theorem 3

Our main goal is to demonstrate that

$$\inf_{G\in\mathcal{G}_{H^*,K,N^*}(\Theta)} \|g_G - g_{G^*}\|_{L^2(\mu)}/\mathcal{L}_2(G, G^*) > 0 \tag{37}$$

By following the same line of reasoning as in Theorem 2, we decompose the inequality (37) into a global and a local component. The local part is to establish equation 37 when $\mathcal{L}_2(G, G_*)$ is small enough as follows

$$\lim_{\varepsilon\to 0}\inf_{G\in\mathcal{G}_{H^*,K,N^*}(\Theta):\mathcal{L}_2(G,G^*)\leq\varepsilon} \|g_G - g_{G^*}\|_{L^2(\mu)}/\mathcal{L}_2(G, G^*) > 0. \tag{38}$$

By contrast, the global part of the proof focuses on the behavior of this property in the regime where $\mathcal{L}_2(G, G_*)$ becomes sufficiently large.

$$\inf_{G\in\mathcal{G}_{H^*,K,N^*}(\Theta):\mathcal{L}_2(G,G^*)>\varepsilon} \|g_G - g_{G^*}\|_{L^2(\mu)}/\mathcal{L}_2(G, G^*) > 0. \tag{39}$$

Proof of the local part (Equation (38)**).**

Assume that this local inequality does not hold. Consequently, we have a sequence of mixing measures $(G_n) \in \mathcal{G}_{H^*,K,N^*}(\Theta)$ such that $\mathcal{L}_2(G_n, G^*) \to 0$ and $\|g_{G_n} - g_{G^*}\|_{L^2(\mu)}/\mathcal{L}_2(G_n, G^*) \to 0$ when $n \to \infty$.

Step 1 (Decomposing the discrepancy between regression functions). First of all, we recall that

$$f^{h',k'}_{G_n}(x) = \sum_{i=1}^{N^*} \frac{\exp(x^\top M^n_{h',i}x)}{\sum_{j=1}^{N^*}\exp(x^\top M^n_{h',j}x)} \cdot (a^n_{h',i,k'})^\top x;$$

$$f^{h,k}_{G^*}(x) = \sum_{i=1}^{N^*} \frac{\exp(x^\top M^*_{h,i}x)}{\sum_{j=1}^{N^*}\exp(x^\top M^*_{h,j}x)} \cdot (a^*_{h,i,k})^\top x.$$

Now, we have

$$g_{G_n}(x) = \sum_{h'=1}^{H^*}\sum_{k'=1}^{K} \omega^n_{h',k'}\varphi\left(f^{h',k'}_{G_n}(x)\right)$$

$$g_{G^*}(x) = \sum_{h=1}^{H^*}\sum_{k=1}^{K^*} \omega^*_{h,k}\varphi\left(f^{h,k}_{G^*}(x)\right)$$

As a result, the Voronoi loss function between G_n and G^* becomes

$$\begin{aligned}
\mathcal{L}_{2n} := \mathcal{L}_2(G_n, G_*) = &\sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\left|\sum_{(h',k')\in V_{h,k}} \omega^n_{h',k'} - \omega_{h,k}\right| \\
&+ \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\sum_{\substack{(h',k')\in V_{h,k},\\ |V_{h,k}|=1}} \omega^n_{h',k'}\Big(\sum_{i=1}^{N^*}\|\Delta M_{h'h,i}\| + \|\Delta a_{h'h,i,k'k}\|\Big) \\
&+ \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\sum_{\substack{(h',k')\in V_{h,k},\\ |V_{h,k}|>1}} \omega^n_{h',k'}\Big(\sum_{i=1}^{N^*}\|\Delta M_{h'h,i}\|^2 + \|\Delta a_{h'h,i,k'k}\|^2\Big),
\end{aligned}$$

where $\Delta M_{h'h,i} = M^n_{h',i} - M^*_{h,i}$ and $\Delta a_{h'h,i,k'k} = a^n_{h',i,k'} - a^*_{h,i,k}$ for any $i \in [N^*]$. Next, we decompose the following difference as follows

$$\begin{aligned}
\mathcal{D}_n(x) &:= g_{G_n}(x) - g_{G^*}(x) \\
&= \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\sum_{(h',k')\in V_{h,k}} \omega^n_{h',k'}\varphi\left(f^{h',k'}_{G_n}(x)\right) - \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\omega^*_{h,k}\varphi\left(f^{h,k}_{G^*}(x)\right) \\
&= \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\sum_{(h',k')\in V_{h,k}} \omega^n_{h',k'}\left[\varphi\left(f^{h',k'}_{G_n}(x)\right) - \varphi\left(f^{h,k}_{G^*}(x)\right)\right] \\
&\quad + \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\left(\sum_{(h',k')\in V_{h,k}} \omega^n_{h',k'} - \omega^*_{h,k}\right)\varphi\left(f^{h,k}_{G^*}(x)\right) \\
&= \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\sum_{(h',k')\in V_{h,k}} \omega^n_{h',k'}\mathcal{D}^{h'h,k'k}_n + \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\left(\sum_{(h',k')\in V_{h,k}} \omega^n_{h',k'} - \omega^*_{h,k}\right)\varphi\left(f^{h,k}_{G^*}(x)\right),
\end{aligned} \tag{40}$$

where $\mathcal{D}^{h'h,k'k}_n(x) := \varphi\left(f^{h',k'}_{G_n}(x)\right) - \varphi\left(f^{h,k}_{G^*}(x)\right)$. Next, we denote

$$E^{h'}_n(x) = \sum_{j=1}^{N^*}\exp(x^\top M^n_{h',j}x);$$

$$E_*^h(x) = \sum_{j=1}^{N^*} \exp(x^\top M_{h,j}^* x);$$
$$u(x; M, a) = \exp(x^\top M x) \cdot a^\top x;$$
$$v_n^{h',k'}(x; M) = \exp(x^\top M x) \cdot f_{G_n}^{h',k'}(x),$$

for all $h, h' \in [H^*]$ and $k' \in [K]$. We now decompose each term $\mathcal{D}_n^{h'h,k'k}(x)$ as follows

$$\begin{aligned}
\mathcal{D}_n^{h'h,k'k}(x) &= \varphi\left(\sum_{i=1}^{N^*} \frac{1}{E_n^{h'}(x)} u(x; M_{h',i}^n, a_{h',i,k'}^n)\right) - \varphi\left(\sum_{i=1}^{N^*} \frac{1}{E_*^h(x)} u(x; M_{h,i}^*, a_{h,i,k}^*)\right) \\
&= \varphi\left(\frac{1}{E_*^h(x)} \sum_{i=1}^{N^*} u(x; M_{h',i}^n, a_{h',i,k'}^n)\right) - \varphi\left(\frac{1}{E_*^h(x)} \sum_{i=1}^{N^*} u(x; M_{h,i}^*, a_{h,i,k}^*)\right) \\
&\quad - \left[\varphi\left(\frac{1}{E_*^h(x)} \sum_{i=1}^{N^*} v_n^{h',k'}(x; M_{h',i}^n)\right) - \varphi\left(\frac{1}{E_*^h(x)} \sum_{i=1}^{N^*} v_n^{h',k'}(x; M_{h,i}^*)\right)\right] \\
&:= A_n^{h'h,k'k}(x) - B_n^{h'h,k'k}(x)
\end{aligned}$$

We now decompose each term $A_n^{h'h,k'k}(x)$ and $B_n^{h'h,k'k}(x)$ using Taylor expansion. In particular, by means of second-order Taylor expansion, we have

$$\begin{aligned}
&A_n^{h'h,k'k}(x) \\
&= \varphi\left(\frac{1}{E_*^h(x)} \sum_{i=1}^{N^*} u(x; M_{h',i}^n, a_{h',i,k'}^n)\right) - \varphi\left(\frac{1}{E_*^h(x)} \sum_{i=1}^{N^*} u(x; M_{h,i}^*, a_{h,i,k}^*)\right) \\
&= \varphi'\left(\frac{1}{E_*^h(x)} \sum_{i=1}^{N^*} u(x; M_{h,i}^*, a_{h,i,k}^*)\right) \times \left[\frac{1}{E_*^h(x)} \sum_{i=1}^{N^*} \sum_{1 \le |t| \le 2} \frac{1}{t!}\right. \\
&\qquad \left. \times \left(\Delta M_{h'h,i}\right)^{t_1} \left(\Delta a_{h'h,i,k'k}\right)^{t_2} \frac{\partial^{|t_1|+|t_2|} u}{\partial M^{t_1} \partial a^{t_2}}(x; M_{h,i}^*, a_{h,i,k}^*)\right] \\
&\quad + \frac{1}{2}\varphi''\left(\frac{1}{E_*^h(x)} \sum_{i=1}^{N^*} u(x; M_{h,i}^*, a_{h,i,k}^*)\right) \left[\frac{1}{E_*^h(x)} \sum_{i=1}^{N^*} \sum_{|t|=1} \frac{1}{t!}\right. \\
&\qquad \left. \times \left(\Delta M_{h'h,i}\right)^{t_1} \left(\Delta a_{h'h,i,k'k}\right)^{t_2} \frac{\partial^{|t_1|+|t_2|} u}{\partial M^{t_1} \partial a^{t_2}}(x; M_{h,i}^*, a_{h,i,k}^*)\right]^2 + \mathcal{R}_{n,1}^{h'h,k'k}(x),
\end{aligned}$$

where $t = (t_1, t_2) \in \mathbb{N}^{\overline{d} \times \overline{d}} \times \mathbb{N}^{\overline{d}}$ and $\mathcal{R}_{n,1}^{h'h,k'k}(x)$ are Taylor remainders such that $\mathcal{R}_{n,1}^{h'h,k'k}(x)/\mathcal{L}_{2n} \to 0$ when $n \to \infty$. Recall that the Voronoi loss function $\mathcal{L}_{2n}$ is defined as the sum of first-order terms $\|\Delta M_{h'h,i}\|$ and $\|\Delta a_{h'h,i,k'k}\|$ for singleton Voronoi cells ($|V_{h,k}| = 1$), and second-order terms $\|\Delta M_{h'h,i}\|^2$ and $\|\Delta a_{h'h,i,k'k}\|^2$ for Voronoi cells with $|V_{h,k}| > 1$. Consequently, we can provide a reduced representation of $A_n^{h'h,k'k}$ by focusing on the dominant terms as follows

$$\begin{aligned}
&A_n^{h'h,k'k}(x) \\
&= \sum_{|V_{h,k}|=1} \sum_{i=1}^{N^*} \sum_{|t|=1} \frac{1}{t!} \left(\Delta M_{h'h,i}\right)^{t_1} \left(\Delta a_{h'h,i,k'k}\right)^{t_2} \frac{\partial^{|t_1|+|t_2|} u}{\partial M^{t_1} \partial a^{t_2}}(x; M_{h,i}^*, a_{h,i,k}^*) \varphi_{1n}^{h,k}(x) \\
&\quad + \sum_{|V_{h,k}|>1} \sum_{i=1}^{N^*} \sum_{1 \le |t| \le 2} \frac{1}{t!} \left(\Delta M_{h'h,i}\right)^{t_1} \left(\Delta a_{h'h,i,k'k}\right)^{t_2} \frac{\partial^{|t_1|+|t_2|} u}{\partial M^{t_1} \partial a^{t_2}}(x; M_{h,i}^*, a_{h,i,k}^*) \varphi_{1n}^{h,k}(x) \\
&\quad + \frac{1}{2} \sum_{|V_{h,k}|>1} \left[\sum_{i=1}^{N^*} \sum_{|t|=1} \frac{1}{t!} \left(\Delta M_{h'h,i}\right)^{t_1} \left(\Delta a_{h'h,i,k'k}\right)^{t_2} \frac{\partial^{|t_1|+|t_2|} u}{\partial M^{t_1} \partial a^{t_2}}(x; M_{h,i}^*, a_{h,i,k}^*)\right]^2 \varphi_{2n}^{h,k}(x)
\end{aligned}$$

$$+ \mathcal{R}_{n,2}^{h'h,k'k}(x),$$

where $\mathcal{R}_{n,2}^{h'h,k'k}(x)$ is the remainder such that $\mathcal{R}_{n,2}^{h'h,k'k}(x)/\mathcal{L}_{2n} \to 0$ when $n \to \infty$ and

$$\varphi_{1n}^{h,k}(x) := \frac{1}{E_*^h(x)}\varphi'\left(\frac{1}{E_*^h(x)}\sum_{i=1}^{N^*} u(x; M_{h,i}^*, a_{h,i,k}^*)\right);$$
$$\varphi_{2n}^{h,k}(x) := \frac{1}{E_*^h(x)^2}\varphi''\left(\frac{1}{E_*^h(x)}\sum_{i=1}^{N^*} u(x; M_{h,i}^*, a_{h,i,k}^*)\right)$$

Next, we represent $A_n^{h'h,k'k}(x)$ as follows

$$\begin{aligned}
&A_n^{h'h,k'k}(x)\\
&= \sum_{|V_{h,k}|=1}\sum_{i=1}^{N^*}\sum_{|t|=1}\frac{1}{t!}\left(\Delta M_{h'h,i}\right)^{t_1}\left(\Delta a_{h'h,i,k'k}\right)^{t_2}\frac{\partial^{|t_1|+|t_2|}u}{\partial M^{t_1}\partial a^{t_2}}(x; M_{h,i}^*, a_{h,i,k}^*)\varphi_{1n}^{h,k}(x)\\
&\quad+\sum_{|V_{h,k}|>1}\sum_{i=1}^{N^*}\sum_{1\le|t|\le 2}\frac{1}{t!}\left(\Delta M_{h'h,i}\right)^{t_1}\left(\Delta a_{h'h,i,k'k}\right)^{t_2}\frac{\partial^{|t_1|+|t_2|}u}{\partial M^{t_1}\partial a^{t_2}}(x; M_{h,i}^*, a_{h,i,k}^*)\varphi_{1n}^{h,k}(x)\\
&\quad+\frac{1}{2}\sum_{|V_{h,k}|>1}\sum_{i_1=1}^{N^*}\sum_{i_2=1}^{N^*}\sum_{\substack{|t_1|=1;\\|t_2|=1}}\left(\Delta M_{h'h,i_1}\right)^{t_{1,1}}\left(\Delta M_{h'h,i_2}\right)^{t_{2,1}}\left(\Delta a_{h'h,i_1,k'k}\right)^{t_{1,2}}\left(\Delta a_{h'h,i_2,k'k}\right)^{t_{2,2}}\\
&\qquad\times\frac{\partial^{|t_{1,1}|+|t_{1,2}|}u}{\partial M^{t_{1,1}}\partial a^{t_{1,2}}}(x; M_{h,i_1}^*, a_{h,i_1,k}^*)\cdot\frac{\partial^{|t_{2,1}|+|t_{2,2}|}u}{\partial M^{t_{2,1}}\partial a^{t_{2,2}}}(x; M_{h,i_2}^*, a_{h,i_2,k}^*)\varphi_{2n}^{h,k}(x) + \mathcal{R}_{n,2}^{h'h,k'k}(x).
\end{aligned}$$

Similarly, by applying the same arguments for decomposing $A_n^{h'h,k'k}(x)$, we can express $B_n^{h'h,k'k}(x)$ as

$$\begin{aligned}
&B_n^{h'h,k'k}(x)\\
&= \sum_{|V_{h,k}|=1}\sum_{i=1}^{N^*}\sum_{|r|=1}\frac{1}{r!}\left(\Delta M_{h'h,i}\right)^{r}\frac{\partial^{|r|}v_n^{h',k'}}{\partial M^r}(x; M_{h,i}^*)\varphi_{3n}^{h',k'}(x)\\
&\quad+\sum_{|V_{h,k}|>1}\sum_{i=1}^{N^*}\sum_{1\le|r|\le 2}\frac{1}{r!}\left(\Delta M_{h'h,i}\right)^{r}\frac{\partial^{|r|}v_n^{h',k'}}{\partial M^r}(x; M_{h,i}^*)\varphi_{3n}^{h',k'}(x)\\
&\quad+\frac{1}{2}\sum_{|V_{h,k}|>1}\sum_{i_1=1}^{N^*}\sum_{i_2=1}^{N^*}\sum_{|r_1|=1,|r_2|=1}\left(\Delta M_{h'h,i_1}\right)^{r_1}\left(\Delta M_{h'h,i_2}\right)^{r_2}\\
&\qquad\times\frac{\partial^{|r_1|}v_n^{h',k'}}{\partial M^{r_1}}(x; M_{h,i_1}^*)\frac{\partial^{|r_2|}v_n^{h',k'}}{\partial M^{r_2}}(x; M_{h,i_2}^*)\varphi_{4n}^{h',k'}(x) + \mathcal{R}_{n,3}^{h'h,k'k}(x),
\end{aligned}$$

where $r, r_1, r_2 \in \mathbb{N}^{\overline{d}\times\overline{d}}$ and $\mathcal{R}_{n,3}^{h'h,k'k}(x)$ is Taylor remainder such that $\mathcal{R}_{n,3}^{h'h,k'k}(x)/\mathcal{L}_{2n} \to 0$ as $n \to \infty$ and

$$\varphi_{3n}^{h',k'}(x) := \frac{1}{E_*^h(x)}\varphi'\left(\frac{1}{E_*^h(x)}\sum_{i=1}^{N^*} v_n^{h',k'}(x; M_{h,i}^*)\right)$$
$$\varphi_{4n}^{h',k'}(x) := \frac{1}{E_*^h(x)^2}\varphi''\left(\frac{1}{E_*^h(x)}\sum_{i=1}^{N^*} v_n^{h',k'}(x; M_{h,i}^*)\right)$$

Next, let us denote $\mathcal{J}_t := \{(t_1, t_2) \in \mathbb{N}^{\overline{d}\times\overline{d}} \times \mathbb{N}^{\overline{d}} : |t_1| + |t_2| = |t|\}$. It is worth noting that $\dfrac{\partial^2 u}{\partial a^2} \equiv 0$. Furthermore, we also have $a^\top \cdot \dfrac{\partial u}{\partial a} = u$. Then, putting all the above results together, the function

$\mathcal{D}_n^{h'h,k'k}(x)$ can be represented as

$$
\begin{aligned}
&\mathcal{D}_n^{h'h,k'k}(x)\\
&=\sum_{i=1}^{N^*}\sum_{|t|=0}^{1+\mathbf{1}_{\{|V_{h,k}|>1\}}}\overline{Y}_t^{h'h,k'k}(i)\frac{\partial^{|t|}u}{\partial M^t}(x;M_{h,i}^*,a_{h,i,k}^*)\varphi_{1n}^{h,k}(x)\\
&\quad-\sum_{i=1}^{N^*}\sum_{|r|=1}^{1+\mathbf{1}_{\{|V_{h,k}|>1\}}}\overline{Z}_r^{h'h}(i)\frac{\partial^{|r|}v_n^{h',k'}}{\partial M^r}(x;M_{h,i}^*)\varphi_{3n}^{h',k'}(x)\\
&\quad+\sum_{\substack{|V_{h,k}|>1,\\ i_1,i_2\in[N^*]}}\sum_{0\le|t_1|,|t_2|\le1}\overline{P}_{t_1,t_2}^{h'h,k'k}(i_1,i_2)\\
&\qquad\times\frac{\partial^{|t_1|}u}{\partial M^{t_1}}(x;M_{h,i_1}^*,a_{h,i_1,k}^*)\frac{\partial^{|t_2|}u}{\partial M^{t_2}}(x;M_{h,i_2}^*,a_{h,i_2,k}^*)\varphi_{2n}^{h,k}(x)\\
&\quad-\sum_{\substack{|V_{h,k}|>1,\\ i_1\in[N^*],i_2\in[N^*]}}\sum_{|r_1|=1,|r_2|=1}\overline{Q}_{r_1,r_2}^{h'h}(i_1,i_2)\\
&\qquad\times\frac{\partial^{|r_1|}v_n^{h',k'}}{\partial M^{r_1}}(x;M_{h,i_1}^*)\frac{\partial^{|r_2|}v_n^{h',k'}}{\partial M^{r_2}}(x;M_{h,i_2}^*)\varphi_{4n}^{h',k'}(x)+\mathcal{R}_{n,4}^{h'h,k'k}(x),
\end{aligned}
\tag{41}
$$

where $\mathcal{R}_{n,4}^{h'h,k'k}(x):=\mathcal{R}_{n,2}^{h'h,k'k}(x)-\mathcal{R}_{n,3}^{h'h,k'k}(x)$ and we denote

$$
\overline{Y}_t^{h'h,k'k}(i)=\begin{cases}\sum_{|t_2|=1}\left(\Delta a_{h'h,i,k'k}\right)^{t_2}, & \text{if } |t|=0\\ \left(\Delta M_{h'h,i}\right)^t+\sum_{|t_2|=1}(a_{h,i,k}^*)^{t_2}\cdot\left(\Delta a_{h'h,i,k'k}\right)^{t_2}\cdot\mathbf{1}_{\{|V_{h,k}|>1\}}, & \text{if } |t|=1\\ \left(\Delta M_{h'h,i}\right)^t, & \text{if } |t|=2\end{cases}
$$

$$
\overline{Z}_r^{h'h}(i)=\frac{1}{r!}\left(\Delta M_{h'h,i}\right)^r
$$

$$
\overline{P}_{t_1,t_2}^{h'h,k'k}(i_1,i_2)=\begin{cases}1/2\left[\sum_{|s|=1}(a_{h,i_1,k}^*)^s\cdot\left(\Delta a_{h'h,i_1,k'k}\right)^s\right]\cdot\left[\sum_{|s|=1}(a_{h,i_2,k}^*)^s\cdot\left(\Delta a_{h'h,i_2,k'k}\right)^s\right]\\ \qquad\qquad\text{if } (|t_1|,|t_2|)=(0,0)\\ 1/2\left[\sum_{|s|=1}(a_{h,i_1,k}^*)^s\cdot\left(\Delta a_{h'h,i_1,k'k}\right)^s\right]\cdot\left(\Delta M_{h'h,i_2}\right)^{t_2},\ \text{if } (|t_1|,|t_2|)=(0,1)\\ 1/2\left(\Delta M_{h'h,i_1}\right)^{t_1}\cdot\left[\sum_{|s|=1}(a_{h,i_2,k}^*)^s\cdot\left(\Delta a_{h'h,i_2,k'k}\right)^s\right],\ \text{if } (|t_1|,|t_2|)=(1,0)\\ 1/2\left(\Delta M_{h'h,i_1}\right)^{t_1}\left(\Delta M_{h'h,i_2}\right)^{t_2},\ \text{if } (|t_1|,|t_2|)=(1,1)\end{cases}
$$

$$
\left(\Delta M_{h'h,i_1}\right)^{t_{1,1}}\left(\Delta M_{h'h,i_2}\right)^{t_{2,1}}\left(\Delta a_{h'h,i_1,k'k}\right)^{t_{1,2}}\left(\Delta a_{h'h,i_2,k'k}\right)^{t_{2,2}}
$$

$$
\overline{Q}_{r_1,r_2}^{h'h}(i_1,i_2)=\frac{1}{2}\left(\Delta M_{h'h,i_1}\right)^{r_1}\left(\Delta M_{h'h,i_2}\right)^{r_2}
$$

where e_u denote the u-th canonical basis vector in $\mathbb{R}^{\overline{d}}$ with a 1 in the u-th component and 0 elsewhere.

Step 2 (Non-vanishing coefficients). Therefore, the equations (40) and (41) show that the ratio $\mathcal{D}_n(x)/\mathcal{L}_{2n}(G_n,G^*)$ can be decomposed as a linear combination of the following functions

$$
\begin{aligned}
&\Bigg\{\varphi\left(f_{G_*}^{h,k}(x)\right),\frac{\partial^{|t|}u}{\partial M^t}(x;M_{h,i}^*,a_{h,i,k}^*)\varphi_{1n}^{h,k}(x),\frac{\partial^{|r|}v_n^{h',k'}}{\partial M^r}(x;M_{h,i}^*)\varphi_{3n}^{h',k'}(x)\\
&\frac{\partial^{|t_1|}u}{\partial M^{t_1}}(x;M_{h,i_1}^*,a_{h,i_1,k}^*)\cdot\frac{\partial^{|t_2|}u}{\partial M^{t_2}}(x;M_{h,i_2}^*,a_{h,i_2,k}^*)\varphi_{2n}^{h,k}(x),\\
&\frac{\partial^{|r_1|}v_n^{h',k'}}{\partial M^{r_1}}(x;M_{h,i_1}^*)\frac{\partial^{|r_2|}v_n^{h',k'}}{\partial M^{r_2}}(x;M_{h,i_2}^*)\varphi_{4n}^{h',k'}(x)\Bigg\},
\end{aligned}
$$

for any $h\in[H^*],k\in[K^*],(h',k')\in V_{h,k},0\le|t|\le2,0\le|t_1|,|t_2|\le1,1\le|r|\le2,|r_1|=1$, and $|r_2|=1$. In this step, we will prove that at least one coefficient of these functions does not

go to 0 as $n \to \infty$. Assume by contrary that all these coefficients of these linear independent functions go to 0. Taking the summation with respect to the coefficient of $\varphi\left(f_{G_*}^{h,k}(x)\right)$ for all $1 \leq h \leq H^*, 1 \leq k \leq K^*$, we have

$$\frac{1}{\mathcal{L}_{2n}} \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} | \sum_{(h',k') \in V_{h,k}} \omega_{h',k'}^n - \omega_{h,k}^* | \to 0. \tag{42}$$

For index $(h,k) \in [H^*] \times [K^*]$ such that $|V_{h,k}| = 1$, taking $t = \mathbf{0}_{\overline{d}}$ and $t = e_{uv}$, in which e_{uv} denotes the canonical basis matrix in $\mathbb{R}^{\overline{d} \times \overline{d}}$ with a 1 in the (u,v)-th entry and 0 elsewhere. Taking the summation with respect to the limits $\overline{Y}_t^{h'h,k'k}(i)/\mathcal{L}_2 \to 0$ for all $i \in [N^*]$, we have

$$\frac{1}{\mathcal{L}_{2n}} \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \sum_{\substack{(h',k') \in V_{h,k}, \\ |V_{h,k}|=1}} \omega_{h',k'}^n \Big(\sum_{i=1}^{N^*} \|\Delta M_{h'h,i}\| + \|\Delta a_{h'h,i,k'k}\| \Big) \to 0 \tag{43}$$

For index $(h,k) \in [H^*] \times [K^*]$ such that $|V_{k|h}| > 1$, by setting $t_1 = t_2 = e_{uv}$ and taking the summation with respect to the limits $\overline{P}_{t_1,t_2}^{h'h,k'k}(i_1,i_1)/\mathcal{L}_2 \to 0$ for all $i_1 \in [N^*]$, we have

$$\frac{1}{\mathcal{L}_{2n}} \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \sum_{\substack{(h',k') \in V_{h,k}, \\ |V_{h,k}|>1}} \omega_{h',k'}^n \sum_{i=1}^{N^*} \|\Delta M_{h'h,i}\|^2 \to 0 \tag{44}$$

Similarly, by setting $t_1 = t_2 = \mathbf{0}_{\overline{d} \times \overline{d}}$ and taking the summation with respect to the limits $\overline{P}_{t_1,t_2}^{h'h,k'k}(i_1,i_1)/\mathcal{L}_2 \to 0$ for all $i_1 \in [N^*]$, we have

$$\frac{1}{\mathcal{L}_{2n}} \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \sum_{\substack{(h',k') \in V_{h,k}, \\ |V_{h,k}|>1}} \omega_{h',k'}^n \sum_{i=1}^{N^*} \|\Delta a_{h'h,i,k'k}\|^2 \to 0 \tag{45}$$

By putting all the results in Equations (42), (43), (44), (45) together, we achieve that $1 = \dfrac{\mathcal{L}_2}{\mathcal{L}_2} \to 0$, which is a contradiction. As a result, at least one of the coefficients of the linearly independent functions $\mathcal{D}_n(x)/\mathcal{L}_2$ does not vanish as $n \to \infty$.

Step 3 (Application of the Fatou's lemma). Denote $\overline{m}_n$ as the maximum of the absolute values of the coefficients of the linear independent functions in $\mathcal{D}_n(x)/\mathcal{L}_{2n}$. Given that at least one of these coefficients does not vanish, we have $1/\overline{m}_n \not\to \infty$ as $n \to \infty$. By invoking Fatou's lemma, we have that

$$0 = \lim_{n \to \infty} \frac{\|g_{G_n} - g_{G^*}\|_{L^2(\mu)}}{\overline{m}_n \mathcal{L}_{2n}} \geq \int \liminf_{n \to \infty} \frac{|g_{G_n} - g_{G^*}|}{\overline{m}_n \mathcal{L}_{2n}} d\mu(x) \geq 0.$$

As a consequence, we achieve that

$$\liminf_{n \to \infty} \frac{|g_{G_n} - g_{G^*}|}{\overline{m}_n \mathcal{L}_{2n}} = 0.$$

When $n \to \infty$, we denote

$$\begin{aligned}
\frac{\sum_{(h',k') \in V_{h,k}} \omega_{h',k'}^n - \omega_{h,k}^*}{\overline{m}_n \mathcal{L}_{2n}} &\to \overline{\lambda}_{\omega}^{h,k}; \\
\frac{\overline{Y}_t^{h'h,k'k}(i)}{\overline{m}_n \mathcal{L}_{2n}} &\to \overline{\lambda}_{y,t,i}^{h'h,k'k}; \\
\frac{\overline{Z}_r^{h'h}(i)}{\overline{m}_n \mathcal{L}_{2n}} &\to \overline{\lambda}_{z,r,i}^{h'h}; \\
\frac{\overline{P}_{t_1,t_2}^{h'h,k'k}(i_1,i_2)}{\overline{m}_n \mathcal{L}_{2n}} &\to \overline{\lambda}_{p,t_1,t_2,i_1,i_2}^{h'h,k'k};
\end{aligned}$$

$$\frac{\overline{Q}^{h'h}_{r_1,r_2}(i_1,i_2)}{\overline{m}_n\mathcal{L}_{2n}} \to \overline{\lambda}^{h'h}_{q,r_1,r_2,i_1,i_2},$$

for any $h \in [H^*], k \in [K^*], (h',k') \in V_{h,k}, 0 \le |t| \le 2, 0 \le |t_1|, |t_2| \le 1, 1 \le |r| \le 2, |r_1| = 1,$ and $|r_2| = 1$. Now, since $\sum_{(h',k')\in V_{h,k}} \omega^n_{h',k'} \to \omega^*_{h,k}$ and $f^{h',k'}_{G_n}(x) \to f^{h,k}_{G^*}(x)$ as $n \to \infty$ for all $(h',k') \in V_{h,k}$, we have

$$\begin{aligned}
v^{h',k'}_n(x;M) &\to \exp(x^\top Mx)\cdot f^{h,k}_{G^*}(x) := v^{h,k}_*(x;M);\\
\varphi^{h,k}_{1n}(x) &\to \frac{1}{E^h_*(x)}\varphi'\left(f^{h,k}_{G^*}(x)\right);\\
\varphi^{h,k}_{2n}(x) &\to \frac{1}{E^h_*(x)^2}\varphi''\left(f^{h,k}_{G^*}(x)\right);\\
\varphi^{h',k'}_{3n}(x) &\to \frac{1}{E^h_*(x)}\varphi'\left(f^{h,k}_{G^*}(x)\right);\\
\varphi^{h',k'}_{4n}(x) &\to \frac{1}{E^h_*(x)^2}\varphi''\left(f^{h,k}_{G^*}(x)\right).
\end{aligned}$$

Given the above notation, the limit $\liminf_{n\to\infty} \frac{|g_{G_n} - g_{G^*}|}{\overline{m}_n\mathcal{L}_2}$ can be expressed as

$$\begin{aligned}
&\sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\overline{\lambda}^{h,k}_{\omega}\varphi\left(f^{h,k}_{G^*}(x)\right) + \sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\sum_{(h',k')\in V_{h,k}}\omega^n_{h',k'}\times\\
&\Bigg[\sum_{i=1}^{N^*}\sum_{|t|=0}^{1+\mathbf{1}_{\{|V_{h,k}|>1\}}}\overline{\lambda}^{h'h,k'k}_{y,t,i}\frac{\partial^{|t|}u}{\partial M^t}(x;M^*_{h,i},a^*_{h,i,k})\frac{1}{E^h_*(x)}\varphi'\left(f^{h,k}_{G^*}(x)\right)\\
&\quad-\sum_{i=1}^{N^*}\sum_{|r|=1}^{1+\mathbf{1}_{\{|V_{h,k}|>1\}}}\overline{\lambda}^{h'h}_{z,r,i}\frac{\partial^{|r|}v^{h,k}_*}{\partial M^r}(x;M^*_{h,i})\frac{1}{E^h_*(x)}\varphi'\left(f^{h,k}_{G^*}(x)\right)\\
&\quad+\sum_{\substack{|V_{h,k}|>1,\\ i_1\in[N^*],i_2\in[N^*]}}\sum_{0\le|t_1|,|t_2|\le 1}\overline{\lambda}^{h'h,k'k}_{p,t_1,t_2,i_1,i_2}\\
&\qquad\times\frac{\partial^{|t_1|}u}{\partial M^{t_1}}(x;M^*_{h,i_1},a^*_{h,i_1,k})\frac{\partial^{|t_2|}u}{\partial M^{t_2}}(x;M^*_{h,i_2},a^*_{h,i_2,k})\frac{1}{E^h_*(x)^2}\varphi''\left(f^{h,k}_{G^*}(x)\right)\\
&\quad-\sum_{\substack{|V_{h,k}|>1,\\ i_1\in[N^*],i_2\in[N^*]}}\sum_{|r_1|=1,|r_2|=1}\overline{\lambda}^{h'h}_{q,r_1,r_2,i_1,i_2}\\
&\qquad\times\frac{\partial^{|r_1|}v^{h,k}_*}{\partial M^{r_1}}(x;M^*_{h,j_1})\frac{\partial^{|r_2|}v^{h,k}_*}{\partial M^{r_2}}(x;M^*_{h,j_2})\frac{1}{E^h_*(x)^2}\varphi''\left(f^{h,k}_{G^*}(x)\right)\Bigg] = 0, \qquad (46)
\end{aligned}$$

for almost every x. Since the function $\varphi(\cdot)$ is type-2 strong identifiable, the set

$$\begin{aligned}
&\Bigg\{\varphi\left(f^{h,k}_{G_*}(x)\right), \frac{\partial^{|t|}u}{\partial M^t}(x;M^*_{h,i},a^*_{h,i,k})\frac{1}{E^h_*(x)}\varphi'\left(f^{h,k}_{G_*}(x)\right),\\
&\frac{\partial^{|r|}v^{h,k}_*}{\partial M^r}(x;M^*_{h,i})\frac{1}{E^h_*(x)}\varphi'\left(f^{h,k}_{G_*}(x)\right),\\
&\frac{\partial^{|t_1|}u}{\partial M^{t_1}}(x;M^*_{h,i_1},a^*_{h,i_1,k})\cdot\frac{\partial^{|t_2|}u}{\partial M^{t_2}}(x;M^*_{h,i_2},a^*_{h,i_2,k})\frac{1}{E^h_*(x)^2}\varphi''\left(f^{h,k}_{G_*}(x)\right),\\
&\frac{\partial^{|r_1|}v^{h,k}_*}{\partial M^{r_1}}(x;M^*_{h,i_1})\frac{\partial^{|r_2|}v^{h,k}_*}{\partial M^{r_2}}(x;M^*_{h,i_2})\frac{1}{E^h_*(x)^2}\varphi''\left(f^{h,k}_{G_*}(x)\right)\Bigg\},
\end{aligned}$$

is linearly independent for any $h \in [H^*], k \in [K^*], 0 \leq |t| \leq 2$, $0 \leq |t_1|, |t_2| \leq 2, |r| \in \{1, 2\}, |r_1| = 1$, and $|r_2| = 1$. Therefore, equation (46) indicates that all the coefficients

$$\left\{\overline{\lambda}_{\omega}^{h,k}, \overline{\lambda}_{y,t,i}^{h'h,k'k}, \overline{\lambda}_{z,r,i}^{h'h}, \overline{\lambda}_{p,t_1,t_2,i_1,i_2}^{hk'}, \overline{\lambda}_{q,r_1,r_2,i_1,i_2}^{h'h}\right\}$$

are 0's, which is a contradiction. Hence, we obtain that

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{G}_{H^*,K,N^*}(\Theta): \mathcal{L}_2(G,G^*) \leq \varepsilon} \|g_G - g_{G^*}\|_{L^2(\mu)} / \mathcal{L}_2(G, G^*) > 0.$$

Proof of the global part (Equation (39)**).**

Following the same approach as in the proof of the global part in Theorem 2 at Appendix A.2, we will demonstrate the identifiability of mixing measure in $\mathcal{G}_{H^*,K,N^*}$. In particular, we prove that the equality $g_G(x) = g_{G^*}(x)$ for almost every x implies that $\mathcal{L}_2(G, G^*) = 0$. Now, we express the equation $g_G(x) = g_{G^*}(x)$ as

$$\sum_{h'=1}^{H^*} \sum_{k'=1}^{K} \omega_{h',k'} \varphi\left(f_G^{h',k'}(x)\right) = \sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \omega_{h,k}^* \varphi\left(f_{G^*}^{h,k}(x)\right)$$

Since the set of functions $x \to \varphi(q(x))$ is linearly independent with different functions $q(x)$, we deduce that for each pair $(h, k) \in [H^*] \times [K^*]$, there exists a set $V_{h,k} = \left\{(h', k') \in [H^*] \times [K] : \varphi(f_G^{h',k'}(x)) = \varphi\left(f_{G^*}^{h,k}(x)\right)\right\}$. Since the function $\varphi(\cdot)$ is injective, we deduce that for each pair $(h, k) \in [H^*] \times [K^*]$, we have

$$f_G^{h',k'}(x) = f_{G_*}^{h,k}(x), \text{for almost every } x, \text{and} \sum_{(h',k') \in [H^*] \times [K^*]} \omega_{h',k'} = \omega_{h,k}, \forall (h', k') \in V_{h,k}. \tag{47}$$

We now prove that if $f_G^{h',k'}(x) = f_{G^*}^{h,k}(x)$ for almost every x, then we deduce

$$\{M_{h',i}, a_{h',i,k'}\}_{i=1}^{N^*} \equiv \{M_{h,i}, a_{h,i,k}\}_{i=1}^{N^*}. \tag{48}$$

For notational simplicity, we will prove that if

$$\sum_{i=1}^{N} \frac{\exp(x^\top M_i' x)}{\sum_{j=1}^{N} \exp(x^\top M_j' x)} \cdot (a_i')^\top x = \sum_{i=1}^{N} \frac{\exp(x^\top M_i x)}{\sum_{j=1}^{N} \exp(x^\top M_j x)} \cdot (a_i)^\top x, \text{for almost every } x, \tag{49}$$

where every element in each set $\{M_i'\}_{i=1}^N, \{M_i\}_{i=1}^N, \{a_i'\}_{i=1}^N$, and $\{a_i\}_{i=1}^N$ is pairwise distinct, then $\{(M_i, a_i)\}_{i=1}^N \equiv \{(M_i', a_i')\}_{i=1}^N$. The equation (49) is equivalent to

$$\sum_{1 \leq i,j \leq N} \exp(x^\top (M_i' + M_j) x) \cdot a_i' = \sum_{1 \leq i,j \leq N} \exp(x^\top (M_j' + M_i) x) \cdot a_i, \tag{50}$$

for almost every x. Since every element in each set $\{M_i'\}_{i=1}^N$ and $\{M_i\}_{i=1}^N$ is pairwise distinct, without loss of generality, there exists an open set $\mathcal{U} \in \mathbb{R}^d$ such that $\exp(u^\top M_1 u) > \exp(u^\top M_2 u) > \ldots > \exp(u^\top M_N u)$ and $\exp(u^\top M_1' u) > \exp(u^\top M_2' u) > \ldots > \exp(u^\top M_N' u)$ for any $u \in \mathcal{U}$. Therefore, if there exist two matrices $A \in \{M_i\}_{i=1}^N$ and $B \in \{M_i'\}_{i=1}^N$ such that $\exp(u^\top A u) = \exp(u^\top B u), \forall\, u \in \mathcal{U}$, we can deduce $A = B$. It is also worth noting that the set of functions $x \to e^{\eta x^2}$ is linearly independent with different parameters η. Now, we choose $x = pu$, where $p \in \mathbb{R}$, and we denote $m_i = u^\top M_i u, m_i' = u^\top M_i' u$ for all $i \in [N]$, the equation (50) can be rewritten as

$$\sum_{1 \leq i,j \leq N} e^{(m_i' + m_j) p^2} a_i' = \sum_{1 \leq i,j \leq N} e^{(m_i' + m_j) p^2} a_j, \tag{51}$$

for almost every p. Now, we will prove by induction that $a_t = a_t'$ and $m_t = m_t' + C$ for all $t = 1, 2, \ldots, N$, where C is a fixed scalar.

For $t = 1$, since $m_1 + m_1' > m_i + m_j', \forall\, (i, j') \neq (1, 1)$, the coefficient of $e^{(m_1 + m_1')p^2}$ is equal to 0, which leads to $a_1 = a_1'$.

For $t = 2$, we first remove the term $e^{(m_1 + m_1')p^2} a_1$ out of equation (51). Then, the maximal value of the remaining $m_i + m_j'$ is achieved by either $m_1' + m_2$ or $m_2' + m_1$. If $m_1' + m_2 > m_2' + m_1$,

the coefficient of $e^{(m_1'+m_2)p^2}$ deduces that $a_1' = a_2$, which is a contradiction, since $a_2 \neq a_1$. Then, $m_1' + m_2 \leq m_2' + m_1$. Similarly, we can also deduce that $m_2' + m_1 \leq m_1' + m_2$. Consequently, we have $m_1' + m_2 = m_2' + m_1$. Consider the coefficient of $e^{(m_1'+m_2)p^2}$, we have $a_1' + a_2' = a_1 + a_2$, which leads to $a_2' = a_2$.

Assume the claim holds for $t \geq 1$. We prove that it also holds for $t+1$. We remove the terms $e^{(m_i+m_j')p^2}a_i$ for any $1 \leq i, j \leq t$ out of equation (51). Then, the maximal value of remaining $m_i + m_j'$ is achieved by either $m_1 + m_t'$ or $m_1' + m_t$. By following the same line of reasoning as in the case $t = 2$, we deduce that $a_{t+1} = a_{t+1}'$ and $m_{t+1} - m_{t+1}' = m_1 - m_1'$.

Since $M_N = M_N' = \mathbf{0}_{\bar{d}\times\bar{d}}$, we deduce that $m_t = m_t'$ for any $t \in [N]$, which leads to $M_t = M_{t'}$ for any $t \in [N]$. The claim in (48) is completely proved. From (47) and (48), we deduce that $\mathcal{L}_2(G, G^*) = 0$, which completes our proof.

B Proofs of Auxiliary Results

B.1 Proof of Proposition 1

We begin the proof by defining notation. To begin with, we define $\mathcal{F}_{H^*,K,N^*}(\Theta)$ as the set of regression functions of all mixing measures in $\mathcal{G}_{H^*,K,N^*}(\Theta)$, or

$$\mathcal{F}_{H^*,K,N^*}(\Theta) := \{f_G(x) : G \in \mathcal{G}_{H^*,K,N^*}(\Theta)\}.$$

Given $\delta > 0$, the local $L^2(\mu)$ ball centered around the regression function $f_{G^*}(x)$ and intersected with the set $\mathcal{F}_{H^*,K,N^*}(\Theta)$ is

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

To quantify the complexity of this set, we employ the bracketing integral introduced by Geer et al. [33]:

$$J_B(\delta, \mathcal{F}_{H^*,K,N^*}(\Theta, \delta)) := \int_{\delta^2/2}^{\delta} H_B^{1/2}\big(t, \mathcal{F}_{H^*,K,N^*}(\Theta, t), \|\cdot\|_{L^2(\mu)}\big)\, dt \vee \delta, \tag{52}$$

where $H_B(t, \mathcal{F}_{H^*,K,N^*}(\Theta, t), \|\cdot\|_{L^2(\mu)})$ denotes the bracketing entropy [33] of $\mathcal{F}_{H^*,K,N^*}(\Theta, t)$ under the L^2-norm, and $t \vee \delta := \max\{t, \delta\}$. Adapting the arguments of Theorem 7.4 and Theorem 9.2 in [33] to our setting yields the following lemma:

Lemma 1. *Let $\Psi(\delta)$ be such that $\Psi(\delta) \geq J_B(\delta, \mathcal{F}_{H^*,K,N^*}(\Theta, \delta))$ and $\Psi(\delta)/\delta^2$ is non-increasing in δ. Then there exists a universal constant c and a sequence (δ_n) satisfying $\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n)$ for which*

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

for all $\delta \geq \delta_n$.

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

$$H_B\big(\varepsilon, \mathcal{F}_{H^*,K,N^*}(\Theta), \|\cdot\|_{L^2(\mu)}\big) \lesssim \log(1/\varepsilon), \tag{53}$$

for any $0 < \varepsilon \leq 1/2$. Indeed, for any function $f_G \in \mathcal{F}_{H^*,K,N^*}(\Theta)$, since the expert functions are bounded, we obtain that $f_G(x) \leq M$ for almost everywhere x, where $M > 0$ is some bounded constant of the expert functions. Choose $\tau \leq \varepsilon$ and take $\xi_1, \ldots, \xi_k$ as a τ-cover of $\mathcal{F}_{H^*,K,N^*}(\Theta)$ under the L^∞ norm of the set $\mathcal{F}_{H^*,K,N^*}(\Theta)$ where $k := N(\tau, \mathcal{F}_{H^*,K,N^*}(\Theta), \|\cdot\|_{L^\infty})$ is the τ-covering number of the metric space $(\mathcal{F}_{H^*,K,N^*}(\Theta), \|\cdot\|_{L^\infty})$. For each $i \in [k]$, define brackets $[L_i, U_i]$ by

$$L_i(x) := \max\{\xi_i(x) - \tau, 0\},$$
$$U_i(x) := \min\{\xi_i(x) + \tau, M\}.$$

This construction ensures that $\mathcal{F}_{H^*,K,N^*}(\Theta) \subset \bigcup_{i=1}^k [L_i(x), U_i(x)]$ and $U_i(x) - L_i(x) \leq \min\{2\tau, M\}$. Consequently,

$$\|U_i - L_i\|_{L^2(\mu)}^2 = \int (U_i - L_i)^2\, d\mu(x) \leq \int 4\tau^2\, d\mu(x) = 4\tau^2,$$

which implies that $\|U_i - L_i\|_{L^2(\mu)} \le 2\tau$. From the definition of bracketing entropy, we therefore obtain

$$H_B\big(2\tau, \mathcal{F}_{H^*,K,N^*}(\Theta), \|\cdot\|_{L^2(\mu)}\big) \le \log k = \log N(\tau, \mathcal{F}_{H^*,K,N^*}(\Theta), \|\cdot\|_{L^\infty}). \tag{54}$$

Consequently, we have to construct an upper bound on the covering number $N(\tau, \mathcal{F}_{H^*,K,N^*}(\Theta), \|\cdot\|_{L^\infty})$. To this end, we decompose the parameter space Θ into the following marginal parameter sets:

$$\Delta := \{M \in \mathbb{R}^{\overline{d}\times\overline{d}} : \ (M, a, \omega) \in \Theta\},$$

$$\Psi := \{a \in \mathbb{R}^{\overline{d}} : \ (M, a, \omega) \in \Theta\}, \qquad \Omega := \{\omega \in \mathbb{R} : \ (M, a, \omega) \in \Theta\}.$$

The compactness of Θ implies Δ, Ψ, and Ω are compact. Therefore, for any $\tau > 0$, there exist finite τ-covers Δ_τ, Ψ_τ, and Ω_τ of Δ, Ψ, and Ω, respectively. We have

$$|\Delta_\tau| \le O\Big(\tau^{-\overline{d}^2 H^* N^*}\Big), \qquad |\Psi_\tau| \le O\Big(\tau^{-\overline{d} H^* K N^*}\Big), \qquad |\Omega_\tau| \le O\Big(\tau^{-H^* K}\Big).$$

For each mixing measure

$$G = \sum_{h=1}^{H^*}\sum_{k=1}^{K} \omega_{h,k} \sum_{i=1}^{N^*} \delta_{(M_{h,i}, a_{h,i,k})},$$

we define the softmax weights by

$$\sigma_{h,i}(x; M_h) := \frac{\exp(x^\top M_{h,i} x)}{\sum_{j=1}^{N} \exp(x^\top M_{h,j} x)}, \qquad h \in [H^*], \ i \in [N].$$

We then consider the following three mixing measures:

$$\check{G} := \sum_{h=1}^{H^*}\sum_{k=1}^{K} \omega_{h,k} \sum_{i=1}^{N^*} \delta_{(M_{h,i}, \bar{a}_{h,i,k})},$$

$$\tilde{G} := \sum_{h=1}^{H^*}\sum_{k=1}^{K} \bar{\omega}_{h,k} \sum_{i=1}^{N^*} \delta_{(M_{h,i}, \bar{a}_{h,i,k})},$$

$$\bar{G} := \sum_{h=1}^{H^*}\sum_{k=1}^{K} \bar{\omega}_{h,k} \sum_{i=1}^{N^*} \delta_{(\bar{M}_{h,i}, \bar{a}_{h,i,k})}.$$

Here $\bar{a}_{h,i,k} \in \Psi\tau$ is the nearest point to $a_{h,i,k}$, $\bar{\omega}h, k \in \Omega\tau$ is the nearest to $\omega_{h,k}$, and $(\bar{M}_{h,i}, \bar{c}_{h,i}) \in \Delta_\tau$ is the nearest to $(M_{h,i}, c_{h,i})$.

From these definitions, we obtain

$$\begin{aligned}
\|f_G - f_{\check{G}}\|_{L^\infty} &= \sup_{x\in\mathcal{X}} \left| \sum_{h=1}^{H^*}\sum_{k=1}^{K} \omega_{h,k} \sum_{i=1}^{N^*} \sigma_{h,i}(x; M_h) \cdot \Big[(a_{h,i,k})^\top x - (\bar{a}_{h,i,k})^\top x \Big] \right| \\
&\le \sum_{h=1}^{H^*}\sum_{k=1}^{K}\sum_{i=1}^{N^*} \sup_{x\in\mathcal{X}} \Big|\omega_{h,k}\Big| \, \sigma_{h,i}(x; M_h) \, \Big|\big(a_{h,i,k} - \bar{a}_{h,i,k}\big)^\top x\Big| \\
&\le \sum_{h=1}^{H^*}\sum_{k=1}^{K}\sum_{i=1}^{N^*} \sup_{x\in\mathcal{X}} \Big|\omega_{h,k}\Big| \, \Big\|a_{h,i,k} - \bar{a}_{h,i,k}\Big\|_2 \, \|x\|_2 \\
&\lesssim \sum_{h=1}^{H^*}\sum_{k=1}^{K}\sum_{i=1}^{N^*} \Big\|a_{h,i,k} - \bar{a}_{h,i,k}\Big\|_2 \lesssim \tau.
\end{aligned}$$

The first inequality uses the triangle inequality, the second uses that the softmax weight is at most 1, and the final bound follows because $|x| \le B$ for some $B > 0$ and ω is bounded over Θ.

Next, we have

$$\|f_{\check{G}} - f_{\tilde{G}}\|_{L^\infty} = \sup_{x\in\mathcal{X}} \left| \sum_{h=1}^{H^*}\sum_{k=1}^{K}\sum_{i=1}^{N^*} \Big(\omega_{h,k} - \bar{\omega}_{h,k}\Big) \, \sigma_{h,i}(x; M_h) \, (\bar{a}_{h,i,k})^\top x \right|$$

$$\leq \sum_{h=1}^{H^*}\sum_{k=1}^{K}\sum_{i=1}^{N^*}\sup_{x\in\mathcal{X}}\left|\omega_{h,k}-\bar{\omega}_{h,k}\right|\left|(\bar{a}_{h,i,k})^\top x\right|$$
$$\lesssim \sum_{h=1}^{H^*}\sum_{k=1}^{K}\left|\omega_{h,k}-\bar{\omega}_{h,k}\right| \lesssim \tau.$$

Again, the triangle inequality gives the first bound, and the second uses that each expert $(\bar{a}_{h,i,k})^\top x$ is uniformly bounded.

Finally, we have

$$\|f_{\tilde{G}}-f_{\bar{G}}\|_{L^\infty} = \sup_{x\in\mathcal{X}}\left|\sum_{h=1}^{H^*}\sum_{k=1}^{K}\sum_{i=1}^{N^*}\bar{\omega}_{h,k}\left[\sigma_{h,i}(x;M_h)-\sigma_{h,i}(x;\bar{M}_h)\right](\bar{a}_{h,i,k})^\top x\right|$$
$$\leq \sum_{h=1}^{H^*}\sum_{k=1}^{K}\sum_{i=1}^{N^*}\sup_{x\in\mathcal{X}}|\bar{\omega}_{h,k}|\left|\sigma_{h,i}(x;M_h)-\sigma_{h,i}(x;\bar{M}_h)\right|\left|(\bar{a}_{h,i,k})^\top x\right|.$$

For this term, we use the Lipschitz continuity of the softmax function. In particular, for each fixed h and i, the map

$$(M_h)\mapsto \sigma_{h,i}(x;M_h)=\frac{\exp(x^\top M_{h,i}x)}{\sum_{j=1}^{N}\exp(x^\top M_{h,j}x)}$$

is Lipschitz uniformly over $x\in\mathcal{X}$ since $\mathcal{X}$ is bounded and (M) ranges over a compact set. Therefore,

$$\|f_{\tilde{G}}-f_{\bar{G}}\|_{L^\infty}\lesssim \sum_{h=1}^{H^*}\sum_{i=1}^{N^*}\left(\|M_{h,i}-\bar{M}_{h,i}\|_F\cdot\|x\|^2\right)\lesssim\tau.$$

Above, the last inequality occurs as the input space is bounded, that is, $\|x\|\leq B$, and $(\bar{M}_{h,i},\bar{c}_{h,i})$ is chosen from a τ-cover.

According to the triangle inequality, we have

$$\|f_G-f_{\bar{G}}\|_{L^\infty}\leq\|f_G-f_{\check{G}}\|_{L^\infty}+\|f_{\check{G}}-f_{\bar{G}}\|_{L^\infty}+\|f_{\tilde{G}}-f_{\bar{G}}\|_{L^\infty}\lesssim\tau.$$

By the definition of the covering number, we deduce that

$$N(\tau,\mathcal{F}_{H^*,K,N^*}(\Theta),\|\cdot\|_{L^\infty})\leq|\Delta_\tau|\times|\Psi_\tau|\times|\Omega_\tau|$$
$$\leq O\Big(\tau^{-\overline{d}^2H^*N^*}\Big)\times O\Big(\tau^{-\bar{d}H^*KN^*}\Big)\times O\Big(\tau^{-H^*K}\Big)$$
$$\leq O\left(\tau^{-\left[\overline{d}^2H^*N+\bar{d}H^*KN+H^*K\right]}\right). \tag{55}$$

Substituting the covering number bound (55) into (54), we obtain

$$H_B\big(2\tau,\mathcal{F}_{H^*,K,N^*}(\Theta),\|\cdot\|_{L^2(\mu)}\big)\leq\log N(\tau,\mathcal{F}_{H^*,K,N^*}(\Theta),\|\cdot\|_{L^\infty})$$
$$\leq\log\Big(C\,\tau^{-\left[d^2H^*N^*+dH^*KN^*+H^*K\right]}\Big)$$
$$\leq C_1\log(1/\tau),$$

where $C_1>0$ is a constant depending on $(H^*,K,N^*,\overline{d})$. Replacing 2τ by ε yields

$$H_B\big(\varepsilon,\mathcal{F}_{H^*,K,N^*}(\Theta),\|\cdot\|_{L^2(\mu)}\big)\lesssim\log(1/\varepsilon).$$

As a result, it follows that

$$J_B(\delta,\mathcal{F}_{H^*,K,N^*}(\Theta,\delta))=\int_{\delta^2/2}^{\delta}H_B^{1/2}\big(t,\mathcal{F}_{H^*,K,N^*}(\Theta,t),\|\cdot\|_{L^2(\mu)}\big)\,dt\vee\delta \tag{56}$$
$$\lesssim\int_{\delta^2/2}^{\delta}\sqrt{\log(1/t)}\,dt\vee\delta. \tag{57}$$

Define $\Psi(\delta)=\delta\sqrt{\log(1/\delta)}$. Then $\Psi(\delta)/\delta^2$ is non-increasing in δ. Equation (57) shows that $\Psi(\delta)\geq J_B(\delta,\mathcal{F}_{H^*,K,N^*}(\Theta,\delta))$. Take $\delta_n=\sqrt{\log(n)/n}$. Then for some universal constant c, we have $\sqrt{n}\delta_n^2\geq c\Psi(\delta_n)$. Finally, applying Lemma 1 yields the statement of the theorem.

B.2 Sigmoid, Tanh, and Softplus Are Type-1/Type-2 Strongly Identifiable Functions

Part I. Sigmoid, Tanh, and Softplus Are Type-1 Strongly Identifiable

To prove that $x \mapsto \varphi(x) \in \{\mathrm{sigmoid}(x+b), \tanh(x+b), \mathrm{softplus}(x+b)\}$ with $b \neq 0$ satisfies the Type-1 strong identifiability condition, we note that these functions are injective, twice differentiable, and Lipschitz continuous. We then show that all the coefficients of the following equation equal to 0:

$$\sum_{i=1}^{N}\sum_{k=0}^{2}\sum_{\alpha\in\mathcal{I}_k}\gamma_{\alpha,k,i}\,x^{\alpha}\exp(x^{\top}M_i x)\varphi^{(k)}(a_i^{\top}x)=0, \text{for almost every } x.$$

Since the left-hand side is real analytic in x, its vanishing almost everywhere implies that it vanishes identically. Collecting terms according to the derivative order of φ, the identity can be rewritten as

$$\sum_{i=1}^{N}\exp(x^{\top}M_i x)\left[A_i(x)\varphi(a_i^{\top}x)+B_i(x)\varphi'(a_i^{\top}x)+C_i(x)\varphi''(a_i^{\top}x)\right]=0,$$

where $A_i(x) = Q_{4,i} + Q_{2,a,i} + Q_{0,i}$, $B_i(x) = Q_{3,i} + Q_{1,i}$, and $C_i(x) = Q_{2,b,i}$, where $Q_{4,i}, Q_{3,i}, Q_{2,a,i}, Q_{2,b,i}, Q_{1,i}, Q_{0,i}$ denote the homogeneous polynomial contributions corresponding to the index pairs $(|\alpha|, k) = (4,0), (3,1), (2,0), (2,2), (1,1), (0,0)$, respectively.

For each i, the functions $\varphi(a_i^{\top}x)$, $\varphi'(a_i^{\top}x)$, and $\varphi''(a_i^{\top}x)$ become singular on the complex hyperplanes

$$a_i^{\top}x + b = z_m, \qquad m \in \mathbb{Z},$$

where $z_m = (m + 1/2)\pi\mathrm{i}$ for tanh and $z_m = (2m+1)\pi\mathrm{i}$ for sigmoid and softplus.

For sigmoid and tanh, these functions have poles of orders $1, 2, 3$, respectively. For softplus, they have a logarithmic singularity, a simple pole, and a double pole, respectively. Hence, their singularities have increasing strengths.

Since $b > 0$ and the vectors a_i are pairwise distinct and nonzero, the singular hyperplanes associated with different indices i do not coincide. At a generic point of a hyperplane associated with a_i, all terms indexed by $j \neq i$ remain holomorphic. Hence, the coefficient $C_i(x)$ of the strongest singular term must vanish on that hyperplane. Since this holds for infinitely many parallel hyperplanes, we deduce that $C_i(x) \equiv 0$. Applying the same argument to the remaining terms successively gives $B_i(x) \equiv 0$ and $A_i(x) \equiv 0$.

Finally, since the homogeneous polynomials appearing in $A_i(x), B_i(x)$, and $C_i(x)$ have different degrees, we deduce that

$$Q_{4,i} = Q_{3,i} = Q_{2,a,i} = Q_{2,b,i} = Q_{1,i} = Q_{0,i} = 0, \quad \forall\, 1 \leq i \leq N.$$

By the definitions of these polynomial components and the linear independence of distinct monomials x^{α}, it follows that

$$\gamma_{\alpha,k,i} = 0, \quad \forall\, i,\ k \in \{0,1,2\},\ \alpha \in \mathcal{I}_k.$$

Therefore, the collection of functions in Definition 1 is linearly independent, completing the proof.

Part II. Type-2 Strong Identifiability of Sigmoid, Tanh, and Softplus

To prove that $z \mapsto \varphi(z) \in \{\mathrm{sigmoid}(z+b), \tanh(z+b), \mathrm{softplus}(z+b)\}$, with $b \neq 0$, satisfies the Type-2 strong identifiability condition, we show that all the coefficients of the following equation equal to 0:

$$\sum_{h=1}^{H^*}\sum_{k=1}^{K^*}\Bigg[A_{h,k}\varphi(f_{G^*}^{h,k}(x)) + \sum_{i=1}^{N^*}\sum_{\alpha\in\mathcal{J}}\gamma_{h,k,i,\alpha}x^{\alpha}u_{h,k,i}(x)\varphi'(f_{G^*}^{h,k}(x))$$
$$+\sum_{1\leq i\leq j\leq N^*}\sum_{\alpha\in\mathcal{J}}\eta_{h,k,i,j,\alpha}x^{\alpha}u_{h,k,i}(x)u_{h,k,j}(x)\varphi''(f_{G^*}^{h,k}(x))\Bigg] = 0, \text{ for almost every } x,$$

where we denote $u_{h,k,i}(x) := \exp(x^\top M^*_{h,i} x) \cdot ((a^*_{h,i,k})^\top x)$. Since the left-hand side is real analytic in x, its vanishing almost everywhere implies that it vanishes identically. Collecting terms according to the derivative order of φ, the identity can be rewritten as

$$\sum_{h=1}^{H^*} \sum_{k=1}^{K^*} \Big[A_{h,k} \varphi(f^{h,k}_{G^*}(x)) + B_{h,k}(x) \varphi'(f^{h,k}_{G^*}(x)) + C_{h,k}(x) \varphi''(f^{h,k}_{G^*}(x)) \Big] = 0,$$

where

$$B_{h,k}(x) = \sum_{i=1}^{N^*} \sum_{\alpha \in \mathcal{J}} \gamma_{h,k,i,\alpha} x^\alpha u_{h,k,i}(x)$$
$$C_{h,k}(x) = \sum_{1 \le i \le j \le N^*} \sum_{\alpha \in \mathcal{J}} \eta_{h,k,i,j,\alpha} x^\alpha u_{h,k,i}(x) u_{h,k,j}(x).$$

For sigmoid and tanh, the functions φ, φ', and φ'' have poles of orders $1, 2, 3$, respectively. For softplus, they have a logarithmic singularity, a simple pole, and a double pole, respectively.

For each (h, k), the singularities of these functions occur when

$$f^{h,k}_{G^*}(x) + b = z_m,$$

where $z_m = (m + 1/2)\pi\mathrm{i}$ for tanh and $z_m = (2m + 1)\pi\mathrm{i}$ for sigmoid and softplus. We denote $f^{h,k}_{G^*}(x) = \frac{P_{h,k}(x)}{Z_h(x)}$, where

$$P_{h,k}(x) := \sum_{i=1}^{N^*} u_{h,k,i}(x), \qquad Z_h(x) := \sum_{i=1}^{N^*} \exp(x^\top M^*_{h,i} x).$$

For each (h, k), consider a smooth genuine pole hypersurface

$$\mathcal{H}_{h,k} := \{x : Z_h(x) = 0,\ P_{h,k}(x) \neq 0,\ \nabla Z_h(x) \neq 0\}$$

of $f^{h,k}_{G^*}$. Near $\mathcal{H}_{h,k}$, the equations $f^{h,k}_{G^*}(x) + b = z_m$ define singular hypersurfaces that accumulate at $\mathcal{H}_{h,k}$, where every smooth point $x_0 \in \mathcal{H}_{h,k}$ is the limit of a sequence (x_m) satisfying $f^{h,k}_{G^*}(x_m) + b = z_m$ as $|m| \to \infty$. At generic points of these hypersurfaces that are not shared by other components, the strongest singularity therefore forces $C_{h,k}$ to vanish.

Since these singular hypersurfaces accumulate at $\mathcal{H}_{h,k}$ and $C_{h,k}(x)$ is holomorphic in its neighborhood, the identity theorem applied along the transverse complex direction implies $C_{h,k}(x) \equiv 0$. Repeating this argument for the remaining terms yields $B_{h,k}(x) \equiv 0$ and $A_{h,k} = 0$. Consequently, we obtain that all the coefficients $\gamma_{h,k,i,\alpha}$ and $\eta_{h,k,i,j,\alpha}$ are equal to zero, completing the proof.

C Experiment Details and Results

This appendix is organized into two parts. Section C.1 provides the ground-truth parameters used in the three experiments described in Section 6. Section C.2 reports the detailed results of Experiments 2 and 3, which were deferred from the main text.

C.1 Experiment Details

Ground-truth parameters for Experiments 1 and 2

For Experiments 1 and 2 ($d = 2$, $H^* = N^* = K^* = 2$), the gating matrices M^* are specified as follows:

$$M^*_{0,0} = \begin{bmatrix} 2.0 & 0.0 \\ 0.0 & 0.5 \end{bmatrix}, \quad M^*_{0,1} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$
$$M^*_{1,0} = \begin{bmatrix} 1.5 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}, \quad M^*_{1,1} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Table 2: Ground-truth parameters used in Experiments 1 and 2 for gating weights $\omega^*_{h,k}$ and expert coefficients $a^*_{h,i,k} \in \mathbb{R}^2$ with $H^* = 2$ heads, $N^* = 2$ experts per head, and $K^* = 2$ channels.

Head h	Expert i	Channel k	$\omega^*_{h,k}$	$a^*_{h,i,k}$
0	0	0	1.0	$(1.0, -0.5)$
0	0	1	0.5	$(0.5, 0.8)$
0	1	0	1.0	$(-1.0, 0.8)$
0	1	1	0.5	$(0.2, -0.3)$
1	0	0	0.8	$(0.6, 0.4)$
1	0	1	0.3	$(-0.2, 0.5)$
1	1	0	0.8	$(-0.7, -0.2)$
1	1	1	0.3	$(0.3, -0.4)$

The corresponding gating weights $\omega^*_{h,k}$ and expert coefficients $a^*_{h,i,k} \in \mathbb{R}^2$ are listed in Table 2.

Ground-truth parameters for Experiment 3

For Experiment 3 ($d = 3$, $H^* = N^* = K^* = 3$), the gating matrices $M^* \in \mathbb{R}^{3\times 3}$ are specified as follows:

$$M^*_{0,0} = \begin{bmatrix} 2.0 & 0.0 & 0.3 \\ 0.0 & 0.5 & 0.1 \\ 0.3 & 0.1 & 1.0 \end{bmatrix}, \quad M^*_{0,1} = \begin{bmatrix} 0.5 & 0.0 & 0.2 \\ 0.0 & 2.0 & 0.1 \\ 0.2 & 0.1 & 0.8 \end{bmatrix}, \quad M^*_{0,2} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$M^*_{1,0} = \begin{bmatrix} 1.5 & 0.2 & 0.1 \\ 0.2 & 0.3 & 0.0 \\ 0.1 & 0.0 & 1.0 \end{bmatrix}, \quad M^*_{1,1} = \begin{bmatrix} 0.3 & -0.1 & 0.0 \\ -0.1 & 1.2 & 0.2 \\ 0.0 & 0.2 & 0.7 \end{bmatrix}, \quad M^*_{1,2} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$M^*_{2,0} = \begin{bmatrix} 1.0 & 0.3 & 0.1 \\ 0.3 & 0.8 & 0.0 \\ 0.1 & 0.0 & 0.6 \end{bmatrix}, \quad M^*_{2,1} = \begin{bmatrix} 0.7 & 0.0 & 0.2 \\ 0.0 & 1.3 & 0.1 \\ 0.2 & 0.1 & 0.5 \end{bmatrix}, \quad M^*_{2,2} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

The corresponding gating weights $\omega^*_{h,k}$ and expert coefficients $a^*_{h,i,k} \in \mathbb{R}^3$ are listed in Table 3.

Table 3: Ground-truth parameters used in Experiment 3 for gating weights $\omega^*_{h,k}$ and expert coefficients $a^*_{h,i,k} \in \mathbb{R}^3$ with $H^* = 3$ heads, $N^* = 3$ experts per head, and $K^* = 3$ channels.

Head h	Expert i	Channel k	$\omega^*_{h,k}$	$a^*_{h,i,k}$
0	0	0	1.0	$(1.0, -0.5, 0.3)$
0	0	1	0.5	$(0.5, 0.8, -0.2)$
0	0	2	0.6	$(0.4, -0.3, 0.6)$
0	1	0	1.0	$(-1.0, 0.8, 0.1)$
0	1	1	0.5	$(0.2, -0.3, 0.5)$
0	1	2	0.6	$(-0.5, 0.4, -0.3)$
0	2	0	1.0	$(0.6, 0.3, -0.4)$
0	2	1	0.5	$(-0.2, 0.7, 0.2)$
0	2	2	0.6	$(0.3, -0.5, 0.8)$
1	0	0	0.8	$(0.6, 0.4, -0.2)$
1	0	1	0.3	$(-0.2, 0.5, 0.3)$
1	0	2	0.4	$(0.5, -0.3, 0.7)$
1	1	0	0.8	$(-0.7, -0.2, 0.4)$
1	1	1	0.3	$(0.3, -0.4, 0.1)$
1	1	2	0.4	$(-0.4, 0.6, -0.5)$
1	2	0	0.8	$(0.2, 0.8, -0.3)$
1	2	1	0.3	$(-0.6, 0.1, 0.4)$
1	2	2	0.4	$(0.7, -0.2, 0.3)$
2	0	0	0.7	$(0.5, 0.6, -0.4)$
2	0	1	0.4	$(-0.3, 0.4, 0.7)$
2	0	2	0.2	$(0.2, -0.5, 0.3)$
2	1	0	0.7	$(-0.4, 0.3, 0.5)$
2	1	1	0.4	$(0.6, -0.7, 0.1)$
2	1	2	0.2	$(-0.2, 0.5, -0.6)$
2	2	0	0.7	$(0.8, -0.1, 0.3)$
2	2	1	0.4	$(-0.5, 0.3, -0.2)$
2	2	2	0.2	$(0.1, 0.7, -0.4)$

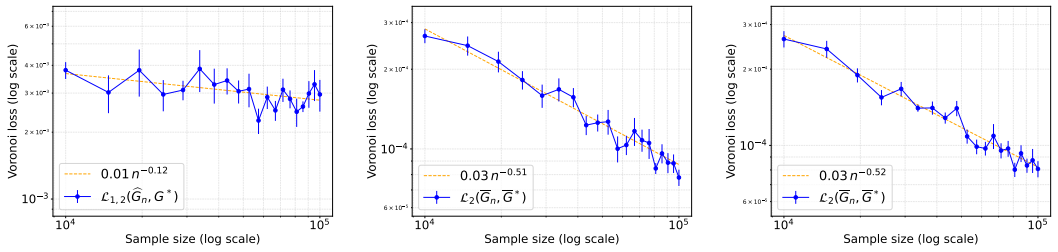
C.2 Experimental Results

Results of Experiment 2

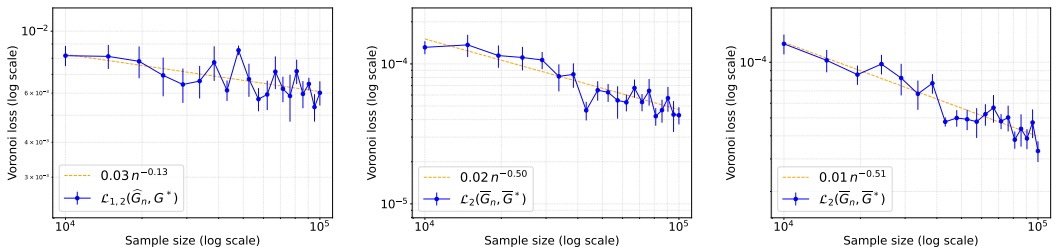
To assess whether the sample-efficiency advantage observed in Experiment 1 is specific to the Sigmoid activation, we repeat the experiment with tanh function. As shown in Figure 2, gated attention with tanh function likewise attains fast convergence in both settings. The Voronoi loss decays at empirical rates of approximately $O(n^{-0.51})$ and $O(n^{-0.50})$ in Setting I (Figures 2b and 2e), and $O(n^{-0.52})$ and $O(n^{-0.51})$ in Setting II (Figures 2c and 2f). Vanilla multi-head self-attention again converges much more slowly, at rates of $O(n^{-0.12})$ and $O(n^{-0.13})$ (Figures 2a and 2d). These results confirm that the sample-efficiency advantage of gated attention is robust to the choice of nonlinearity.

Results of Experiment 3

We further evaluate our theory under higher-dimensional inputs ($d = 3$) and larger head/expert/channel configurations ($H^* = N^* = K^* = 3$, with $K \in \{4, 5\}$ fitted channels). Figure 3 shows that gated attention continues to converge rapidly in this larger configuration. The Voronoi loss decays at empirical rates of approximately $O(n^{-0.53})$ and $O(n^{-0.52})$ in Setting I (Fig-

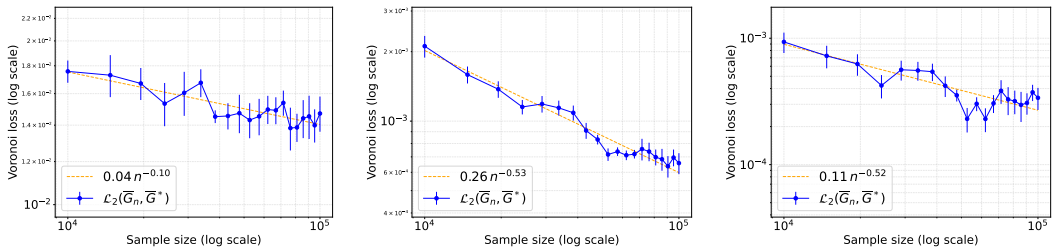


(a) MHA ($K = 3$) (b) Value Gated Attention ($K = 3$) (c) SDPA Gated Attention ($K = 3$)

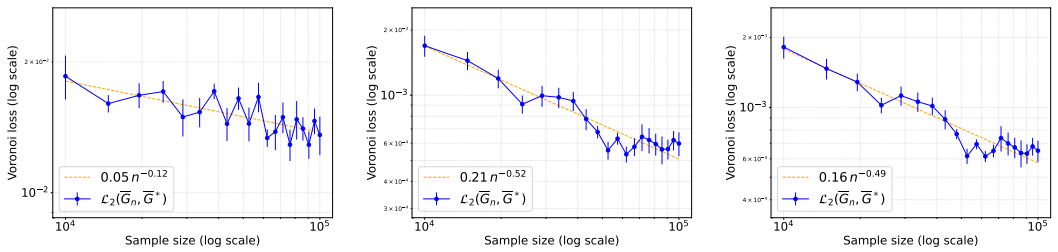


(d) MHA ($K = 4$) (e) Value Gated Attention ($K = 4$) (f) SDPA Gated Attention ($K = 4$)

Figure 2: Experiment 2: log-log plots of empirical Voronoi losses versus sample size n for multi-head self-attention and gated attention mechanisms with tanh activation function.



(a) MHA ($K = 4$) (b) Value Gated Attention ($K = 4$) (c) SDPA Gated Attention ($K = 4$)



(d) MHA ($K = 5$) (e) Value Gated Attention ($K = 5$) (f) SDPA Gated Attention ($K = 5$)

Figure 3: Experiment 3: log-log plots of empirical Voronoi losses versus sample size n for multi-head self-attention and gated attention mechanisms in higher-dimensional inputs ($d = 3$) and larger head/expert/channel configurations ($H^* = N^* = K^* = 3$).

ures 3b and 3e), and $O(n^{-0.52})$ and $O(n^{-0.49})$ in Setting II (Figures 3c and 3f). Vanilla multi-head self-attention, by contrast, exhibits notably slower rates of $O(n^{-0.10})$ and $O(n^{-0.12})$ (Figures 3a and 3d). These results confirm that the sample-efficiency advantage of gated attention persists in higher-dimensional inputs and larger model configurations.

D Additional Experiments on Real-World Data

D.1 Objective

The controlled HMoE experiments in Section 6 isolate the parameter-estimation behavior predicted by our theory. We complement them with an end-to-end language-modeling experiment designed to answer a practical question: when the remainder of a Transformer and its training recipe are fixed, does inserting the nonlinearities studied in Settings I and II improve performance as the number of training tokens increases? Following the language-pretraining pipeline of CompeteSMoE [26], we train standard MHA, Value Gated Attention, and SDPA Gated Attention from scratch and evaluate the same checkpoints throughout pretraining.

D.2 Dataset and Prediction Task

We pretrain on a subset of SlimPajama, a cleaned and deduplicated language corpus assembled from seven public data sources. Documents are tokenized with an 8,000-token SentencePiece vocabulary and packed into sequences of 1,024 tokens. The training objective is causal next-token prediction: for a token sequence $(x_1, \ldots, x_T)$, the model minimizes

$$\mathcal{L}_{\mathrm{LM}} = -\frac{1}{T-1} \sum_{t=1}^{T-1} \log p_\theta(x_{t+1} \mid x_{1:t}).$$

Training shards are traversed in a fixed randomized order shared by all three variants. We use a disjoint SlimPajama holdout subset for validation and report perplexity, $\exp(\mathcal{L}_{\mathrm{LM}})$, so lower values indicate better language modeling.

D.2.1 Model and Attention Variants

The backbone is a 153,723,712-parameter decoder-only sparse Transformer. It contains 16 pre-layer-normalized Transformer layers, hidden dimension 512, four causal self-attention heads per layer, head-projection dimension 82, rotary positional embeddings, and ReLU feed-forward blocks. Each sparse feed-forward block has 64 experts of hidden dimension 128 and routes every token to its top two experts. All three variants use exactly the same embeddings, sparse feed-forward layers, normalization, residual connections, causal masks, and output head.

Only the attention operation is changed. For one head, let

$$Q = XW_Q, \qquad K = XW_K, \qquad V = XW_V, \qquad A = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right).$$

The three compared attention modules are

$$\begin{aligned} \text{MHA (Control):} \quad & Y = AV, \\ \text{Value Gated (VG):} \quad & Y = A\varphi(V), \\ \text{SDPA Gated (SG):} \quad & Y = \varphi(AV), \end{aligned}$$

followed by the same output projection. We apply the activation elementwise with $\varphi(z) = \tanh(z + 0.001)$. This implementation directly matches Settings I and II: VG places the nonlinearity after the value projection, whereas SG places it after the complete scaled dot-product attention output. The activation introduces no learned parameters, so all variants have the same parameter count.

D.3 Training and Evaluation Protocol

We initialize every model with seed 42, 52 and optimize it with AdamW using a learning rate of 2.5×10^{-4}, weight decay 0.01, gradient clipping at 0.1, mixed-precision training, and a cosine learning-rate schedule with minimum multiplier 0.1. We use no model or attention dropout. As in the CompeteSMoE pipeline, the sparse feed-forward layers use balancing and router regularization; these terms and their coefficients are identical for all attention variants. Training is conducted on an NVIDIA A100 80GB GPU.

Each optimization step processes $64 \times 1{,}024 = 65{,}536$ tokens. We save a checkpoint every 10,000 steps and continue training from that checkpoint on the next nonoverlapping block of SlimPajama

Table 4: Architecture and pretraining configuration for the real-world validation.

Configuration	Value
Transformer layers / hidden size	16 / 512
Attention heads / head dimension	4 / 82
Feed-forward experts / active experts	64 / 2
Expert hidden dimension	128
Parameters	153,723,712
Context length / vocabulary size	1,024 / 8,000
Batch size	64 sequences
Optimizer / learning rate	AdamW / 2.5×10^{-4}
Weight decay / gradient clipping	0.01 / 0.1
Learning-rate schedule	Cosine
Dropout	0

tokens. Consequently, the checkpoints at 10,000 through 70,000 steps correspond to approximately 655M, 1,311M, 1,966M, 2,621M, 3,277M, 3,932M, and 4,588M seen tokens.

We evaluate each checkpoint on the fixed SlimPajama validation split and four zero-shot benchmarks. HellaSwag measures commonsense sentence completion, the Children's Book Test (CBT) evaluates contextual cloze prediction, BLiMP tests grammatical knowledge through minimal sentence pairs, and RACE evaluates reading comprehension. For each benchmark item, the model scores the candidate completions autoregressively and selects the candidate with the highest likelihood. Table 5 reports validation perplexity and benchmark accuracy at every token budget.

D.3.1 Results

Language modeling. Both gated-attention variants have lower mean validation perplexity than standard MHA at all eight reported budgets, with VG obtaining the lowest mean throughout. At 327.68M processed tokens, the mean perplexities are 7.2538, 7.2377, and 7.2472 for MHA, VG, and SG, respectively. At 4,587.52M tokens, they are 6.7904, 6.7590, and 6.7710. The improvements over MHA are consistent across the reported budgets for this seed pair.

Zero-shot evaluation. A gated variant has the highest mean accuracy in 31 of the 32 budget–benchmark comparisons; the exception is BLiMP at 10,000 steps. Individually, VG exceeds MHA in 29 comparisons and SG in 26, while both exceed MHA in 24. At the largest budget, VG improves over MHA by 0.26, 0.51, 0.58, and 0.61 percentage points on HellaSwag, CBT, BLiMP, and RACE, respectively; SG improves by 0.35, 0.34, 0.50, and 0.73 points. The downstream results favor gated attention overall for this seed pair, but the ranking depends on the task and budget, and the reported standard errors reflect only two runs.

Table 5: Validation perplexity and zero-shot accuracy (%) throughout SlimPajama pretraining. Entries are mean ± one standard error across seeds 42 and 52 ($n = 2$), with $\mathrm{SE} = s/\sqrt{2}$ using the sample standard deviation s. Control denotes standard MHA; VG and SG apply $\tanh(z + 0.001)$ after the value projection and SDPA output, respectively. Bold entries mark the best unrounded means within each budget (lower perplexity, higher accuracy). Tokens count processed tokens, not necessarily unique tokens.

Tokens (M)	Attention	Perplexity ↓	HellaSwag ↑	CBT ↑	BLiMP ↑	RACE ↑
327.68	Control	7.2538 ± 0.0030	27.99 ± 0.17	57.08 ± 0.30	66.32 ± 0.04	28.78 ± 0.19
	VG	**7.2377 ± 0.0023**	**28.42 ± 0.04**	**57.24 ± 0.07**	66.37 ± 0.48	**29.37 ± 0.24**
	SG	7.2472 ± 0.0104	27.85 ± 0.06	56.71 ± 0.03	**66.51 ± 0.23**	29.17 ± 0.06
655.36	Control	6.3808 ± 0.0037	27.68 ± 0.16	60.93 ± 0.13	**69.58 ± 0.24**	29.44 ± 0.16
	VG	**6.3435 ± 0.0010**	27.93 ± 0.09	60.83 ± 0.07	69.25 ± 0.11	**29.68 ± 0.20**
	SG	6.3507 ± 0.0119	**27.97 ± 0.02**	**60.97 ± 0.00**	69.06 ± 0.71	29.42 ± 0.58
1,310.72	Control	6.7146 ± 0.0192	27.81 ± 0.00	65.66 ± 0.19	72.02 ± 0.03	29.27 ± 0.18
	VG	**6.6759 ± 0.0188**	28.10 ± 0.05	66.25 ± 0.15	72.09 ± 0.14	29.35 ± 0.09
	SG	6.6828 ± 0.0286	**28.11 ± 0.09**	**66.53 ± 0.14**	**72.09 ± 0.61**	**29.68 ± 0.07**
1,966.08	Control	6.8119 ± 0.0645	28.13 ± 0.15	66.87 ± 0.10	72.75 ± 0.16	29.60 ± 0.22
	VG	**6.7771 ± 0.0680**	**28.29 ± 0.06**	**67.59 ± 0.15**	72.77 ± 0.27	**29.84 ± 0.26**
	SG	6.7827 ± 0.0550	28.02 ± 0.02	67.54 ± 0.25	**73.01 ± 0.73**	29.82 ± 0.10
2,621.44	Control	6.7486 ± 0.0028	28.06 ± 0.11	68.17 ± 0.02	73.44 ± 0.23	29.53 ± 0.13
	VG	**6.7090 ± 0.0064**	**28.33 ± 0.08**	**68.54 ± 0.12**	73.21 ± 0.20	**29.87 ± 0.08**
	SG	6.7269 ± 0.0003	28.27 ± 0.05	68.47 ± 0.24	**73.51 ± 0.52**	29.60 ± 0.10
3,276.80	Control	6.8050 ± 0.0414	28.36 ± 0.00	68.86 ± 0.22	73.71 ± 0.25	29.52 ± 0.13
	VG	**6.7791 ± 0.0570**	**28.55 ± 0.16**	69.23 ± 0.28	74.10 ± 0.36	29.84 ± 0.06
	SG	6.7908 ± 0.0382	28.50 ± 0.08	**69.46 ± 0.14**	**74.12 ± 0.19**	**29.90 ± 0.35**
3,932.16	Control	6.7869 ± 0.0161	28.59 ± 0.02	69.38 ± 0.13	73.07 ± 0.08	29.42 ± 0.10
	VG	**6.7518 ± 0.0002**	**28.66 ± 0.00**	69.83 ± 0.03	**73.78 ± 0.66**	29.75 ± 0.05
	SG	6.7702 ± 0.0128	28.56 ± 0.04	**69.91 ± 0.14**	73.66 ± 0.21	**30.01 ± 0.20**
4,587.52	Control	6.7904 ± 0.0328	28.65 ± 0.23	69.67 ± 0.19	73.13 ± 0.06	29.38 ± 0.08
	VG	**6.7590 ± 0.0100**	28.91 ± 0.06	**70.17 ± 0.41**	**73.71 ± 0.74**	29.99 ± 0.20
	SG	6.7710 ± 0.0276	**29.00 ± 0.11**	70.00 ± 0.21	73.62 ± 0.20	**30.11 ± 0.34**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction provide a precise and comprehensive overview of the paper's primary contributions and thematic scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We included in the Discussion Section.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions are already stated.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided experiment descriptions in the main paper and appendix. Furthermore, we will release our GitHub implementation upon the accepteance of this paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will release our GitHub implementation if the paper is accepted.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provided details for experiments in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report it in our experiment results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The experiments are conducted at a very small scale on synthetic data with low-dimensional inputs ($d = 2$ or 3) and modest sample sizes, so reporting detailed compute resources is not necessary.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We included it at the end of the paper.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work does not pose any risks.

    Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have properly cited papers and resources used in our experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: We do not have experiments involving crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: We do not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.