重新思考带有 Sigmoid 门控函数的多项逻辑专家混合模型

Tuan Minh Pham*, ♢ Thinh Cao*, ◇ Viet Nguyen*, ◇
Huy Nguyen† Nhat Ho**, † Alessandro Rinaldo**, †

†德克萨斯大学奥斯汀分校

♢普渡大学

◇河内理工大学

2026年2月3日

arXiv:2602.01466v1 [stat.ML] 1 Feb 2026

摘要

在混合专家（MoE）模型中，Sigmoid 门控函数已通过实证证明，在从近似前馈网络到语言建模的多种任务中，表现优于 softmax 门控。此外，近期研究表明，在回归设定下，Sigmoid 门控在理论上比对应的 softmax 门控具有更高的样本效率。然而，文献中仍有三个值得关注的问题尚未得到解决：(i) 尚未在分类设定下证实 Sigmoid 门控的优势；(ii) 现有的 Sigmoid门控混合专家模型可能无法收敛至真实值；(iii) Sigmoid 门控中温度参数的影响在理论上仍缺乏充分研究。为解决这些开放问题，我们对采用修正的 Sigmoid 门控的多项逻辑专家混合进行了全面分析，以确保模型收敛。结果表明，无论是参数估计还是专家估计，Sigmoid 门控的样本复杂度均低于 Softmax 门控。此外，我们发现，在 Sigmoid 门控中引入温度会导致样本复杂度呈指数级增长，这是由温度与门控参数之间的内在相互作用造成的。为克服这一问题，我们提出在门控函数中以欧几里得评分替代传统的内积评分，从而有效消除这种相互作用，并将样本复杂度显著降低至多项式级别。

1 引言

首次由 [19]，专家混合（MoE）是一种集成框架，通过自适应门控机制整合多个专业化子模型。这些子模型称为专家，可以采用多种形式，包括分类器 [6, 38]，回归模型 [21, 14]，或前馈神经网络（FFN） [23, 20]。门控机制负责根据输入值确定各个专家的贡献或权重，使专家能够专注于输入空间的不同区域或任务的不同方面。近期， [34]提出了专家混合架构的稀疏版本，其中每个输入只会被路由到少数相关专家，而非所有专家。由于这种稀疏专家激活为提升模型容量、同时避免计算量同比例增加提供了切实可行的途径，专家混合在大规模应用中重新受到关注，具体包括大语言模型 [15, 16, 1, 32]，计算机视觉 [25, 33]，

*同等贡献， **共同末位作者。

多模态学习 [18, 39], 领域泛化 [28, 24], 迁移学习 [13, 22, 35], 等领域。

尽管稀疏混合专家模型取得了成功， [7] 指出，这类模型会面临一种称为表征坍塌的新挑战：部分专家主导计算，而其他专家则未得到充分利用。这一问题归因于广泛使用的 softmax 门控所引发的专家竞争，即一个专家的权重增加时，其他专家的权重会自动降低。为此， [7] 选择采用带有可学习温度参数的 Sigmoid 门控，作为应对表征坍塌的稳健替代方案。随后， [8] 通过实验证明，在逼近两层前馈神经网络的任务中，Sigmoid 门控的表现优于 Softmax 门控。这一发现与 DeepSeek 语言模型技术报告中的结果一致：采用 sigmoid 门控混合专家模型的 DeepSeek-v3 [11] 优于采用 softmax门控混合专家模型的 DeepSeek-v2 [10] 。

从理论角度来看，文献中已有少数研究尝试阐明 Sigmoid 门控相较于 Softmax 门控的优势。 [30] 评估了一种回归框架的样本复杂度，其中回归函数采用 sigmoid 门控混合专家模型的形式。他们证明，与采用 Softmax 门控相比，采用 Sigmoid 门控时，只需更少的数据，就能以给定误差逼近专家。然而，仍有几个重要问题尚未得到解决。首先，上述 Sigmoid 门控的优势仅在回归任务中得到验证，其在分类任务中的优势尚未得到论证。其次，当专家数量未知且设定过多时（这种情况很常见），由于 Sigmoid 结构的内在特性，这些研究所考察的 sigmoid 门控混合专家模型无法收敛到真实模型（见 [30]中的推论 1）。最后， [7] 提出的温度参数所产生的影响尚未得到研究。

主要贡献。 本文针对上述问题展开研究，并作出以下贡献。

揭示用于分类的混合专家模型中的 *Sigmoid* 门控机制。 我们研究*Sigmoid* 门控多项逻辑专家混合（*MLMoE*）模型的理论性质。我们推导出参数估计和专家估计的收敛速率，并证明 Sigmoid 门控比文献 [27]研究的 Softmax 门控具有更高的样本效率。

用于 *MLMoE* 收敛的修正 *Sigmoid* 门控。 为确保在过度指定情形下（即部署的专家数量多于真实数量时）MLMoE 模型收敛到其真实值，我们设计了一种修正的 Sigmoid 门控，其中 Sigmoid 函数乘以一个正数常量。我们在第 3节给出 MLMoE 模型及其参数的收敛速率。

带温度参数的 *Sigmoid* 门控。 在第4节中，我们研究温度对模型样本复杂度的影响。我们的结果表明，使用标准内积亲和度分数时，要将参数或专家近似到给定精度，需要指数级数量的数据点。这主要是由于温度与门控参数之间存在内在相互作用。为缓解这一问题，我们建议改用欧几里得亲和度分数，此时样本复杂度可证明降至多项式级别。

在第 5 节中，我们介绍了若干数值实验，以实证检验我们的理论结果。最后，在第 6 节中，我们对论文进行了深入讨论，包括这些理论可能带来的实际意义、我们工作的局限性，以及一些潜在的研究方向。

表 1： Voronoi 型度量下多项逻辑混合专家模型（MLMoE）的参数估计速率汇总。对于 softmax 门控的 MLMoE，我们区分专家截距参数的两种情形 $b^*_{j\ell}$（见 [27]）：机制 *1* 要求 $\forall j \in [k^*],\ \exists \ell \in [K-1]$ 使得 $b^*_{j\ell} \neq 0$; 而情形*2* 假设 $\exists j \in [k^*]$ 使得 $b^*_{j\ell} = 0\ \forall \ell \in [K-1]$。对于 sigmoid 门控，我们展示不同的门控机制：无温度、带内积亲和度的温度机制，以及带欧几里得亲和度的温度机制。结果分别报告了精确指定参数和过度指定参数的估计速率。Voronoi 度量分别为 softmax 下 $\mathcal{D}_2$ in [27]，以及相应 sigmoid 设置下的 $\mathcal{D}_1$、$\mathcal{D}_{2、2}$ 和 $\mathcal{D}_3$。

门控函数	情形	精确指定专家	过度指定专家	定理
softmax	机制 1	$\mathcal{O}_P(\lfloor\log(n)/n\rfloor^{\frac{1}{2}})$	$\mathcal{O}_P(\lfloor\log(n)/n\rfloor^{\frac{1}{4}})$	[27]
	情形2	慢于任何多项式速率		[27]
sigmoid，无温度	All	$\mathcal{O}_P(\lfloor\log(n)/n\rfloor^{\frac{1}{2}})$	$\mathcal{O}_P(\lfloor\log(n)/n\rfloor^{\frac{1}{4}})$	定理 1
Sigmoid 函数、温度与内积		慢于任何多项式速率		定理 2
Sigmoid 函数、温度与欧几里得		$\mathcal{O}_P(\lfloor\log(n)/n\rfloor^{\frac{1}{2}})$	$\mathcal{O}_P(\lfloor对数(n)/n\rfloor^{\frac{1}{4}})$	定理 3

证明和其他结果见附录。

记号。对于任意 $n \in \mathbb{N}$，令 $[n] := \{1, 2, \ldots, n\}$。对于任意集合 S，$|S|$ 表示其基数。对于向量 $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$，$\|v\|$ 表示其欧几里得范数。对于多重指标 $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$，记 $|\alpha| = \sum_{i=1}^d \alpha_i$, $\alpha! = \alpha_1! \cdots \alpha_d!$，并且对于 $v \in \mathbb{R}^d$，$v^\alpha = v_1^{\alpha_1} \cdots v_d^{\alpha_d}$。对于正数序列 (a_n) 和 (b_n)，如果对于所有 n 及某个普适常数$C > 0$，$a_n \leq C b_n$，则记作 $a_n = \mathcal{O}(b_n)$ 或 $a_n \lesssim b_n$。对于随机正数序列 (t_n) 和 (s_n)，若 t_n/s_n 是随机有界的，则记作 $t_n = \mathcal{O}_P(s_n)$：对于任意 $\epsilon > 0$，存在 $M > 0$，使得对于所有充分大的 n，$\mathbb{P}(t_n/s_n > M) < \epsilon$。最后，对于由勒贝格测度 μ支配的任意两个概率密度 p 和 q，定义全变差距离为 $d_V(p, q) := \frac{1}{2}\int |p - q| \mathrm{d}\mu$，海林格距离为$d_H(p, q) := \left(\frac{1}{2}\int |\sqrt{p} - \sqrt{q}|^2 \mathrm{d}\mu\right)^{1/2}$。

2 相关工作

越来越多的文献开始研究采用不同似然函数和门控机制的混合专家模型。早期理论研究针对特定形式建立了一致性：[5] 通过发现期望最大化内循环中的不稳定性，并提出一种使用精确海森矩阵的牛顿-拉夫森方法，改进了多分类混合专家的学习算法；[31] 则提出了一种用于分类的函数型混合专家框架，该框架利用稀疏正则化来识别数据中的关键特征。近期，[27] 针对 Softmax 门控多项逻辑混合专家模型建立了一般收敛理论，并表明门控参数与专家参数之间的内在相互作用从根本上决定了估计速率。作为对这些分析的补充，[2] 对用于学习大规模 Softmax 混合模型的 EM 算法进行了全面的理论分析。

从深度学习的角度来看，专家混合架构因其表达能力和可扩展性而受到广泛研究。[6] 对神经网络中的混合专家层进行了理论分析，刻画了其逼近能力和基于梯度的训练动态。[37] 分析了混合专家模型在结构化任务中的表达能力，表明专家组合能够高效地表示层次模式。[3]证明，更细的专家粒度通过提升专门化程度，能够显著提高表征效率。在大语言模型中，DeepSeek-MoE [9, 29] 通过稀疏激活模式的实证结果表明，共享专家隔离能够同时提升

性能和计算效率 .

尽管这些研究极具启发性且颇具价值，但它们对混合专家模型的统计挑战探讨有限，仍存在若干理论空白。具体而言，尚未在多项逻辑分类框架下系统研究Sigmoid门控混合专家模型，也尚未证明过度指定情形下混合测度的收敛。此外，温度缩放在sigmoid门控下的统计作用仍未得到充分理解。我们将在后续章节中探讨这些问题。

3 修正的Sigmoid门控函数

本节将分析采用修正的Sigmoid门控函数的MLMoE模型中条件密度和模型参数的收敛性。

问题设定。 假设数据 $\{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^d \times [K]$ 以独立同分布的方式从*Sigmoid*门控*MLMoE*模型中抽取，该模型的条件密度函数对任意 $s \in [K]$ 定义如下:

$$p_{G^*}(Y = s \mid X) := \sum_{i=1}^{k^*} \frac{\exp(\gamma_i^*)\,\sigma\big((\alpha_i^*)^\top X + \beta_i^*\big)}{\sum_{j=1}^{k^*} \exp(\gamma_j^*)\,\sigma\big((\alpha_j^*)^\top X + \beta_j^*\big)} \cdot f(Y = s|X; a_i^*, b_i^*), \tag{1}$$

其中， $t \in \mathbb{R} \mapsto \sigma(t) := \frac{1}{1+\exp(-t)}$ 是Sigmoid函数， k^* 表示真实专家的数量。上文将 $\exp(\gamma_i^*)$ 项纳入标准 Sigmoid 门控，并将其称为修正的Sigmoid门控函数；我们将在本节后文对此作进一步说明。接下来，对于每个$i \in [k^*]$，第 i 个专家采用多项逻辑函数的形式:

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

其中 $a_i^* := (a_{i1}^*, \ldots, a_{iK}^*) \in \mathbb{R}^{d\times K}$与 $b_i^* := (b_{i1}^*, \ldots, b_{iK}^*) \in \mathbb{R}^K$为专家参数。相比之下，我们将 $(\gamma_i^*, \alpha_i^*, \beta_i^*)_{i\in[k^*]} \subset \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$ 称为门控参数。最后， $G^* := \sum_{i=1}^{k^*} \exp(\gamma_i^*)\,\delta_{(\alpha_i^*, \beta_i^*, a_i^*, b_i^*)}$ 是真实但未知的混合测度，由狄拉克测度 δ的混合构成，这些狄拉克测度与专家参数和门控参数 $(\gamma_i^*, \alpha_i^*, \beta_i^*, a_i^*, b_i^*) \in \Theta \subset \mathbb{R} \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^{d\times K} \times \mathbb{R}^K$相关联，对所有 $i \in [k^*]$，其中 Θ 表示参数空间。

最大似然估计（MLE）。我们通过估计真实混合测度来估计模型参数 G^* ，使用最大似然方法。然而，由于真实专家的数量 k^* 在实践中通常是未知的，我们将考虑一种过度指定的设定，在该设定下，我们用一个包含 k 个专家的混合模型（ $k > k^*$）来拟合模型，然后考虑最大似然估计

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

其中， $\mathcal{G}_k(\Theta)$表示所有至多包含 k 个分量、形式为$G = \sum_{i=1}^{k'} \exp(\gamma_i)\delta_{(\alpha_i,\beta_i,a_i,b_i)}$ 的混合测度所构成的集合，其中 $1 \le k' \le k$ 和（γ_i、 α_i、 β_i、 a_i、 b_i） $\in \Theta$。

修正的 Sigmoid 门控的优势。 引入$\exp(\gamma_i)$ 这一项对于准确刻画条件密度估计的收敛至关重要。具体而言，由于我们拟合真实值 MLMoE模型（1），采用 $k > k^*$ 个专家的混合，某些真实值参数必须

$\widehat{\ }$ $\widehat{\ }$至少由两个分量拟合。例如，假设 $(\widehat{\alpha}_i^n,\ \beta_i^n,\ \widehat{a}_i^n,\ b_i^n) \to (\alpha_1^*,\ \beta_1^*,\ a_1^*,\ b_1^*)$ 依概率成立，对于 $i \in \{1,2\}$。为使 $p_{\widehat{G}_n}$ 收敛到 p_{G^*} 成立，我们必须有$\sum_{i=1}^2 \exp(\widehat{\gamma}_i^n)\sigma\left((\widehat{\alpha}_i^n)^\top X + \widehat{\beta}_i^n\right) \to \exp(\gamma_i^*)\sigma\left((\alpha_i^*)^\top X + \beta_i^*\right)$，对几乎必然的 X。若缺少 $\exp(\gamma_i)$，则前述收敛不再成立。因此，正项 $\exp(\gamma_i)$ 的引入在条件密度收敛中起着至关重要的作用。

在介绍本节的主要结果之前，我们先提出一些贯穿全文的必要且温和的假设，除非另有说明。

假设 。在我们的分析中，我们对模型参数作出四个主要假设 .
(A.1) 参数空间 Θ 是 $\mathbb{R} \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^{d\times K} \times \mathbb{R}^K$ 的紧子集，且输入空间 $\mathcal{X}$ 有界；**(A.2)** $(a_1^*,\ b_1^*)$ .. . ,$(a_{k^*}^*,\ b_{k^*}^*)$ 是不同的参数；**(A.3)** $a_{iK}^* = \mathbf{0}_d$, $b_{iK}^* = 0$，对所有 $i \in [k^*]$ 均成立，且 $\gamma_{k^*} = 0$；
(A.4) $\alpha_1^*, \ldots, \alpha_{k^*}^*$ 中至少有一个非零。

第一个假设是最大似然估计收敛所必需的。其次，第二个假设确保存在不同专家；它与第三个假设共同使 Sigmoid门控混合专家模型具有可识别性。最后一个假设则确保门控值依赖于输入。

现在，我们开始分析。首先，我们在如下命题中证明 Sigmoid 门控 MLMoE 模型是可辨识的，其证明见附录B.2。

命题1. （可识别性）假设 G 与 G' 是 $\mathcal{G}_k(\Theta)$中的两个混合测度。若$p_G(Y|X) = p_{G'}(Y|X)$ 以概率一成立（关于 X 与 Y的联合分布），则$G \equiv G'$。

$\widehat{\ }$该结果表明，最大似然估计 G_n 收敛到真实混合测度 G^* 可由密度估计量 $p_{\widehat{G}_n}$ 收敛到真实密度 p_{G^*}推出。接下来，我们推导后者的收敛速率。

命题 2. （密度估计速率）密度估计量 $p_{\widehat{G}_n}(Y|X)$ 在海林格距离下收敛到真实密度 $p_{G^*}(Y|X)$ 的速率为

$$\mathbb{E}_X[d_H(p_{\widehat{G}_n}(\cdot|X), p_{G^*}(\cdot|X))] = \mathcal{O}_P(\sqrt{Kdk\log(n)/n}).$$

命题2的证明见附录B.1。上述界表明，在保持 K、 k 和d固定的情况下，真实密度的估计速率相对于样本量而言是参数速率。因此，如果我们能够构造一个参数间的度量 $\mathcal{D}$ ，使其满足不等式 $\widehat{\ }$ $\sim\mathbb{E}_X[d_H(p_{\widehat{G}_n}(\cdot|X)\ p_{G^*}(\cdot|X))] \gtrsim \mathcal{D}(G_n, G^*)$，那么最大似然估计 G_n也将以参数速率收敛到真实混合测度 G^* 。为此，我们采用文献 [26]中引入的 Voronoi 损失函数概念。

Voronoi 损失。假设混合测度 G 有 k' 个原子 $\theta_i := (\alpha_i,\ \beta_i,\ a_i,\ b_i)$。我们将这些原子分配到由 G^* 的分量 $\theta_j^* := (\alpha_i^*,\ \beta_i^*,\ a_i^*,\ b_i^*)$ 所诱导的 Voronoi 单元 $\mathcal{A}_j := \mathcal{A}_j(G)$ 中，这些单元定义如下：

$$\mathcal{A}_j := \{i \in [k'] : \|\theta_i - \theta_j^*\| \leq \|\theta_i - \theta_\ell^*\|, \forall\, \ell \neq j\}. \tag{3}$$

接下来，我们定义 $R_{ij}(\kappa_1,\ \kappa_2,\ \kappa_3,\ \kappa_4) := \|\Delta\alpha_{ij}\|^{\kappa_1} + |\Delta\beta_{ij}|^{\kappa_2} + \sum_{\ell=1}^K \left(\|\Delta a_{ij\ell}\|^{\kappa_3} + |\Delta b_{ij\ell}|^{\kappa_4}\right)$, 其中$\{\kappa_l, l = 1, \ldots, 4\}$ 为非负整数， $\Delta\alpha_{ij} := \alpha_i - \alpha_j^*$, $\Delta\beta_{ij} := \beta_i - \beta_j^*$, $\Delta a_{ij\ell} := a_{i\ell} - a_{j\ell}^*$,

以及 $\Delta b_{ij\ell} := b_{i\ell} - b^*_{j\ell}$。我们考虑如下 Voronoi 损失：

$$\mathcal{D}_1(G, G^*) := \sum_{j=1}^{k^*} \left| \sum_{i \in \mathcal{A}_j} \exp(\gamma_i) - \exp(\gamma_i^*) \right| + \sum_{\substack{j:|\mathcal{A}_j|>1, \\ i \in \mathcal{A}_j}} \exp(\gamma_i)\, R_{ij}(2,2,2,2) + \sum_{\substack{j:|\mathcal{A}_j|=1, \\ i \in \mathcal{A}_j}} \exp(\gamma_i)\, R_{ij}(1,1,1,1). \tag{4}$$

上述 Voronoi 损失函数使我们能够根据以下下界，刻画 sigmoid 门控 MLMoE 模型中参数估计与专家估计不同的收敛行为。

定理 1. 对于所有混合测度 $G \in \mathcal{G}_k(\Theta)$，均有

$$\mathbb{E}_X \left[d_H\big(p_G(\cdot \mid X),\, p_{G^*}(\cdot \mid X)\big) \right] \gtrsim \mathcal{D}_1(G, G^*).$$

因此，将该不等式与命题 2 中的密度估计速率结合，可得

$$\mathcal{D}_1(\widehat{G}_n, G^*) = \mathcal{O}_P\left(\sqrt{Kdk\log(n)/n}\right).$$

定理 1 的证明见附录 A.1。根据 $\mathcal{D}_1$ 的构造可知，精确设定的参数 $(\alpha_j^*, \beta_j^*, a_j^*, b_j^*)$（即其 Voronoi 单元 $\mathcal{A}_j$ 中仅包含一个元素）可以达到参数速率 $\mathcal{O}_P([\log(n)/n]^{\frac{1}{2}})$，同时固定其他所有参数。我们回顾一下，每个专家 $f(\cdot|X;\, a_j^*, b_j^*)$ 都是一个 Lipschitz 函数。令 $G_n := \sum_{i=1}^{k_n} \exp(\widehat{\gamma}_i^n)\, \delta_{(\widehat{\alpha}_i^n, \widehat{\beta}_i^n, \widehat{a}_i^n, \widehat{b}_i^n)}$，则可得

$$\sup_x |f(\cdot|x;\widehat{a}_i^n, \widehat{b}_i^n) - f(\cdot|x;a_j^*, b_j^*)| \leq L_1 \left(\|\widehat{a}_i^n - a_j^*\| + \|\widehat{b}_i^n - b_j^*\| \right) \lesssim \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}}), \tag{5}$$

对于任意 $i \in \mathcal{A}_j$，其中 L_1 是 Lipschitz 常数。上述界表明，专家 $f(\cdot|X;\, a_j^*, b_j^*)$ 同样具有阶为 $\mathcal{O}_P([\log(n)/n]^{\frac{1}{2}})$ 的参数估计速率。与此同时，Voronoi 单元 $\mathcal{A}_j$ 中包含多个元素的过度指定参数，其估计速率较慢，阶为 $\mathcal{O}_P([\log(n)/n]^{\frac{1}{4}})$。通过参照公式 (5) 进行论证，我们推导出：如果专家 $f(\cdot|X;\, a_j^*, b_j^*)$ 至少由两个专家拟合，那么其估计速率为 $\mathcal{O}_P([\log(n)/n]^{\frac{1}{4}})$。

Sigmoid 门控相较于 Softmax 门控的优势。对于 Softmax 门控 MLMoE 模型，[27]证明了，如果存在 $j \in [k^*]$ 使得 $a_{j\ell}^* = 0$，对于任意 $\ell \in [K-1]$，那么门控参数 α_j^* 将通过某些偏微分方程与专家参数 $a_{j\ell}^*$ 发生交互。值得注意的是，由于这种交互，参数估计速率比任何多项式速率都慢，例如 $\mathcal{O}_P(n^{-1/(2r)})$，对于任意 $r \geq 1$。相比之下，使用 sigmoid 门控函数时，这种交互不再存在。因此，参数和专家估计速率显著提升至多项式阶，如定理1所示，并总结于表1。因此，我们认为，在 MLMoE 模型中，Sigmoid 门控比 Softmax 门控更具样本效率。

4 带温度参数的 SigmoidGating函数

本节研究在 MLMoE 模型中使用带温度参数的 sigmoid 门控函数时，参数估计的收敛性质。在第4.1节中，我们首先证明，直接将温度应用于内积亲和度分数会导致内在参数交互，从而使参数估计的收敛速率显著变慢。为此，我们在第4.2 节提出使用一种新颖的欧几里得亲和度分数，并证明它确实能提高参数估计速率。

4.1 内积亲和度分数

首先，我们给出将温度参数应用于标准内积亲和度分数的 sigmoid 门控多层混合专家模型的形式。

问题设定。 设数据 $\{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^d \times [K]$ 是来自 sigmoid 门控多层混合专家模型的独立同分布样本，其条件概率密度对任意 $s \in [K]$ 均为，

$$\widetilde{p}_{\widetilde{G}^*}(Y = s|X) := \sum_{i=1}^{k^*} \frac{\exp(\gamma_i^*)\sigma(\frac{(\alpha_i^*)^\top X+\beta_i^*}{\tau^*})}{\sum_{j=1}^{k^*}\exp(\gamma_j^*)\sigma(\frac{(\alpha_j^*)^\top X+\beta_j^*}{\tau^*})} \cdot \frac{\exp((a_{is}^*)^\top X + b_{is}^*)}{\sum_{\ell=1}^{K}\exp((a_{i\ell}^*)^\top X + b_{il}^*)}, \tag{6}$$

这里，$\widetilde{G}^* := \sum_{i=1}^{k^*}\exp(\gamma_i^*)\delta_{(\alpha_i^*,\beta_i^*,\tau^*,a_i^*,b_i^*)}$ 表示与参数 相关联的真实但未知的 混合测度 $(\alpha_i^*, \beta_i^*, \tau^*, a_i^*, b_i^*) \in \Upsilon$，其中 Υ 是 $\mathbb{R}^d \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{d\times K} \times \mathbb{R}^K$的紧子集。在模型的门控网络中，线性层的输出会先按温度进行缩放，再经过 Sigmoid 函数。除非另有说明，我们沿用第 3节中关于参数的假设。

最大似然估计。的最大似然估计为 $\widetilde{G}^*$，其表达式为

$$\widetilde{G}_n \in \arg\max_{G\in\mathcal{G}_k(\Upsilon)} \sum_{i=1}^{n} \log(\widetilde{p}_G(Y_i|X_i)), \tag{7}$$

其中，$\widetilde{p}_G$ 具有方程 (6) 中指定的形式，且 $\mathcal{G}_k(\Upsilon)$ 表示所有至多包含 $k \geq k^*$ 个形如 $G = \sum_{i=1}^{k'} \exp(\gamma_i)\delta_{(\alpha_i,\beta_i,\tau,a_i,b_i)}$ 的分量的混合测度构成的集合，其中 $1 \leq k' \leq k$，且 $(\alpha_i, \beta_i, \tau, a_i, b_i) \in \Upsilon$。值得注意的是，采用内积亲和度分数的温度 sigmoid 门控混合专家模型仍然是可辨识的。换言之，如果密度估计 $\widetilde{p}_{\widetilde{G}_n}$ 收敛到真实密度 $\widetilde{p}_{\widetilde{G}^*}$，那么最大似然估计 $\widetilde{G}_n$ 也会收敛到其真实对应值 $\widetilde{G}^*$。与命题 2 类似，在海林格距离下，估计密度 $\widetilde{p}_{\widetilde{G}_n}(Y|X)$ 以 $\mathcal{O}_P(\sqrt{Kdk\log(n)/n})$ 的阶收敛到真实密度~~$\widetilde{p}_{\widetilde{G}^*}(Y|X)$~~。

为了根据上述密度估计速率，推导方程（6）中模型的参数估计速率，我们将密度差异 $\widetilde{p}_{\widetilde{G}_n}(\cdot|X) - \widetilde{p}_{\widetilde{G}^*}(\cdot|X)$ 分解为线性无关分量之和。该分解通过对 Sigmoid 分子与专家函数 $\widetilde{u}(Y = s|X; \alpha, \beta, \tau, a, b)$ 的乘积进行泰勒展开来实现：$= \sigma(\frac{\alpha^\top X+\beta}{\tau})f(Y = s|X; a, b)$。遗憾的是，在 Sigmoid 门控中使用温度 τ 时，它与门控参数 α, β 之间存在内在交互，这种交互由以下偏微分方程刻画：

$$\frac{\partial\widetilde{u}}{\partial\tau} = -\frac{1}{\tau}\cdot\alpha^\top\frac{\partial\widetilde{u}}{\partial\alpha} - \beta\frac{\partial\widetilde{u}}{\partial\beta}. \tag{8}$$

因此，我们构造 Voronoi 损失函数，以分析这种交互如何影响定理2中的参数估计速率。

Voronoi 损失函数。对于任意混合测度 G in $\mathcal{G}_k(\Upsilon)$ 以及非负整数 $\{\kappa_l, l = 1, \ldots, 5\}$，采用上一节引入的记号，定义 $K_{ij}(\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5) := \|\Delta\alpha_{ij}\|^{\kappa_1} + |\Delta\beta_{ij}|^{\kappa_2} + |\Delta\tau|^{\kappa_3} + \sum_{\ell=1}^{K}\left(\|\Delta a_{ij\ell}\|^{\kappa_4} + |\Delta b_{ij\ell}|^{\kappa_5}\right)$，其中记 $\Delta\tau := \tau - \tau^*$。于是，对于每个$r \geq 1$，我们关注的 Voronoi 损失函数为

$$\mathcal{D}_{2,r}(G, \widetilde{G}^*) := \sum_{j=1}^{k^*}\left|\sum_{i\in\mathcal{A}_j}\exp(\gamma_i) - \exp(\gamma_i^*)\right| + \sum_{j=1}^{k^*}\sum_{i\in\mathcal{A}_j}\exp(\gamma_i)K_{ij}(r, r, r, r, r). \tag{9}$$

偏微分方程 (8) 会产生显著且不理想的影响：海林格距离 $\mathbb{E}_X[d_H(\widetilde{p}_{\widetilde{G}_n}(\cdot|X)\ \widetilde{p}_{\widetilde{G}^*}(\cdot|X))]$ 不再有下界 $\mathcal{D}_{2,r}(G_n, G^*)$。因此，我们无法使用上述论证得出 $\mathcal{D}_{2,r}(\widetilde{G}_n, \widetilde{G}^*) = \mathcal{O}_P(n^{-1/2})$，如定理 1 中所述。相反，我们在附录 A.2 中证明

$$\inf_{G\in\mathcal{G}_k(\Upsilon):\mathcal{D}_{2,r}(G,\widetilde{G}^*)\leq\varepsilon}\frac{\mathbb{E}_X[d_V(\widetilde{p}_G(\cdot|X), \widetilde{p}_{\widetilde{G}^*}(\cdot|X))]}{\mathcal{D}_{2,r}(G, \widetilde{G}^*)} \to 0,$$

当 $\varepsilon \to 0$ 对任意 $r \geq 1$。
上述结果意味着参数估计的以下极小极大下界：

定理 2. 在过度设定情形下，对于任意 $r \geq 1$，估计 G^*的极小极大下界如下：

$$\inf_{G_n\in\mathcal{G}_k(\Upsilon)}\sup_{G\in\mathcal{G}_k(\Upsilon)\setminus\mathcal{G}_{k^*-1}(\Upsilon)}\mathbb{E}_{\widetilde{p}_G}[\mathcal{D}_{2,r}(G_n, G)] \gtrsim n^{-1/2}.$$

此处，记号 $\mathbb{E}_{\widetilde{p}_G}$ 表示关于混合密度为 p_G^n的乘积测度所取的期望。

定理 2 的证明见附录 A.2。上述极小极大下界与方程（$\mathcal{D}_{2,r}$）中的表述9表明，参数 $\alpha_j^*, \beta_j^*, \tau^*, a_j^*, b_j^*$的估计速率慢于 $\mathcal{O}_P(n^{-1/(2r)})$，对任意 $r \geq 1$均如此。这意味着这些速率慢于任何多项式速率，衰减速度甚至可能慢至 $\mathcal{O}_P(1/\log(n))$。

采用方程（5后描述的 Lipschitz 函数的相同论证可知，估计专家 $f(\cdot|X; a_j^*, b_j^*)$ 的速率不会优于参数a_j^*和 b_j^*的估计速率，且也可能慢至 $\mathcal{O}_P(1/\log(n))$。这一结果表明，在 MLMoE 模型的标准 Sigmoid 门控中加入温度项并非样本高效，因为正如我们的分析所示，这可能会产生不良影响。

4.2 欧几里得亲和度得分

为消除 PDE（8）中观察到的内积亲和度分数与温度参数之间不理想的相互作用，我们提出用一种新颖的欧几里得亲和度分数替代它。我们证明，这种欧几里得亲和度公式可确保温度 sigmoid 门控混合专家模型具有多项式阶的参数估计速率。该模型的公式定义如下：

问题设定。 同样地，我们假设数据 $\{(X_i,\ Y_i)\}_{i=1}^n \subset \mathbb{R}^d \times [K]$ 独立同分布，但这次数据来自一个修正的 Sigmoid 门控多项逻辑专家混合模型，其条件密度函数为，对于每个 $s \in [K]$,

$$\overline{p}_{\overline{G}^*}(Y=s|X) := \sum_{i=1}^{k^*} \frac{\exp(\gamma_i^*)\sigma(\frac{\|\alpha_i^* - X\| + \beta_i^*}{\tau^*})}{\sum_{j=1}^{k^*}\exp(\gamma_j^*)\sigma(\frac{\|\alpha_j^* - X\| + \beta_j^*}{\tau^*})} \cdot \frac{\exp((a_{is}^*)^\top X + b_{is}^*)}{\sum_{\ell=1}^{K}\exp((a_{i\ell}^*)^\top X + b_{il}^*)}, \tag{10}$$

通常， $G^* := \sum_{i=1}^{k^*} \exp(\gamma_i^*)\delta_{(\alpha_i^*, \beta_i^*, \tau^*, a_i^*, b_i^*)}$ 表示真实但未知的混合测度，它是与真实参数（$\alpha_i^*, \beta_i^*, \tau^*, a_i^*, b_i^*$） $\in \Upsilon$相关联的狄拉克测度 δ的组合。记$u(X;\ \alpha, \beta, \tau, a, b) := \sigma\left(\frac{\|\alpha - X\| + \beta}{\tau}\right) f(Y=s|X;\ a, b)$ 。那么，欧几里得亲和度分数的一阶独立性条件保证了门控参数 α_i, β_i 与温度 τ 之间不再存在相互作用，即

$$\frac{\partial \overline{u}}{\partial \tau} \neq -\frac{1}{\tau} \cdot \alpha^\top \frac{\partial \overline{u}}{\partial \alpha} - \beta \frac{\partial \overline{u}}{\partial \beta}. \tag{11}$$

因此，与第4.1节中的结果相比，参数 $\alpha_j^*, \beta_j^*, \tau^*, a_j^*, b_j^*$的估计速率应有所提高。

最大似然估计。 根据门控函数的变化，我们重新定义模型（10）对应的最大似然估计为

$$\overline{G}_n := \arg\max_{G \in \mathcal{G}_k(\Upsilon)} \sum_{i=1}^{n} \log(\overline{p}_G(Y|X)).$$

沿用内积亲和度分数情形下的推理过程，条件密度 $p_G(\cdot|X)$ 仍然是可辨识的。此外，密度估计量 $p_{G_n}(\cdot|X)$ 收敛到 $p_{G^*}(\cdot|X)$ 的速率仍为 $\mathcal{O}_P(\sqrt{Kdk\log(n)/n})$。基于这一观察，如果我们能够构造一个参数间的损失函数 $\mathcal{D}_3$，使得 $\mathbb{E}_X[d_H(p_{G_n}(\cdot|X), p_{G^*}(\cdot|X))] \gtrsim \mathcal{D}_3(G_n,\ G^*)$，那么最大似然估计 G_n 也将以 $\mathcal{O}_P(n^{-1/2})$ 的速率收敛到真实混合测度 G^* 。

Voronoi 损失。 我们提出 Voronoi 损失

$$\begin{aligned}\mathcal{D}_3(G, \overline{G}^*) := &\sum_{j=1}^{k^*}\left|\sum_{i \in \mathcal{A}_j}\exp(\gamma_i) - \exp(\gamma_i^*)\right| \\ &+ \sum_{j:|\mathcal{A}_j|>1,\ i\in\mathcal{A}_j} \exp(\gamma_i)K_{ij}(2,2,2,2,2) + \sum_{j:|\mathcal{A}_j|=1,\ i\in\mathcal{A}_j}\exp(\gamma_i)K_{ij}(1,1,1,1,1).\end{aligned} \tag{12}$$

在以下定理中，我们给出温度 sigmoid 门控 MLMoE 模型在欧几里得亲和度分数下的参数收敛速率。

定理 3。 对于任意 $G \in \mathcal{G}_k(\Upsilon)$，以下海灵格下界成立：

$$\mathbb{E}_X[d_H(\overline{p}_G(\cdot|X), \overline{p}_{\overline{G}^*}(\cdot|X))] \gtrsim \mathcal{D}_3(G, \overline{G}^*).$$

该界给出了最大似然估计的参数收敛速率： $\mathcal{D}_3(G_n, G^*) = \mathcal{O}_P(\sqrt{Kdk\log(n)/n})$。

定理 3 的证明见附录 A.3。该结果表明，与单个拟合分量 ($|\mathcal{A}_j(G_n)| = 1$) 相关的参数 $\alpha_j^*, \beta_j^*, \tau^*, a_j^*, b_j^*$ 的估计速率为 $\mathcal{O}_P([\log(n)/n]^{\frac{1}{2}})$ 阶。相比之下，由多个分量 ($|\mathcal{A}_j(G_n)| > 1$) 近似的参数收敛得更慢，速率为 $\mathcal{O}_P([\log(n)/n]^{\frac{1}{4}})$。这些速率与定理 1 中建立的、无温度参数的 Sigmoid 门控多层混合专家模型的速率一致。

此外，沿用公式（5）之后的相同论证，我们得到，专家估计量在采用单个专家拟合时也以 $\mathcal{O}_P$（$[\log（n）/n]^{\frac{1}{2}}$）的速率收敛，而在采用多个专家进行近似时，则以 $\mathcal{O}_P$（$[\log（n）/n]^{\frac{1}{4}}$）的速率收敛。因此，用我们提出的欧几里得亲和度分数替代标准的内积亲和度分数，可显著提高专家估计速率，使其从约 $\mathcal{O}_P（1/\log（n）$）提升至多项式阶。这凸显了欧几里得亲和度分数相较于内积运算的优势。

5 数值实验

本节将通过数值实验，实证展示关于 sigmoid 门控多层混合专家模型收敛行为的理论结果。

合成数据。 对于每个样本量 n，我们生成独立同分布的样本$\{(X_i, Y_i)\}_{i=1}^n$：首先从 $[0,1]$ 上的均匀分布中抽取 X_i，然后从真实条件密度 $p_{G^*}(Y = s|X)$、$\widetilde{p}_{\widetilde{G}^*}(Y = s|X)$ 和 $p_{G^*}(Y = s|X)$ 中抽取 Y_i，这些条件密度分别对应于每个定理配置中指定的混合专家模型。

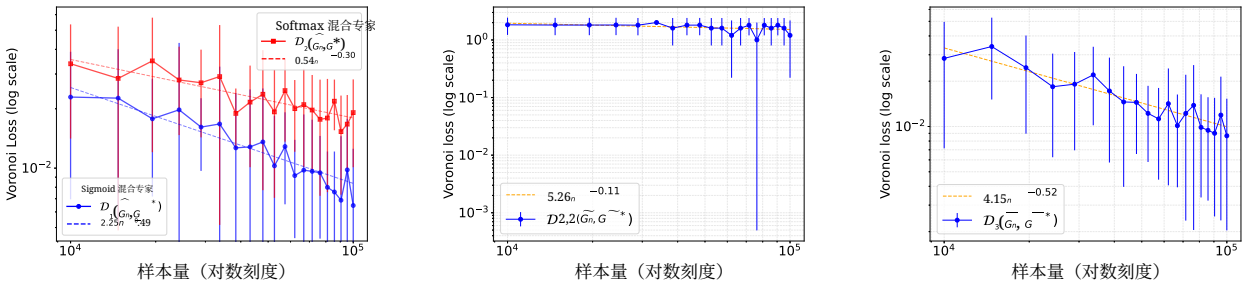
最大似然估计。 计算每组样本的最大似然 $\widehat{\ }$ $\widetilde{\ }$估计量 G （或 G，G）的一种广泛使用的方法是采用期望最大化（EM）算法[12]。由于最大化步骤中无法获得门控参数的闭式更新公式，我们改为采用此前在 [4]中使用过的基于 EM 的数值过程。我们将收敛容差设为 $\varepsilon = 10^{-6}$，并将算法最多运行 1000 次迭代。

实验设计。 我们的实证研究包含三种实验配置，每种配置均与主论文中提出的理论设定直接对应。对于每种配置，我们针对 20 个不同的样本量（范围从 10^4 到 10^5）分别生成 20 个独立数据集。为确保一致性并便于比较各项实验，我们采用包含两个真实专家（$k^* = 2$）和两个类别（$K = 2$）的模型架构，同时在实验设定中拟合包含三个和四个专家的模型。

初始化。对于每个 $k \in \{k^* + 1, k^* + 2\}$，集合 1，2，..... k 中的索引会被随机分配到 k^* 个 Voronoi 单元中，记为 $\mathcal{A}_1$，....., $\mathcal{A}_{k^*}$，并确保每个单元至少包含一个元素。每次重复实验都会独立进行此分配。随后，从以真实值为中心、方差较小的高斯分布中采样，初始化模型参数，以确保稳定收敛。

5.1 MLMoE 中的 Sigmoid 门控与 Softmax 门控

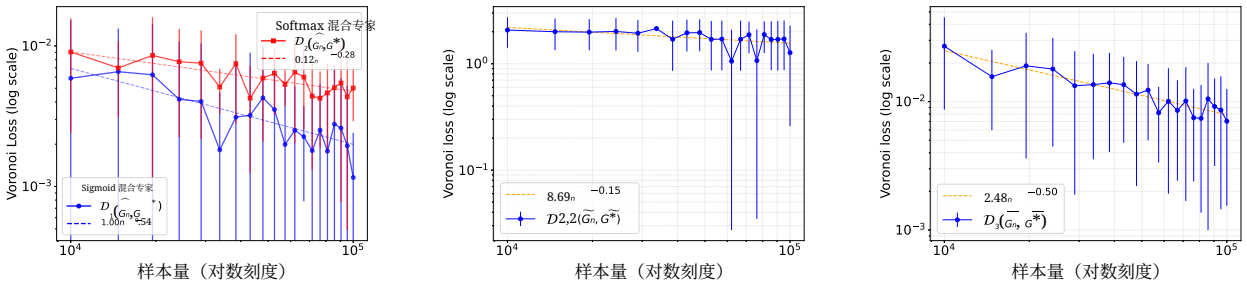
如图（1a）和（1d）所示，在不使用温度参数的情况下采用 sigmoid 门控时，拟合 $\mathcal{O}（n^{-0.49}）$ 个和 $\mathcal{O}（n^{-0.54}）$ 个$k = 3$ 和 $k = 4$ 个专家时，Voronoi 损失分别以 $\mathcal{O}（n^{-0.49}）$ 和 $\mathcal{O}（n^{-0.54}）$ 的阶数收敛至零。这些经验收敛速率与定理1给出的理论保证一致。



(a) 收敛速率 $\mathcal{D}_1(\hat{G}_n, G^*)$ 与 $\mathcal{D}_2(\hat{G}_n, G^*)$ 当 $k = 3$

(b) 收敛速率 $\mathcal{D}_{2,2}(\widetilde{G}_n, \widetilde{G}^*)$ 当 $k = 3$

(c) 收敛速率 $\mathcal{D}_3(\overline{G}_n, \overline{G}^*)$ 当 $k = 3$



(d) 收敛速率 $\mathcal{D}_1(\hat{G}_n, G^*)$ 与 $\mathcal{D}_2(\hat{G}_n, G^*)$ 当 $k = 4$

(e) 的收敛速率 $\mathcal{D}_{2,2}(\widetilde{G}_n, \widetilde{G}^*)$ 当 $k = 4$

(f) 的收敛速率 $\mathcal{D}_3(\overline{G}_n, \overline{G}^*)$ 当 $k = 4$

图1： 经验收敛速率的双对数刻度图。左列：比较不带温度参数的 sigmoid 门控（G_n）与用于 Voronoi 度量的 softmax 门控。中列和右列：~~最~~大似然估计 G_n 的收敛速率，以及内积亲和度分数和欧几里得亲和度分数设定下的 G_n。在这些图中，蓝色曲线表示 sigmoid 门控对应的经验差异，红色曲线则表示 softmax 门控对应的经验差异。橙色点划线表示最小二乘拟合线性回归线。误差棒表示两倍经验标准差。

我们进一步研究了在 [27]中将 sigmoid 门控替换为 softmax 门控配置时的收敛速率。在这两种情况下，sigmoid 门控模型中的参数都比 softmax 门控模型中的参数收敛得更快；对于 $\mathcal{O}(n^{-0.49})$，其收敛速率分别为 $\mathcal{O}(n^{-0.49})$ 和 $\mathcal{O}(n^{-0.30})$；对于 $k = 3$，则分别为 $\mathcal{O}(n^{-0.54})$ 和 $\mathcal{O}(n^{-0.28})$。这些经验结果与第3节中的论断一致，凸显了 sigmoid 门控函数在提高模型样本效率方面的作用。

5.2 带温度参数的 sigmoid门控函数

在标准sigmoid门控MLMoE所使用的内积亲和度分数中引入温度参数后，收敛速率显著变慢。如图（1b）和（1e）所示，Voronoi 损失分别以 $\mathcal{O}(n^{-0.11})$ 和 $\mathcal{O}(n^{-0.15})$ 的速率下降，对应于$k = 3$ 和 $k = 4$，这与定理 2中的理论结果一致。

相比之下，在温度 sigmoid 门控混合专家模型中采用欧几里得亲和度分数时，收敛表现有了显著改善。如图 (1c) 和 (1f) 所示，对于 $\mathcal{O}(n^{-0.52})$，Voronoi 损失以 和 $\mathcal{O}(n^{-0.50})$ 阶的速率收敛，分别对应 $k = 3$ 和 $k = 4$。这些实证结果有力地支持了定理 3中的理论结果。

6 讨论

本文研究了 sigmoid 门控混合专家模型中的参数估计和专家估计的收敛表现。我们首先在标准 sigmoid 门控函数中引入一个正尺度参数，以解决其在混合专家模型过度设定情形中的根本局限：无法确保密度收敛。基于这一改动，我们证明，在 MLMoE 中，就参数和专家估计而言，Sigmoid 门控比 softmax 门控更加样本高效。我们进一步刻画了温度参数在采用内积亲和度分数的 Sigmoid 门控中的作用。我们发现，温度参数与门控参数之间的相互作用会导致专家估计速率缓慢，且不具有多项式阶。通过采用一种新颖的欧几里得亲和度分数作为替代方案，我们证明上述相互作用得以消除，从而恢复专家估计的多项式阶收敛速率。

实际意义。 我们的理论对 MLMoE 模型的设计提出了三项重要的实际意义。

(I.1) 在用于分类的混合专家模型中，有理论保证 *Sigmoid* 门控比 *softmax* 门控具有更高的样本效率。 如表1所示，在过度指定的情形下（即专家数量超过真实数量这一常见的实际场景），与 Softmax 门控函数相比，sigmoid 门控函数能够显著加快参数和专家的估计速率。

(I.2) 在内积亲和度分数中引入温度参数会降低模型的样本效率。 我们的分析表明，将温度参数与传统的内积亲和度分数结合，会与门控参数产生内在的相互作用。这种相互作用会使估计速率慢于任何多项式阶。因此，这意味着若要以给定精度估计模型参数或专家，在引入温度参数的情况下，所需样本数可能呈指数增长，而非像不引入温度参数时那样呈多项式增长。

(I.3) 欧几里得亲和度分数适用于温度控制门控。 用欧几里得亲和度分数替代内积运算，可以消除温度与门控参数之间的不利相互作用，从而恢复多项式收敛速率。这为如何将温度控制纳入 sigmoid 门控函数，同时保持可靠的参数估计，提供了有理论依据的指导。

局限性与未来方向。 就模型表述而言，在式（1）的MLMoE中，专家被设定为简单线性网络的多项逻辑函数。在未来的发展中，我们可以将本文结果推广到用任意神经网络替代线性网络的情形。就问题设定而言，我们的分析是在设定正确情形下进行的，其中数据由Sigmoid 门控 MLMoE 模型生成。在实践中，真实条件分布可能位于该模型类之外，从而导致模型设定错误情形：最大似然估计收敛到在非凸参数空间上最小化Kullback–Leibler散度的伪真混合测度，其唯一性与收敛速率难以刻画。此外，在模型设定错误下，我们证明的几何基础——尤其是用于控制混合测度差异的Voronoi胞元分解——可能会失效，因为伪真参数未必对应于分离良好的专家。因此，为模型设定错误下的Sigmoid门控混合专家模型建立理论仍然是一个开放问题。

《Sigmoid门控函数下多项逻辑专家混合的再思考》补充材料

在本补充材料中，我们给出理论结果的严格证明。Sigmoid 门控的多项逻辑专家混合模型在多种设定下参数估计的收敛速率在附录A中得到严格建立。关于密度估计速率与模型可辨识性的辅助结果的支持性证明见附录B。最后，我们在附录C中详细说明数值研究中使用的实验设定与配置。

A 主要结果的证明

A.1 定理1的证明

为证明定理1中的结论，只需证明

$$\inf_{G\in\mathcal{G}_k(\Theta)} \left(\mathbb{E}_X\left[d_H\big(p_G(\cdot\mid X),\, p_{G^*}(\cdot\mid X)\big)\right] / \mathcal{D}_1(G,G^*)\right) > 0.$$

由于全变差距离上界为海林格距离（$d_H(p,\ q) \gtrsim d_V(p,\ q)$），因此只需证明：

$$\liminf_{G\in\mathcal{G}_k(\Theta)} \left(\mathbb{E}_X\left[d_V\big(p_G(\cdot\mid X),\, p_{G^*}(\cdot\mid X)\big)\right] / \mathcal{D}_1(G,G^*)\right) > 0.$$

局部结构。我们首先通过反证法考察该不等式的局部行为：

$$\liminf_{\substack{\epsilon\to 0\\ G\in\mathcal{G}_k(\Theta)\\ \mathcal{D}_1(G,G^*)\le\epsilon}} \left(\mathbb{E}_X\left[d_V\big(p_G(\cdot\mid X),\, p_{G^*}(\cdot\mid X)\big)\right] / \mathcal{D}_1(G,G^*)\right) > 0.$$

反设上述结论不成立。那么必然存在一列混合测度

$$G_n := \sum_{j=1}^{k_n} \exp(\gamma_j^n)\,\delta_{(\alpha_j^n,\beta_j^n,a_j^n,b_j^n)} \in \mathcal{G}_k(\Theta)$$

使得 $\mathbb{E}_X\left[d_V\big(p_{G_n}(\cdot\mid X)\ p_{G^*}(\cdot\mid X)\big)\right] / \mathcal{D}_1(G_n,\ G^*)$ 和 $\mathcal{D}_1(G_n,\ G^*)$ 均随 $n\to\infty$ 趋于零。接下来，我们将剩余论证分为三个不同的步骤。**步骤 1**。 在这一步中，我们利用泰勒展开来分解差值

$$T_n(s) := \left[\sum_{j=1}^{k_*} \exp(\gamma_j^*)\sigma((\alpha_j^*)^\top X+\beta_j^*)\right]\left[p_{G_n}(Y=s|X) - p_{G^*}(Y=s|X)\right].$$

我们定义

$$u(Y=s|X;\alpha,\beta,a,b) := \sigma(\alpha^\top X+\beta) f(Y=s|X;a,b);$$
$$v(Y=s|X;\alpha,\beta) := \sigma(\alpha^\top X+\beta) p_{G_n}(Y=s|X),$$

对于每个 $s \in [K]$。那么 $T_n(s)$ 可以表示为

$$\begin{aligned}
T_n(s) &= \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp(\gamma_i^n) \left[u(Y = s|X; \alpha_i^n, \beta_i^n, a_i^n, b_i^n) - u(Y = s|X; \alpha_j^*, \beta_j^*, a_j^*, b_j^*) \right] \\
&\quad - \sum_{j=1}^{k_*} \sum_{i \in \mathcal{A}_j} \exp(\gamma_i^n) \left[v(Y = s|X; \alpha_i^n, \beta_i^n) - v(Y = s|X; \alpha_j^*, \beta_j^*) \right] \\
&\quad + \sum_{j=1}^{k_*} \left(\sum_{i \in \mathcal{A}_j} \exp(\gamma_i^n) - \exp(\gamma_j^*) \right) \left[u(Y = s|X; \alpha_j^*, \beta_j^*, a_j^*, b_j^*) - v(Y = s|X; \alpha_j^*, \beta_j^*) \right] \\
&:= A_n - B_n + E_n. \qquad (13)
\end{aligned}$$

基于 A_n 和 B_n 的表达式，我们进一步根据 Voronoi 单元 C_j 的大小对每个表达式进行分解。具体而言，

$$\begin{aligned}
A_n &= \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\gamma_i^n) \left[u(Y = s|X; \alpha_i^n, \beta_i^n, a_i^n, b_i^n) - u(Y = s|X; \alpha_j^*, \beta_j^*, a_j^*, b_j^*) \right] \\
&\quad + \sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} \exp(\gamma_i^n) \left[u(Y = s|X; \alpha_i^n, \beta_i^n, a_i^n, b_i^n) - u(Y = s|X; \alpha_j^*, \beta_j^*, a_j^*, b_j^*) \right] \\
&= A_{n,1} + A_{n,2}.
\end{aligned}$$

接下来，我们定义 $h_\ell(X; a_{j\ell}, b_{j\ell}) := (a_{j\ell})^\top X + b_{j\ell}, \forall\, \ell \in [K]$。于是，

$$f(Y = s|X; a_j, b_j) = \frac{\exp((a_{js})^\top X + b_{js})}{\sum_{\ell=1}^K \exp((a_{j\ell})^\top X + b_{j\ell})} = \frac{\exp(h_s(X; a_{js}, b_{js}))}{\sum_{\ell=1}^K \exp(h_\ell(X; a_{js}, b_{js}))}.$$

利用一阶泰勒展开，$A_{n,1}$ 可写为

$$\begin{aligned}
A_{n,1} &= \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\gamma_i^n) \sum_{|r|=1} \frac{1}{r!} (\Delta\alpha_{ij}^n)^{r_1} (\Delta\beta_{ij}^n)^{r_2} \prod_{\ell=1}^K (\Delta a_{ij\ell}^n)^{r_{3\ell}} (\Delta b_{ij\ell}^n)^{r_{4\ell}} \\
&\quad \times \frac{\partial^{|r|} u}{\partial \alpha_j^{r_1} \partial \beta_j^{r_2} \prod_{\ell=1}^K \partial a_{j\ell}^{r_{3\ell}} \partial b_{j\ell}^{r_{4\ell}}} (Y = s|X; \alpha_j^*, \beta_j^*, a_j^*, b_j^*) + R_1(X, Y),
\end{aligned}$$

其中 $r = (r_1, r_2, r_{31}, \ldots, r_{3K}, r_{41}, \ldots, r_{4K})$，对于所有 $\ell \in [K]$，$r_1, r_{3\ell} \in \mathbb{N}^d$ 和 $r_2, r_{4\ell} \in \mathbb{N}$。此外，$R_1(X, Y)$ 是满足 $R_1(X, Y)/\mathcal{D}_1(G_n, G^*) \to 0$（当 $n \to \infty$ 时）的泰勒余项。令 $g(X; \alpha_\ell, \beta_\ell) := (\alpha_\ell)^\top X + \beta_\ell, \forall \ell \in [k^*]$，我们还可以写成

$$\begin{aligned}
A_{n,1} &= \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\gamma_i^n) \sum_{|r|=1} \frac{1}{\alpha!} (\Delta\alpha_{ij}^n)^{r_1} (\Delta\beta_{ij}^n)^{r_2} \prod_{\ell=1}^K (\Delta a_{ij\ell}^n)^{r_{3\ell}} (\Delta b_{ij\ell}^n)^{r_{4\ell}} \\
&\quad \times \frac{\partial^{|r_1|+r_2} \sigma((\alpha_j^*)^\top X + \beta_j^*)}{\partial \alpha_j^{r_1} \partial \beta_j^{r_2}} \times \frac{\partial^{\sum_{\ell=1}^K |r_{3\ell}|+r_{4\ell}} f}{\prod_{\ell=1}^K \partial a_{j\ell}^{r_{3\ell}} \partial b_{j\ell}^{r_{4\ell}}} (Y = s|X; a_j^*, b_j^*) + R_1(X, Y) \\
&= \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\gamma_i^n) \sum_{|r|=1} \frac{1}{\alpha!} (\Delta\alpha_{ij}^n)^{r_1} (\Delta\beta_{ij}^n)^{r_2} \prod_{\ell=1}^K (\Delta a_{ij\ell}^n)^{r_{3\ell}} (\Delta b_{ij\ell}^n)^{r_{4\ell}}
\end{aligned}$$

$$\times X^{r_1+\sum_{\ell=1}^K r_{3\ell}} \times \frac{\partial^{|r_1|+r_2}\sigma((\alpha_j^*)^\top X+\beta_j^*)}{\partial g^{|r_1|+r_2}} \times \frac{\partial^{\sum_{\ell=1}^K |r_{3\ell}|+r_{4\ell}} f}{\prod_{\ell=1}^K \partial h_\ell^{|r_{3\ell}|+r_{4\ell}}}(Y=s|X;a_j^*,b_j^*)+R_1(X,Y)$$

令 $\mathcal{I}^t := \{(q_1,q_2,q_3):\ q_2+|q_3|=t, |q_1|\le t\}$，且

$$\mathcal{J}_{q_1,q_2,q_3} := \left\{ r=(r_1,r_2,r_{31},\ldots,r_{3K},r_{41},\ldots,r_{4K}) : r_1+\sum_{\ell=1}^K r_{3\ell}=q_1, |r_1|+r_2=q_2, (r_{4\ell}+|r_{3\ell}|)_{\ell=1}^K=q_3 \right\}.$$

$A_{n,1}$ 可以表示为

$$\begin{aligned} A_{n,1} = &\sum_{j:|\mathcal{A}_j|=1}\sum_{t=1}\sum_{(q_1,q_2,q_3)\in\mathcal{I}^t}\sum_{r\in\mathcal{J}_{q_1,q_2,q_3}}\sum_{i\in\mathcal{A}_j}\frac{\exp(\gamma_i^n)}{r!}(\Delta\alpha_{ij}^n)^{r_1}(\Delta\beta_{ij}^n)^{r_2}\prod_{\ell=1}^K(\Delta a_{ij\ell}^n)^{r_{3\ell}}(\Delta b_{ij\ell}^n)^{r_{4\ell}} \\ &\times X^{q_1}\times\frac{\partial^{q_2}\sigma((\alpha_j^*)^\top X+\beta_j^*)}{\partial g^{q_2}}\times\frac{\partial^{|q_3|}f}{\prod_{\ell=1}^K\partial h_\ell^{q_{31}}}(Y=s|X;a_j^*,b_j^*)+R_1(X,Y). \end{aligned}$$

类似地，利用二阶泰勒展开，$A_{n,2}$ 可以写成

$$\begin{aligned} A_{n,2} = &\sum_{j:|\mathcal{A}_j|>1}\sum_{t=1}^2\sum_{(q_1,q_2,q_3)\in\mathcal{I}^t}\sum_{r\in\mathcal{J}_{q_1,q_2,q_3}}\sum_{i\in\mathcal{A}_j}\frac{\exp(\gamma_i^n)}{r!}(\Delta\alpha_{ij}^n)^{r_1}(\Delta\beta_{ij}^n)^{r_2}\prod_{\ell=1}^K(\Delta a_{ij\ell}^n)^{r_{3\ell}}(\Delta b_{ij\ell}^n)^{r_{4\ell}} \\ &\times X^{q_1}\times\frac{\partial^{q_2}\sigma((\alpha_j^*)^\top X+\beta_j^*)}{\partial g^{q_2}}\times\frac{\partial^{|q_3|}f}{\prod_{\ell=1}^K\partial h_\ell^{q_{31}}}(Y=s|X;a_j^*,b_j^*)+R_2(X,Y), \end{aligned}$$

这里 $R_2(X,Y)$表示一个泰勒余项，满足 $R_2(X,Y)/\mathcal{D}_1(G_n,G^*)\to 0$（当 $n\to\infty$时）。对 B_n应用类似的推理进行分解，我们得到

$$\begin{aligned} B_n = &\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\gamma_i^n)\sum_{|r|=1}\frac{1}{r!}(\Delta\alpha_i^n)^{r_1}(\Delta\beta_i^n)^{r_2}\frac{\partial^{|r|}\sigma((\alpha_j^*)^\top X+\beta_j^*)}{\partial\alpha_j^{r_1}\partial\beta_j^{r_2}}g_{G^*}(Y=s|X)+R_3(X,Y) \\ &+\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\gamma_i^n)\sum_{|r|=1}^2\frac{1}{r!}(\Delta\alpha_i^n)^{r_1}(\Delta\beta_i^n)^{r_2}\frac{\partial^{|r|}\sigma((\alpha_j^*)^\top X+\beta_j^*)}{\partial\alpha_j^{r_1}\partial\beta_j^{r_2}}g_{G^*}(Y=s|X)+R_4(X,Y) \\ = &\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\gamma_i^n)\sum_{|r|=1}\frac{1}{r!}\times X^{r_1}\times(\Delta\alpha_i^n)^{r_1}(\Delta\beta_i^n)^{r_2}\frac{\partial^{|r|}\sigma((\alpha_j^*)^\top X+\beta_j^*)}{\partial g^{|r|}}g_{G^*}(Y=s|X)+R_3(X,Y) \\ &+\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\gamma_i^n)\sum_{|r|=1}^2\frac{1}{r!}\times X^{r_1}\times(\Delta\alpha_i^n)^{r_1}(\Delta\beta_i^n)^{r_2}\frac{\partial^{|r|}\sigma((\alpha_j^*)^\top X+\beta_j^*)}{\partial g^{|r|}}g_{G^*}(Y=s|X)+R_4(X,Y) \end{aligned}$$

综合以上所有结果，我们得到

$$\begin{aligned} T_n(s) = &\sum_{j=1}^{k_*}\sum_{t=0}^{1+\mathbf{1}_{\{|\mathcal{A}_j|>1\}}}\sum_{(q_1,q_2,q_3)\in\mathcal{I}^t}U_{q_1,q_2,q_3}^n(j)X^{q_1}\frac{\partial^{q_2}\sigma((\alpha_j^*)^\top X+\beta_j^*)}{\partial g^{q_2}}\times\frac{\partial^{|q_3|}f}{\prod_{\ell=1}^K\partial h_\ell^{q_{31}}}(Y=s|X;a_j^*,b_j^*) \\ &+\sum_{j=1}^{k_*}\sum_{|r|=0}^{1+\mathbf{1}_{\{|\mathcal{A}_j|>1\}}}W_r^n(j)\times X^{r_1}\times\frac{\partial^{|r|}\sigma((\alpha_j^*)^\top X+\beta_j^*)}{\partial g^{|r|}}\times p_{G^*}(Y=s|X)+R(X,Y), \end{aligned}$$

其中 $R(X,Y)$ 是泰勒余项之和，满足 $R(X,Y)/\mathcal{D}_1(X,Y) \to 0$ 当 $n \to \infty$ 以及

$$U^n_{q_1,q_2,q_3}(j) = \begin{cases} \sum_{r\in\mathcal{J}_{q_1,q_2,q_3}} \sum_{i\in\mathcal{A}_j} \frac{\exp(\gamma_i^n)}{r!} (\Delta\alpha_{ij}^n)^{r_1} (\Delta\beta_{ij}^n)^{r_2} \prod_{\ell=1}^K (\Delta a_{ij\ell}^n)^{r_{3\ell}} (\Delta b_{ij\ell}^n)^{r_{4\ell}}, & (q_1,q_2,q_3) \neq (\mathbf{0}_d, \mathbf{0}, \mathbf{0}_K) \\ \sum_{i\in\mathcal{A}_j} \exp(\gamma_i^n) - \exp(\gamma_j^*), & (q_1,q_2,q_3) = (\mathbf{0}_d, \mathbf{0}, \mathbf{0}_K) \end{cases}$$

and

$$W_r^n(j) = \begin{cases} \sum_{i\in\mathcal{A}_j} \frac{\exp(\gamma_i^n)}{r!} (\Delta\alpha_i^n)^{r_1} (\Delta\beta_i^n)^{r_2}, & |r| \neq \mathbf{0}_d, \\ -\sum_{i\in\mathcal{A}_j} \exp(\gamma_i^n) + \exp(\gamma_j^*), & |r| = \mathbf{0}_d, \end{cases}$$

对于任意 $j \in [k_*]$。

步骤 2。现在我们证明，在 $U^n_{q_1、\ q_2、\ q_3}(j)/\mathcal{D}_1(G_n、\ G^*)$ 中，至少有一个在 $n \to \infty$ 时不趋于零。反设它们在 $n \to \infty$ 时全部趋于零。对所有 $j \in \{j\colon\ |\mathcal{A}_j| = 1\}$、$q_2 = 1$、$q_3 = \mathbf{0}_K$ 和 $q_1 \in \{e_1, \ldots, e_d\}$，将 $|U^n_{q_1、\ q_2、\ q_3}(j)|/\mathcal{D}_1(G_n、G^*)$ 求和，其中 $e_m := (0、\ldots\ldots、0、\underbrace{1}_{m\text{-th}}0、\ldots\ldots、0) \in \mathbb{R}^d$，可得

$$\frac{1}{\mathcal{D}_1(G_n, G^*)} \sum_{j:|\mathcal{A}_j|=1} \sum_{i\in\mathcal{A}_j} \exp(\gamma_i^n) \|\Delta\alpha_{ij}^n\| \to 0.$$

同理，对于 $q_1 \in \{e_1, \ldots, e_d\}$、$q_2 = 0$，并对 $q_3 \in \{e'_1, \ldots, e'_K\}$ 迭代，其中 $e'_\ell\colon = (0、\ldots\ldots、0、1、\underbrace{0、\ldots\ldots、0}_{\text{阶}}) \in \mathbb{R}^K$，我们有

$$\frac{1}{\mathcal{D}_1(G_n, G^*)} \sum_{j:|\mathcal{A}_j|=1} \sum_{i\in\mathcal{A}_j} \sum_{\ell=1}^K \exp(\gamma_i^n) \|\Delta a_{ij\ell}^n\| \to 0.$$

另一方面，取 $(q_1, q_2, q_3) = (\mathbf{0}_d, \mathbf{1}, \mathbf{0}_K)$ 并考虑标量分量，可得

$$\frac{1}{\mathcal{D}_1(G_n, G^*)} \sum_{j:|\mathcal{A}_j|=1} \sum_{i\in\mathcal{A}_j} \exp(\gamma_i^n) |\Delta\beta_{ij}^n| \to 0.$$

类似地，对每个 $\ell \in [K]$，取 $(q_1, q_2, q_3) = (\mathbf{0}_d, \mathbf{0}, e'_\ell)$，可得

$$\frac{1}{\mathcal{D}_1(G_n, G^*)} \sum_{j:|\mathcal{A}_j|=1} \sum_{i\in\mathcal{A}_j} \sum_{\ell=1}^K \exp(\gamma_i^n) |\Delta b_{ij\ell}^n| \to 0.$$

结合上述 4 个结果，我们推得

$$\frac{1}{\mathcal{D}_1(G_n, G^*)} \sum_{j:|\mathcal{A}_j|=1} \sum_{i\in\mathcal{A}_j} \exp(\gamma_i^n) \left[\|\Delta\alpha_{ij}^n\| + |\Delta\beta_{ij}^n| + \sum_{\ell=1}^K (\|\Delta a_{ij\ell}^n\| + |\Delta b_{ij\ell}^n|) \right] \to 0. \tag{14}$$

By a对满足 $|\mathcal{A}_j| > 1$ 的索引应用相同的论证，我们还可以推得

$$\frac{1}{\mathcal{D}_1(G_n, G^*)} \sum_{j:|\mathcal{A}_j|>1} \sum_{i\in\mathcal{A}_j} \exp(\gamma_i^n) \left[\|\Delta\alpha_{ij}^n\|^2 + |\Delta\beta_{ij}^n|^2 + \sum_{\ell=1}^K (\|\Delta a_{ij\ell}^n\|^2 + |\Delta b_{ij\ell}^n|^2) \right] \to 0. \tag{15}$$

此外，对于元组 $(q_1, q_2, q_3) = (\mathbf{0}_d, 0, \mathbf{0}_K)$，可得

$$\frac{1}{\mathcal{D}_1(G_n, G^*)} \cdot \left(\sum_{j=1}^{k_*} \left| \sum_{i \in \mathcal{A}_j} \exp(\gamma_i^n) - \exp(\gamma_j^*) \right| \right) \to 0. \tag{16}$$

由方程（14）、（15）、（16）中的极限可知，当 $n \to \infty$ 时，$1 = \mathcal{D}_1(G_n, G^*)/\mathcal{D}_1(G_n, G^*) \to 0$，这就产生了矛盾。因此，当 $n \to \infty$ 时，$U^n_{q_1,\, q_2,\, q_3}(j)/\mathcal{D}_1(G_n, G^*)$ 中至少有一个不趋于零。

步骤 3。 最后，我们应用Fatou引理，得到与步骤 2 中结果相矛盾的结论。令 m_n 表示 $U^n_{q_1,\, q_2,\, q_3}(j)/\mathcal{D}_1(G_n, G^*)$ 和 $W^n_r(j)/\mathcal{D}_1(G_n, G^*)$ 的绝对值中的最大值。那么，根据Fatou引理，

$$0 = \lim_{n\to\infty} \frac{\mathbb{E}_X\left[2 d_V(p_{G_n}(\cdot|X), p_{G^*}(\cdot|X))\right]}{m_n \mathcal{D}_1(G_n, G^*)} \geq \int \sum_{s=1}^{K} \liminf_{n\to\infty} \left| \frac{p_{G_n}(Y=s|X) - p_{G^*}(Y=s|X)}{m_n \mathcal{D}_1(G_n, G^*)} \right| \mathrm{d}X \geq 0.$$

因此，我们得到

$$\frac{|p_{G_n}(Y=s|X) - p_{G^*}(Y=s|X)|}{m_n \mathcal{D}_1(G_n, G^*)} \to 0,$$

这意味着

$$\frac{T_n(s)}{m_n \mathcal{D}_1(G_n, G^*)} \to 0 \quad \text{as } n \to \infty, \quad \forall s \in [K], \text{ a.s. } X.$$

Let

$$\frac{U^n_{q_1,q_2,q_3}(j)}{m_n \mathcal{D}_1(G_n, G^*)} \to \tau_{q_1,q_2,q_3}(j), \quad \frac{W^n_r(j)}{m_n \mathcal{D}_1(G_n, G^*)} \to \eta_r(j) \quad \text{as } n \to \infty.$$

那么，前面的结果表明

$$T(s) = \sum_{j=1}^{k_*} \sum_{t=0}^{1+\mathbf{1}_{\{|\mathcal{A}_j|>1\}}} \sum_{(q_1,q_2,q_3)\in\mathcal{I}^t} \tau_{q_1,q_2,q_3}(j)\, X^{q_1} \frac{\partial^{q_2}\sigma((\alpha_j^*)^\top X + \beta_j^*)}{\partial g^{q_2}} \times \frac{\partial^{|q_3|} f}{\prod_{\ell=1}^K \partial h_\ell^{q_{3\ell}}}(Y=s|X; a_j^*, b_j^*)$$
$$+ \sum_{j=1}^{k_*} \sum_{|r|=0}^{1+\mathbf{1}_{\{|\mathcal{A}_j|>1\}}} \eta_r(j)\, X^{r_1} \frac{\partial^{|r|}\sigma((\alpha_j^*)^\top X + \beta_j^*)}{\partial g^{|r|}}\, p_{G^*}(Y=s|X) = 0,$$

对于任意 $s \in [K]$，几乎必然有 X。然而，这意味着所有系数 $\tau_{q_1,\, q_2,\, q_3}(j)$ 和$\eta_r(j)$ 都必须为零，这就产生了矛盾。

全局结构 **ure:** 由于局部不等式成立，存在一个常数 $\epsilon' > 0$ 使得

$$\inf_{\substack{G\in\mathcal{G}_k(\Theta) \\ \mathcal{D}_1(G,G^*)\leq\epsilon'}} \left(\mathbb{E}_X\left[d_V\big(p_G(\cdot \mid X),\, p_{G^*}(\cdot \mid X)\big)\right] / \mathcal{D}_1(G, G^*)\right) > 0.$$

反设该结论不成立。那么存在一个序列 $G'_n \in \mathcal{G}_k(\Theta)$，使得$\mathcal{D}_1(G'_n, G^*) > \epsilon'$ 且

$$\mathbb{E}_X[d_V(p_{G'_n}(\cdot|X), p_{G^*}(\cdot|X))] \to 0 \quad \text{as } n \to \infty.$$

由于 Θ 是紧的，我们可以抽取一个子序列，使得 $G'_n \to G' \in \mathcal{G}_k(\Theta)$。因此， $\mathcal{D}_1(G', G^*) > \epsilon'$。应用 Fatou引理可得

$$0 = \lim_{n\to\infty} \frac{\mathbb{E}_X[2d_V(p_{G'_n}(\cdot|X), p_{G^*}(\cdot|X))]}{\mathcal{D}_1(G'_n, G^*)} \geq \int \sum_{s=1}^{K} \liminf_{n\to\infty} |p_{G'_n}(Y=s|X) - p_{G^*}(Y=s|X)| \, \mathrm{d}X.$$

因此,

$$\int \sum_{s=1}^{K} |p_{G'}(Y=s|X) - p_{G^*}(Y=s|X)| \, \mathrm{d}X = 0,$$

这意味着对于 X 中几乎所有的 $s \in [K]$， $p_{G'}(Y=s|X) = p_{G^*}(Y=s|X)$。根据可识别性，我们得出 $G' = G^*$，这与 $\mathcal{D}_1(G', G^*) > \epsilon'$ 相矛盾。因此，全局不等式成立，证明完毕。

A.2 定理2的证明

Fir首先，我们给出一个将在该证明中使用的有用引理如下 s:

引理1. 对于任意 $r \geq 1$，如果以下条件成立:

$$\lim_{\varepsilon\to 0} \inf_{G\in\mathcal{G}_k(\Upsilon):\mathcal{D}_{2,r}(G,\widetilde{G}^*)\leq\varepsilon} \frac{\mathbb{E}_X[d_V(\widetilde{p}_G(\cdot|X), \widetilde{p}_{\widetilde{G}^*}(\cdot|X))]}{\mathcal{D}_{2,r}(G, \widetilde{G}^*)} = 0,$$

则我们得到

$$\inf_{G_n\in\mathcal{G}_k(\Upsilon)} \sup_{G\in\mathcal{G}_k(\Upsilon)\setminus\mathcal{G}_{k^*-1}(\Upsilon)} \mathbb{E}_{\widetilde{p}_G}[\mathcal{D}_{2,r}(G_n, G)] \gtrsim n^{-1/2}.$$

引理的证明 1 推迟至附录 A.2.2。现在我们准备给出定理 2的主要证明。

A.2.1 主要证明

基于引理1的结果，只需构造一列混合测度 G_n 使得 $\widetilde{\mathcal{D}}_{2,\,r}(G_n,\ G^*) \to 0$ 且

$$\frac{\mathbb{E}_X[d_V(\widetilde{p}_{G_n}(\cdot|X), \widetilde{p}_{\widetilde{G}^*}(\cdot|X))]}{\mathcal{D}_{2,r}(G, \widetilde{G}^*)} \to 0, \tag{17}$$

当 $n \to \infty$。为此，我们选取 $G_n = \sum_{i=1}^{k^*+1} \exp(\gamma_i^n)\delta_{(\alpha_i^n, \beta_i^n, \tau^n, a_i^n, b_i^n)}$ 使得

- $a_i^n = a_j^*, \ \ b_i^n = b_j^*$对任意 $i \in \mathcal{A}_j$;
- $\tau^n = \tau^* + t_n$;
- $\beta_i^n = \beta_j^* + s_{n,\ j}$ 对于任意 $i \in \mathcal{A}_j$;
- $a_i^n = a_j^* + v_{n,\ j}$ 对于任意 $i \in \mathcal{A}_j$;
- $\exp(\gamma_i^n) = \frac{1}{|\mathcal{A}_j|}\exp(\gamma_j^*)$ 对任意 $i \in \mathcal{A}_j$,

其中 $v_{n,j} := (v_n^{(1)},_j, \ldots, v_n^{(d)},_j) \in \mathbb{R}^d$ 与 $t_n, s_{n,j} \in \mathbb{R}$ 将在稍后选取，使得 $v_n^{(u)},_j, s_{n,j}, t_n \to 0$ 当 $n \to \infty$ 对任意 $u \in [d]$ 与 $j \in [k^*]$。然后，损失函数 $\mathcal{D}_{2,r}(G_n, G^*)$ 变为

$$\mathcal{D}_{2,r}(G_n, \widetilde{G}^*) = \sum_{j=1}^{k^*} |\mathcal{A}_j| \cdot (\|v_{n,j}\|^r + |t_n|^r + |s_{n,j}|^r)$$

显然，当 $n \to \infty$ 时，$\mathcal{D}_{2,r}(G_n, G^*) \to 0$。

现在，我们将证明，当 $n \to \infty$ 时，$\mathbb{E}_X[d_H(\widetilde{p}_{G_n}(\cdot|X), \widetilde{p}_{\widetilde{G}^*}(\cdot|X))]/\mathcal{D}_{2,r}(G_n, G^*)$ 趋于零。根据上述对 G_n 的设定，考虑如下量：$T_n(s) = \Big[\sum_{i=1}^{k^*} \exp(\gamma_i^*)\sigma\Big(\frac{(\alpha_i^*)^\top X + \beta_i^*}{\tau^*}\Big)\Big] [\widetilde{p}_{G_n}(Y = s|X) - \widetilde{p}_{\widetilde{G}^*}(Y = s|X)$

$$\begin{aligned}
T_n(s) :=& \sum_{j=1}^{k^*} \sum_{i \in \mathcal{A}_j} \exp(\gamma_i^n)[\widetilde{u}(Y = s|X; \alpha_i^n, \beta_i^n, \tau^n, a_i^n, b_i^n) - \widetilde{u}(Y = s|X; \alpha_j^*, \beta_j^*, \tau^*, a_j^*, b_j^*)] \\
&- \sum_{j=1}^{k^*} \sum_{i \in \mathcal{A}_j} \exp(\gamma_i^n)[\widetilde{v}(Y = s|X; \alpha_i^n, \beta_i^n, \tau^n) - \widetilde{v}(Y = s|X; \alpha_j^*, \beta_j^*, \tau^*)] \\
&+ \sum_{j=1}^{k^*} \Big[\sum_{i \in \mathcal{A}_j} \exp(\gamma_i^n) - \exp(\gamma_j^*) \Big] \Big[\widetilde{u}(Y = s|X; \alpha_j^*, \beta_j^*, \tau^*, a_j^*, b_j^*) - \widetilde{u}(Y = s|X; \alpha_j^*, \beta_j^*, \tau^*)\Big] \\
&:= A_n - B_n + E_n,
\end{aligned}$$

其中，我们记 $f(Y = s|X;\ a, b) := \exp(a_s^\top X + b_s)/\sum_{\ell=1}^K \exp(a_\ell^\top X + b_\ell)$，$\widetilde{u}(Y = s|X;\ \alpha, \beta, \tau, a, b) := \sigma\Big(\frac{\alpha^\top X + \beta}{\tau}\Big) f(Y = s|X;\ a, b)$，以及 $\widetilde{v}(Y = s|X;\ \alpha, \beta, \tau) := \sigma\Big(\frac{\alpha^\top X + \beta}{\tau}\Big) p_{G_n}(Y = s|X)$，对每个 $s \in [K]$ 均成立。根据上述对 G_n 的表述，我们可以将项 A_n 重写为

$$A_n = \sum_{j=1}^{k^*} \sum_{i \in \mathcal{A}_j} \exp(\gamma_i^n) \left[\sigma\left(\frac{(\alpha_i^n)^\top X + \beta_i^n}{\tau^n}\right) - \sigma\left(\frac{(\alpha_j^*)^\top X + \beta_j^*}{\tau^*}\right)\right] f(Y = s|X; a^*, b^*)$$

通过一阶泰勒展开，我们可以将 A_n 改写为

$$\begin{aligned}
A_n = \sum_{j=1}^{k^*} \sum_{i \in \mathcal{A}_j} \exp(\gamma_i^n) \Bigg\{ \sum_{u=1}^{d} \Big[\frac{v_{n,j}^{(u)}}{\tau^*} - \frac{t_n(\alpha_j^*)^{(u)})}{(\tau^*)^2}\Big] \cdot X^{(u)} + \Big[\frac{s_{n,j}}{\tau^*} - \frac{t_n\beta_j^*}{(\tau^*)^2}\Big] \Bigg\} \sigma\left(\frac{(\alpha_j^*)^\top X + \beta_j^*}{\tau^*}\right) \\
\times \left[1 - \sigma\left(\frac{(\alpha_j^*)^\top X + \beta_j^*}{\tau^*}\right)\right] f(Y = s|X; a^*, b^*) + R_1(X, Y),
\end{aligned}$$

其中，$R_1(X, Y)$ 是泰勒余项，满足 $R_1(X, Y)/\mathcal{D}_{2,r}(G_n, G^*) \to 0$ 当 $n \to \infty$ 时。因此，通过选择

$$t_n = \frac{1}{n}; \qquad v_{n,j} = \frac{t_n\alpha_j^*}{\tau^*} = \frac{\alpha_j^*}{n\tau^*}; \qquad s_{n,j} = \frac{t_n\beta_j^*}{\tau^*} = \frac{\beta_j^*}{n\tau^*},$$

按照相同的推理，我们得到 $A_n/\mathcal{D}_{2,r}(G_n, G^*) \to 0$ 当 $n \to \infty$ 时。

~按照相同的推理，我们还得到 $B_n/\mathcal{D}_{2,r}(G_n, G^*) \to 0$ 当 $n \to \infty$ 时。~结合 $E_n = 0$，可知 $T_n(s)/\mathcal{D}_{2,r}(G_n, G^*) \to 0$ 当 $n \to \infty$ 时。由于项 $\sum_{i=1}^{k^*} \exp(\gamma_i^*)\sigma\left((\alpha_i^*)^\top X + \beta_i^*\right)$ 有界，我们推得 $|\widetilde{p}_{G_n}(Y|X) - \widetilde{p}_{\widetilde{G}^*}(Y|X)|/\mathcal{D}_{2,r}(G_n, \widetilde{G}^*) \to 0$，当 $n \to \infty$ 时，(X, Y) 几乎必然成立。因此，我们满足方程 (17) 中的条件。证明完毕。

A.2.2 引理 A.1 的证明。

对于固定常数 $M_1 > 0$ 以及稍后将选取的充分小的 $\varepsilon > 0$，由~假设可知，我们可以找到一个混合测度 $G_1^* \in \mathcal{G}_k(\Upsilon)$，使其满足 $\mathcal{D}_{2,r}(G_1^*, G^*) = 2\varepsilon$ 且 $\mathbb{E}_X[d_V(\widetilde{p}_{G_1^*}(\cdot|X), \widetilde{p}_{\widetilde{G}^*}(\cdot|X))] \leq M_1\varepsilon$。此外，对于任意序列 $G_n' \in \mathcal{G}_k(\Upsilon)$，我们有

$$2 \max_{G \in \{G_1^*, \widetilde{G}^*\}} \mathbb{E}_{\widetilde{p}_G}[\mathcal{D}_{2,r}(G_n', G)] \geq \mathbb{E}_{\widetilde{p}_{\widetilde{G}^*}}[\mathcal{D}_{2,r}(G_n', \widetilde{G}^*)] + \mathbb{E}_{\widetilde{p}_{G_1^*}}[\mathcal{D}_{2,r}(G_n', G_1^*)],$$

其中，$\mathbb{E}_{\widetilde{p}_G}$ 表示关于密度为 $\widetilde{p}_G$ 的乘积测度所取的期望。由于损失函数 $\mathcal{D}_{2,r}$ 满足弱三角不等式，因此我们可以找到一个常数 $M_2 > 0$，使得

$$\mathcal{D}_{2,r}(G_n', \widetilde{G}^*) + \mathcal{D}_{2,r}(G_n', G_1^*) \geq M_2\mathcal{D}_{2,r}(G_1^*, \widetilde{G}^*) = 2M_2\varepsilon.$$

因此可得

$$\begin{aligned}\max_{G \in \{G_1^*, \widetilde{G}^*\}} \mathbb{E}_{\widetilde{p}_G}[\mathcal{D}_{2,r}(G_n', G)] &\geq \frac{1}{2}\left(\mathbb{E}_{\widetilde{p}_{\widetilde{G}^*}}[\mathcal{D}_{2,r}(G_n', \widetilde{G}^*)] + \mathbb{E}_{\widetilde{p}_{G_1^*}}[\mathcal{D}_{2,r}(G_n', G_1^*)]\right) \\ &\geq M_2\varepsilon \cdot \inf_{f_1, f_2}\left(\mathbb{E}_{\widetilde{p}_{\widetilde{G}^*}}[f_1] + \mathbb{E}_{\widetilde{p}_{G_1^*}}[f_2]\right),\end{aligned}$$

其中，上述下确界中的 f_1 和 f_2 是关于 $X_1, X_2, \ldots, X_n$ 的可测函数，且满足 $f_1 + f_2 = 1$。根据全变差距离的定义，下确界值 $\inf_{f_1, f_2}\left(\mathbb{E}_{\widetilde{p}_{\widetilde{G}^*}}[f_1] + \mathbb{E}_{\widetilde{p}_{G_1^*}}[f_2]\right)$ 等于 $1 - d_V(\widetilde{p}_{\widetilde{G}^*}(\cdot|X), \widetilde{p}_{G_1^*}(\cdot|X))$。因此，我们得到

$$\begin{aligned}\max_{G \in \{G_*', \widetilde{G}^*\}} \mathbb{E}_{p_G}[\mathcal{D}_{2,r}(G_n', G)] &\geq M_2\varepsilon\Big(1 - d_V(p_{\widetilde{G}^*}(\cdot|X), p_{G_1^*}(\cdot|X))\Big) \\ &\geq M_2\varepsilon\left[1 - \sqrt{1 - (1 - M_1^2\varepsilon^2)^n}\right].\end{aligned}$$

通过选择 $\varepsilon = n^{-1/2}/M_1$，我们有 $M_1^2\varepsilon^2 = \frac{1}{n}$ 和 $\sqrt{1 - (1 - M_1^2\varepsilon^2)^n} = \sqrt{1 - (1 - 1/n)^n} \to \sqrt{1 - e^{-1}} < 1$，这意味着

$$\inf_{G_n' \in \mathcal{G}_k(\Upsilon)} \sup_{G \in \mathcal{G}_k(\Upsilon) \setminus \mathcal{G}_{k^*-1}(\Upsilon)} \mathbb{E}_{\widetilde{p}_G}[\mathcal{D}_{2,r}(G_n', G)] \geq \max_{G \in \{G_1^*, \widetilde{G}^*\}} \mathbb{E}_{\widetilde{p}_G}[\mathcal{D}_{2,r}(G_n', G)] \gtrsim n^{-1/2}.$$

因此，我们得出定理2的结论，即

$$\inf_{G_n \in \mathcal{G}_k(\Upsilon)} \sup_{G \in \mathcal{G}_k(\Upsilon) \setminus \mathcal{G}_{k^*-1}(\Upsilon)} \mathbb{E}_{\widetilde{p}_G}[\mathcal{D}_{2,r}(G_n, G)] \gtrsim n^{-1/2},$$

对于任意 $r \geq 1$。

A.3 定理证明3

为达到所需的　定理中的结论A.3，我们需要证明以下关键不等式：

$$\inf_{G\in\mathcal{G}_k(\Upsilon)} \mathbb{E}_X[d_V(\overline{p}_G(\cdot|X), \overline{p}_{\overline{G}^*}(\cdot|X))]/\mathcal{D}_3(G, \overline{G}^*) > 0, \tag{18}$$

然后将其分为两部分，分别称为局部结构和全局结构。由于全局结构可以类比附录A.1中的情形处理，为简洁起见，此处省略其证明。

局部结构。在这一部分，我们旨在通过反证法建立以下不等式：

$$\lim_{\varepsilon\to 0} \inf_{G\in\mathcal{G}_k(\Upsilon), \mathcal{D}_3(G,\overline{G}^*)\leq\varepsilon} \mathbb{E}_X[d_V(\overline{p}_G(\cdot|X), \overline{p}_{\overline{G}^*}(\cdot|X))]/\mathcal{D}_3(G, \overline{G}^*) > 0.$$

假设这个局部不等式不成立。利用附录A.1中的一些推导，我们按以下三步框架进行：

步骤 1。首先，我们定义

$$\begin{aligned}
\overline{u}(Y=s|X;\alpha,\beta,\tau,a,b) &:= \sigma\left(\frac{\|\alpha-X\|+\beta}{\tau}\right) f(Y=s|X;a,b);\\
\overline{v}(Y=s|X;a,b,\tau) &:= \sigma\left(\frac{\|\alpha-X\|+\beta}{\tau}\right) \overline{p}_{\overline{G}_n}(Y=s|X),
\end{aligned}$$

对于任意 $s\in[K]$。接下来，将以下差值分解如下

$$\begin{aligned}
T_n(s) &:= \Big[\sum_{j=1}^{k^*}\exp(\gamma_j^*)\sigma\left(\frac{\|\alpha_j^*-X\|+\beta_j^*}{\tau^*}\right)\Big](\overline{p}_{\overline{G}_n}(Y=s|X)-\overline{p}_{\overline{G}^*}(Y=s|X))\\
&= \sum_{j=1}^{k^*}\sum_{i\in\mathcal{A}_j}\exp(\gamma_i^n)[\overline{u}(Y=s|X;\alpha_j^n,\beta_j^n,\tau^n,a^n,b^n)-\overline{u}(Y=s|X;\alpha_j^*,\beta_j^*,\tau^*,a^*,b^*)]\\
&\quad-\sum_{j=1}^{k^*}\sum_{i\in\mathcal{A}_j}\exp(\gamma_i^n)[\overline{v}(Y=s|X;\alpha_j^n,\beta_j^n,\tau^n)-\overline{v}(Y=s|X;\alpha_j^*,\beta_j^*,\tau^*)]\\
&\quad+\sum_{j=1}^{k^*}\left(\sum_{i\in\mathcal{A}_j}\exp(\gamma_i^n)-\exp(\gamma_j^*)\right)\overline{u}(Y=s|X;\alpha_j^*,\beta_j^*,\tau^*,a^*,b^*)\\
&\quad-\sum_{j=1}^{k^*}\left(\sum_{i\in\mathcal{A}_j}\exp(\gamma_i^n)-\exp(\gamma_j^*)\right)\overline{v}(Y=s|X;\alpha_j^*,\beta_j^*,\tau^*)\\
&:= A_n-B_n+E_{n,1}-E_{n,2}.
\end{aligned}$$

给定上述 A_n 与 B_n 的表述，我们进一步根据 Voronoi 单元 $\mathcal{A}_j$ 的基数，将每一项分解为两个更小的分量。具体而言，

$$A_n = \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\gamma_i^n)[\overline{u}(Y=s|X;\alpha_j^n,\beta_j^n,\tau^n,a^n,b^n)-\overline{u}(Y=s|X;\alpha_j^*,\beta_j^*,\tau^*,a^*,b^*)]$$

$$+\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\gamma_i^n)[\overline{u}(Y=s|X;\alpha_j^n,\beta_j^n,\tau^n,a^n,b^n)-\overline{u}(Y=s|X;\alpha_j^*,\beta_j^*,\tau^*,a^*,b^*)]$$
$$:=A_{n,1}+A_{n,2}.$$

对任意 $\ell\in[K]$ 和 $i\in[k]$，我们记 $h_\ell(X;\,a_{i\ell},b_{i\ell}):=(a_{i\ell})^\top X+b_{i\ell}$，$g_i(X;\,\alpha_i,\beta_i,\tau):=(\|\alpha_i-X\|+\beta_i)/\tau$。借助一阶泰勒展开，$A_{n,1}$ 可以表示为

$$A_{n,1}=\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\gamma_i^n)\sum_{|t|=1}\frac{1}{t!}(\Delta\alpha_{ij})^{t_1}(\Delta\beta_{ij})^{t_2}(\Delta\tau_{ij})^{t_3}\prod_{\ell=1}^{K}(\Delta a_{ij\ell}^n)^{t_{4\ell}}(\Delta b_{ij\ell}^n)^{t_{5\ell}}$$
$$\times\frac{\partial^{|t_1|+t_2+t_3}\sigma\circ g_j}{\partial\alpha_j^{t_1}\partial\beta_j^{t_2}\partial\tau^{t_3}}(X;\alpha_j^*,\beta_j^*,\tau^*)\times\frac{\partial^{\sum_{\ell=1}^K(|t_{4\ell}|+t_{5\ell})}f}{\prod_{\ell=1}^K\partial a_{j\ell}^{|t_{4\ell}|}\partial b_{j\ell}^{t_{5\ell}}}(Y=s|X;a_j^*,b_j^*)+R_1(X,Y),$$

其中 $t=(t_1,t_2,t_3,t_{41},\ldots,t_{4K},t_{51},\ldots,t_{5K})$，其中 $t_1,t_{4\ell}\in\mathbb{N}^d$，$t_2,t_3,t_{5\ell}\in\mathbb{N}$对所有 $\ell\in[K]$ 均成立。此外，$R_1(X,Y)$ 是一个泰勒余项，使得当 $n\to\infty$ 时，$R_1(X,Y)/D_3(G_n,G^*)\to\overline{0}$。根据函数 h 的表达式，我们有

$$A_{n,1}=\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\gamma_i^n)\sum_{|t|=1}\frac{1}{t!}(\Delta\alpha_{ij})^{t_1}(\Delta\beta_{ij})^{t_2}(\Delta\tau_{ij})^{t_3}\prod_{\ell=1}^{K}(\Delta a_{ij\ell}^n)^{t_{4\ell}}(\Delta b_{ij\ell}^n)^{t_{5\ell}}\frac{\partial^{|t_1|+t_2+t_3}g_j}{\partial\alpha_j^{t_1}\beta_j^{t_2}\tau^{t_3}}(X;\alpha_j^*,\beta_j^*,\tau_j^*)$$
$$\times X^{\sum_{\ell=1}^K t_{4\ell}}\times\frac{\partial^{|t_1|+t_2+t_3}\sigma\circ g_j}{\partial g_j^{|t_1|+t_2+t_3}}(X;\alpha_j^*,\beta_j^*,\tau_j^*)\times\frac{\partial^{\sum_{\ell=1}^K(|t_{4\ell}|+t_{5\ell})}f}{\prod_{\ell=1}^K\partial h_\ell^{|t_{4\ell}|+t_{5\ell}}}(Y=s|X;a_j^*,b_j^*)+R_1(X,Y).$$

我们定义 $q_1=\sum_{\ell=1}^K t_{4\ell}\in\mathbb{N}^d$，$q_2=|t_1|+t_2+t_3\in\mathbb{N}$，$q_3=(q_{3\ell})_{\ell=1}^K:=(|t_{4\ell}|+t_{5\ell})_{\ell=1}^K\in\mathbb{N}^K$。此外，我们记 $\mathcal{I}^z:=\{(q_1,q_2,q_3)\in\mathbb{N}^d\times\mathbb{N}\times\mathbb{N}^K:\,q_2+|q_3|=z,|q_1|\le|q_3|\}$ 并且

$$\mathcal{J}_{q_1,q_2,q_3}:=\left\{t=(t_1,t_2,t_3,t_{41},\ldots,t_{4K},t_{51},\ldots,t_{5K}):\sum_{\ell=1}^K t_{4\ell}=q_1,|t_1|+t_2+t_3=q_2,(|t_{4\ell}|+t_{5\ell})_{\ell=1}^K=q_3\right\}.$$

项 $A_{n,\ 1}$ 可写为:

$$A_{n,1}=\sum_{j:|\mathcal{A}_j|=1}\sum_{t=1}\sum_{(q_1,q_2,q_3)\in\mathcal{I}^t}\sum_{t\in\mathcal{J}_{q_1,q_2,q_3}}\sum_{i\in\mathcal{A}_j}\frac{\exp(\gamma_i^n)}{t!}(\Delta\alpha_{ij})^{t_1}(\Delta\beta_{ij})^{t_2}(\Delta\tau_{ij})^{t_3}\prod_{\ell=1}^{K}(\Delta a_{ij\ell}^n)^{t_{4\ell}}(\Delta b_{ij\ell}^n)^{t_{5\ell}}$$
$$\times\frac{\partial^{|t_1|+t_2+t_3}g_j}{\partial\alpha_j^{t_1}\partial\beta_j^{t_2}\partial\tau^{t_3}}(X;\alpha_j^*,\beta_j^*,\tau_j^*)\times X^{q_1}\times\sigma^{(q_2)}(g_j(X;\alpha_j^*,\beta_j^*,\tau_j^*))\times\frac{\partial^{|q_3|}f}{\partial h^{q_3}}(Y=s|X;a_j^*,b_j^*)$$
$$+R_1(X,Y).$$

同样，通过二阶泰勒展开，项$A_{n,\ 2}$ 可以表示如下

$$A_{n,2}=\sum_{j:|\mathcal{A}_j|>1}\sum_{t=1}^{2}\sum_{(q_1,q_2,q_3)\in\mathcal{I}^t}\sum_{t\in\mathcal{J}_{q_1,q_2,q_3}}\sum_{i\in\mathcal{A}_j}\frac{\exp(\gamma_i^n)}{t!}(\Delta\alpha_{ij})^{t_1}(\Delta\beta_{ij})^{t_2}(\Delta\tau_{ij})^{t_3}\prod_{\ell=1}^{K}(\Delta a_{ij\ell}^n)^{t_{4\ell}}(\Delta b_{ij\ell}^n)^{t_{5\ell}}$$
$$\times\frac{\partial^{|t_1|+t_2+t_3}g_j}{\partial\alpha_j^{t_1}\partial\beta_j^{t_2}\partial\tau^{t_3}}(X;\alpha_j^*,\beta_j^*,\tau_j^*)\times X^{q_1}\times\sigma^{(q_2)}(g_j(X;\alpha_j^*,\beta_j^*,\tau_j^*))\times\frac{\partial^{|q_3|}f}{\partial h^{q_3}}(Y=s|X;a_j^*,b_j^*)$$

$+ R_2(X,\ Y)$,

其中 $R_2(X,Y)$ 是泰勒余项，使得 $R_2(X,Y)/\mathcal{D}_3(G_n,G^*) \to 0$ 当 $\overline{n} \to \overline{\infty}$。类似地，通过对 A_n 的分解应用相同的论证，我们可以将 $B_n = B_{n,1} + B_{n,2}$表示如下，其中

$$
\begin{aligned}
B_{n,1} = & \sum_{j:|\mathcal{A}_j|=1} \sum_{i\in\mathcal{A}_j} \exp(\gamma_j^n) \sum_{|t|=1} \frac{1}{t!} (\Delta\alpha_{ij}^n)^{t_1} (\Delta\beta_{ij}^n)^{t_2} (\Delta\tau_{ij})^{t_3} \sigma^{(|t|)}(g_j(X;\alpha_j^*,\beta_j^*,\tau^*)) \\
& \times \frac{\partial^{|t|} g_j}{\partial\alpha_j^{t_1}\partial\beta_j^{t_2}\partial\tau^{t_3}}(X;\alpha_j^*,\beta_j^*,\tau^*) \overline{p}_{\overline{G}_n}(Y=s|X) + R_3(X,Y) \\
B_{n,2} = & \sum_{j:|\mathcal{A}_j|>1} \sum_{i\in\mathcal{A}_j} \exp(\gamma_j^n) \sum_{|t|=1}^{2} \frac{1}{t!} (\Delta\alpha_{ij}^n)^{t_1} (\Delta\beta_{ij}^n)^{t_2} (\Delta\tau_{ij})^{t_3} \sigma^{(|t|)}(g_j(X;\alpha_j^*,\beta_j^*,\tau^*)) \\
& \times \frac{\partial^{|t_1|+t_2+t_3} g_j}{\partial\alpha_j^{t_1}\partial\beta_j^{t_2}\partial\tau^{t_3}}(X;\alpha_j^*,\beta_j^*,\tau^*) \overline{p}_{\overline{G}_n}(Y=s|X) + R_4(X,Y),
\end{aligned}
$$

其中 $R_3(X,Y)$ 和 $R_4(X,Y)$ 是泰勒余项，使得
$R_3(X,\ Y)/\mathcal{D}_3(\overline{G}_n,\ \overline{G}^*) \to 0, R_4(X,Y)/\mathcal{D}_3(\overline{G}_n,\overline{G}^*) \to 0$ 当 $n \to \infty$。我们可以验证

$$
\begin{aligned}
A_n + E_{n,1} = & \sum_{j=1}^{k^*} \sum_{t=0}^{1+\mathbf{1}_{\{|\mathcal{A}_j|>1\}}} \sum_{(q_1,q_2,q_3)\in\mathcal{I}^t} U_{q_1,q_2,q_3}^n(j) \times \frac{\partial^{|t_1|+t_2+t_3} g_j}{\partial\alpha_j^{t_1}\partial\beta_j^{t_2}\partial\tau^{t_3}}(X;\alpha_j^*,\beta_j^*,\tau_j^*) \\
& \times X^{q_1} \times \sigma^{(q_2)}(g_j(X,\alpha_j^*,\beta_j^*,\tau_j^*)) \times \frac{\partial^{|q_3|} f}{\partial h^{q_3}}(Y=s|X;a_j^*,b_j^*) + R_1(X,Y) + R_2(X,Y);
\end{aligned} \tag{19}
$$

$$
\begin{aligned}
B_n + E_{n,2} = & \sum_{j=1}^{k^*} \sum_{t=0}^{1+\mathbf{1}_{|\mathcal{A}_j|>1}} W_t^n(j) \frac{\partial^{|t_1|+t_2+t_3} g_j}{\partial\alpha_j^{t_1}\partial\beta_j^{t_2}\partial\tau^{t_3}}(X;\alpha_j^*,\beta_j^*,\tau^*) \\
& \times \sigma^{(|t|)}(g_j(X;\alpha_j^*,\beta_j^*,\tau^*)) \overline{p}_{\overline{G}_n}(Y=s|X) + R_3(X,Y) + R_4(X,Y)
\end{aligned} \tag{20}
$$

其中我们记 $U_{q_1,q_2,q_3}^n(j)$ 等于

$$
\begin{cases}
\sum_{i\in\mathcal{A}_j} \sum_{t\in\mathcal{J}_{q_1,q_2,q_3}} \frac{\exp(\gamma_i^n)}{t!} (\Delta\alpha_{ij})^{t_1} (\Delta\beta_{ij})^{t_2} (\Delta\tau_{ij})^{t_3} \prod_{\ell=1}^{K} (\Delta a_{ij\ell}^n)^{t_{4\ell}} (\Delta b_{ij\ell}^n)^{t_{5\ell}}, & (q_1,q_2,q_3) \neq (\mathbf{0}_d,0,\mathbf{0}_K) \\
\sum_{i\in\mathcal{A}_j} \exp(\gamma_i^n) - \exp(\gamma_j^*), & (q_1,q_2,q_3) = (\mathbf{0}_d,0,\mathbf{0}_K)
\end{cases}
$$

and

$$
W_t^n(j) = \begin{cases}
\sum_{i\in\mathcal{A}_j} \frac{\exp(\gamma_j^n)}{t!} (\Delta\alpha_{ij}^n)^{t_1} (\Delta\beta_{ij}^n)^{t_2} (\Delta\tau_{ij})^{t_3}, & (t_1,t_2,t_3) \neq (\mathbf{0}_d,0,0) \\
\sum_{i\in\mathcal{A}_j} \exp(\gamma_i^n) - \exp(\gamma_j^*) & (t_1,t_2,t_3) = (\mathbf{0}_d,0,0).
\end{cases}
$$

步骤 2. 在这一步中，我们将证明，在 $n \to \infty$ 时，$U_{q_1,q_2,q_3}^n(j)/\mathcal{D}_3(G_n,G^*)$ 中至少有一个不趋于零。反设它们在 $n \to \infty$ 时均趋于零。对以下各项求和

$|U^n_{q_1,q_2,q_3}(j)|/\mathcal{D}_3(G_n, G^*)$，对所有 $j \in \{[k^*]\colon |\mathcal{A}_j| = 1\}$, $q_2 = 0, q_3 = \mathbf{0}_K$ 和 $q_1 \in \{e_1, e_2, \ldots, e_d\}$，其中 $e_i := (\underbrace{0, \ldots, 0}_{i\text{-th}}, 1, 0, \ldots, 0) \in \mathbb{R}^d$，我们得到

$$\frac{1}{\mathcal{D}_3(\overline{G}_n, \overline{G}^*)} \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\gamma_i) \|\Delta a^n_{ij}\| \to 0.$$

对于 $q_1 = \mathbf{0}_d$， $q_2 = 0$ 和 $q_3 \in \{e'_1, \ldots, e'_K\}$ 的相似度，其中 $e'_\ell := (0, \ldots, 0, \underbrace{1, 0}, \ldots, 0) \in \mathbb{R}^K$，我们得到$\ell$ 阶
have

$$\frac{1}{\mathcal{D}_3(\overline{G}_n, \overline{G}^*)} \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\gamma_i) |\Delta b^n_{ij}| \to 0.$$

另一方面，取 $(q_1, q_2, q_3) = (\mathbf{0}_d, \mathbf{1}, \mathbf{0}_K)$，并分别考虑三种情况 $|t_1| = 1$， $t_2 = 1$ 和 $t_3 = 1$ ，可得

$$\frac{1}{\mathcal{D}_3(\overline{G}_n, \overline{G}^*)} \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\gamma_i) \|\Delta \alpha^n_{ij}\| \to 0.$$

$$\frac{1}{\mathcal{D}_3(\overline{G}_n, \overline{G}^*)} \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\gamma_i) |\Delta \beta^n_{ij}| \to 0.$$

$$\frac{1}{\mathcal{D}_3(\overline{G}_n, \overline{G}^*)} \sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} \exp(\gamma_i) |\Delta \tau| \to 0.$$

结合上述 5 个结果，我们推得

$$\sum_{j:|\mathcal{A}_j|=1} \sum_{i \in \mathcal{A}_j} K_{ij}(1,1,1,1,1)/\mathcal{D}_3(\overline{G}_n, \overline{G}^*) \to 0. \tag{21}$$

运用相同的论证，我们还可推得

$$\sum_{j:|\mathcal{A}_j|>1} \sum_{i \in \mathcal{A}_j} K_{ij}(2,2,2,2,2)/\mathcal{D}_3(G_n, \overline{G}^*) \to 0 \tag{22}$$

此外，对于元组 $(q_1, q_2, q_3) = (\mathbf{0}_d, 0, \mathbf{0}_K)$，可得

$$\frac{1}{\mathcal{D}_3(\overline{G}_n, \overline{G}^*)} \cdot \left(\sum_{j=1}^{k^*} \left| \sum_{i \in \mathcal{A}_j} \exp(\gamma_i) - \exp(\gamma^*_i) \right| \right) \to 0 \tag{23}$$

由方程（21）、（22）、（23）中的极限可知，当 $n \to \infty$ 时， $1 = \mathcal{D}_3(\overline{G}_n, \overline{G}^*)/\mathcal{D}_3(\overline{G}_n, \overline{G}^*) \to 0$, 这与结论矛盾。因此，当 $n \to \infty$ 时， $U^n_{q_1,q_2,q_3}(j)/\mathcal{D}_3(G_n, G^*)$ 中至少有一项不会趋于零。

步骤 3. 现在，我们将 m_n 记为$U^n_{q_1,\ q_2,\ q_3}(j)/\mathcal{D}_3(G_n,\ G^*)$与 $W^n_t(j)/\mathcal{D}_3(G_n,\ G^*)$的绝对值的最大值。由于当$n \to \infty$时，$U^n_{q_1,\ q_2,\ q_3}(j)/\mathcal{D}_3(G_n,\ G^*)$中至少有一个不趋于零，我们推得 $m_n \not\to 0$，且等价地有 $1/m_n \not\to \infty$。然后我们将

$$U^n_{q_1,q_2,q_3}(j)/\mathcal{D}_3(\overline{G}_n, \overline{G}^*) \to \zeta_{q_1,q_2,q_3}(j)$$

$$W_t^n(j)/\mathcal{D}_3(\overline{G}_n,\overline{G}^*)\to\eta_t(j),$$

为 $n\to\infty$。利用Fatou引理，我们有

$$0=\lim_{n\to\infty}\frac{\mathbb{E}_X[2d_V(\overline{p}_{\overline{G}_n}(\cdot|X),\overline{p}_{\overline{G}^*}(\cdot|X))]}{\overline{m}_n\mathcal{D}_3(\overline{G}_n,\overline{G}^*)}\geq\int\sum_{s=1}^{K}\liminf_{n\to\infty}\left|\frac{\overline{p}_{\overline{G}_n}(Y=s|X)-\overline{p}_{\overline{G}^*}(Y=s|X)}{m_n\mathcal{D}_3(\overline{G}_n,\overline{G}^*)}\right|\mathrm{d}X\geq 0.$$

因此，这表明对任意 $s\in[K]$ 且几乎必然地 X，有 $|p_{G_n}(Y=s|X)-p_{G^*}(Y=s|X)|/[m_n\mathcal{D}_3(G_n,G^*)]\to 0$。该结果等价于

$$T_n(s)/[m_n\mathcal{D}_3(\overline{G}_n,\overline{G}^*)]\to 0, \tag{24}$$

当 $n\to\infty$ 时，对任意 $s\in[K]$。综合方程（19）、（20）和（24）中的结果，我们有

$$\sum_{j=1}^{k^*}\sum_{t=0}^{1+\mathbf{1}_{\{|\mathcal{A}_j|>1\}}}\left[\sum_{(q_1,q_2,q_3)\in\mathcal{I}^t}\zeta_{q_1,q_2,q_3}(j)X^{q_1}\sigma^{(q_2)}(g_j(X,\alpha_j^*,\beta_j^*,\tau_j^*))\frac{\partial^{|q_3|}f}{\partial h^{q_3}}(Y=s|X;a_j^*,b_j^*)\right.$$
$$\left.+\eta_t(j)\sigma^{(|t|)}(g_j(X;\alpha_j^*,\beta_j^*,\tau^*))\overline{p}_{\overline{G}^*}(Y=s|X)\right]\times\frac{\partial^{|t_1|+t_2+t_3}g_j}{\partial\alpha^{t_1}\partial\beta^{t_2}\partial\tau^{t_3}}(X;\alpha_j^*,\beta_j^*,\tau^*)=0,$$

对几乎必然的 (X,Y) 成立。值得注意的是，参数 α_j^*,β_j^*两两互异，则集合

$$\left\{X^{q_1}\sigma^{(q_2)}(g_j(X,\alpha_j^*,\beta_j^*,\tau_j^*))\frac{\partial^{|q_3|}f}{\partial h^{q_3}}(Y=s|X;a_j^*,b_j^*)\frac{\partial^{|t_1|+t_2+t_3}g_j}{\partial\alpha^{t_1}\partial\beta^{t_2}\partial\tau^{t_3}}(X;\alpha_j^*,\beta_j^*,\tau^*);\right.$$
$$\left.\sigma^{(|t|)}(g_j(X;\alpha_j^*,\beta_j^*,\tau^*))\overline{p}_{\overline{G}^*}(Y=s|X)\frac{\partial^{|t_1|+t_2+t_3}g_j}{\partial\alpha^{t_1}\partial\beta^{t_2}\partial\tau^{t_3}}(X;\alpha_j^*,\beta_j^*,\tau^*)\right\}$$

对任意 $j\in[k^*]$、$0\leq q_2+|q_3|\leq 2, 0\leq|t|\leq 2$ 和 $j\in[k^*]$ 线性无关。因此，我们得到所有系数 $\zeta_{q_1,q_2,q_3}=\eta_t=0$ 均为零，这与 ζ_{q_1,q_2,q_3} 中至少有一个不为零的事实相矛盾。

B 辅助结果的证明

B.1 命题的证明 2

B.1.1 关键结果

首先，我们引入一些将在证明中使用的记号。令 $P_k(\Theta):=\{p_G(Y\mid X):G\in\mathcal{G}_k(\Theta)\}$ 表示 $\mathcal{G}_k(\Theta)$ 中所有混合测度对应的条件密度函数集合。此外，我们将 $N(\varepsilon,P_k(\Theta)\ \|\cdot\|_\infty)$ 定义为度量空间 $(P_k(\Theta)\ \|\cdot\|_\infty)$ 的覆盖数 [36]，而 $H_B(\varepsilon,P_k(\Theta)\ h)$ 则是 $P_k(\Theta)$ 关于海林格距离的括号熵 [36]。接下来，以下结果给出了这些量的上界：

引理 2. 对于任意有界集 Θ 和 $\varepsilon\in(0,1/2)$，我们有

(i) $\log N(\varepsilon,P_k(\Theta),\|\cdot\|_\infty)\lesssim Kdk\log(1/\varepsilon)$;

(*ii*) $H_B(\varepsilon,\ P_k(\Theta)\ d_H) \lesssim Kdk\log(1/\varepsilon)$。

引理的证明 2。**第（i）部分** 首先，我们引入参数集

$$\Omega := \{(a,b) \in \mathbb{R}^{d\times K} \times \mathbb{R}^K : (\gamma,\alpha,\beta,a,b) \in \Theta\}.$$

$$\Delta := \{(\gamma,\alpha,\beta) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R} : (\gamma,\alpha,\beta,a,b) \in \Theta\},$$

因为 Θ 是紧集，所以Ω与 Θ 均为紧集。因此，它们各自存在一个 ε-覆盖；分别记为Ω_ε 与 Δ_ε 。可以验证

$$|\Omega_\varepsilon| \lesssim \mathcal{O}(\varepsilon^{-K(d+1)k}), \qquad |\Delta_\varepsilon| \lesssim \mathcal{O}(\varepsilon^{-(d+2)k}).$$

接下来，考虑一个混合测度 $G = \sum_{i=1}^{k'} \exp(\gamma_i)\delta_{(\alpha_i,\beta_i,a_i,b_i)} \in \mathcal{G}_k(\Theta)\ \forall\, k' \in [k]$. 利用上面引入的 ε-覆盖，我们为 G 关联两个近似测度。

首先，对每个 $i \in [k']$，令$(\bar{a}_i, b_i) \in \Omega_\varepsilon$ 为覆盖中最接近(a_i, b_i)的点，并定义

$$\widetilde{G} = \sum_{i=1}^{k'} \exp(\gamma_i)\delta_{(\alpha_i,\beta_i,\bar{a}_i,\bar{b}_i)}.$$

接下来，取$(\bar{\gamma}_i, \bar{\alpha}_i, \beta_i) \in \Delta_\varepsilon$ 作为覆盖中最接近$(\gamma_i, \alpha_i, \beta_i)$的点，并令

$$\bar{G} = \sum_{i=1}^{k'} \exp(\bar{\gamma}_i)\delta_{(\bar{\alpha}_i,\bar{\beta}_i,\bar{a}_i,\bar{b}_i)}.$$

根据构造可直接得出，$p_{\bar{G}}$ 属于该族

$$\mathcal{H} := \left\{p_G \in P_k(\Theta) : (\bar{\gamma}_i, \bar{\alpha}_i, \bar{\beta}_i) \in \Delta_\varepsilon,\ (\bar{a}_i, \bar{b}_i) \in \Omega_\varepsilon,\ \forall i \in [k']\right\}.$$

现在验证 $\mathcal{H}$ 是度量空间（$P_k(\Theta)\ \|\cdot\|_\infty$）的一个 ε-覆盖，但不一定是最小覆盖。为此，我们对一致距离 $\|p_G - p_{\bar{G}}\|_\infty$进行界定。应用三角不等式可得

$$\|p_G - p_{\overline{G}}\|_\infty \le \|p_G - p_{\widetilde{G}}\|_\infty + \|p_{\widetilde{G}} - p_{\overline{G}}\|_\infty.$$

对于任意混合测度 $G = \sum_{i=1}^{k'} \exp(\gamma_i)\delta_{(\alpha_i,\beta_i,a_i,b_i)}$，回顾其对应的条件密度为

$$p_G(Y = s \mid X) = \sum_{i=1}^{k'} \pi(Y \mid X; \gamma_i, \alpha_i, \beta_i)\, f(Y = s \mid X; a_i, b_i),$$

其中

$$\pi(Y \mid X; \gamma_i, \alpha_i, \beta_i) = \frac{\exp(\gamma_i)\sigma((\alpha_i)^\top X + \beta_i)}{\sum_{j=1}^{k'} \exp(\gamma_j)\sigma((\alpha_j)^\top X + \beta_j)}, \text{ and } f(Y = s \mid X; a_i, b_i) = \frac{\exp((a_{is})^\top X + b_{is})}{\sum_{\ell=1}^{K} \exp((a_{i\ell})^\top X + b_{i\ell})}$$

转而考察右侧的第一项,

$$\begin{aligned}
&\|p_{\widetilde{G}} - p_G\|_\infty \\
&= \sum_{s=1}^{K} \sup_X \left|p_G(Y = s \mid X) - p_{\widetilde{G}}(Y = s \mid X)\right|.
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{s=1}^{K} \sup_{X} \left| \sum_{i=1}^{k'} \pi(Y \mid X; \gamma_i, \alpha_i, \beta_i)\ f(Y = s \mid X; a_i, b_i) - \sum_{i=1}^{k'} \pi(Y \mid X; \gamma_i, \alpha_i, \beta_i)\ f(Y = s \mid X; \bar{a}_i, \bar{b}_i) \right| \\
&\leq \sum_{s=1}^{K} \sup_{X} \sum_{i=1}^{k'} |\pi(Y \mid X; \gamma_i, \alpha_i, \beta_i)| \cdot \left| f(Y = s \mid X; a_i, b_i) - f(Y = s \mid X; \bar{a}_i, \bar{b}_i) \right| \\
&\leq \sum_{s=1}^{K} \sup_{X} \sum_{i=1}^{k'} \left| f(Y = s \mid X; a_i, b_i) - f(Y = s \mid X; \bar{a}_i, \bar{b}_i) \right|.
\end{aligned}
$$

f 的可微性意味着它是利普希茨连续的，且具有某个常数 $L > 0$。此外，由于 $\mathcal{X}$ 是一个有界集，我们可以假设对于某个常数 $B > 0$，有 $\|X\| \leq B$。由此可得

$$
\begin{aligned}
\|p_{\widetilde{G}} - p_G\|_\infty &\leq \sum_{s=1}^{K} \sum_{i=1}^{k'} L \cdot \left(\|a_{is} - \bar{a}_{is}\| \cdot \|X\| + \|b_{is} - \bar{b}_{is}\| \right) \\
&\leq K k' L (B\varepsilon + \varepsilon) \lesssim \varepsilon.
\end{aligned}
$$

类似地，我们对第二项进行如下界定：

$$
\|p_{\overline{G}} - p_{\widetilde{G}}\|_\infty = \sum_{s=1}^{K} \sup_{X} \sum_{i=1}^{k'} \left| \pi(Y \mid X; \overline{\gamma}_i, \overline{\alpha}_i, \overline{\beta}_i) - \pi(Y \mid X; \gamma_i, \alpha_i, \beta_i) \right| f(Y = s | X; \overline{a}_i; \overline{b}_i).
$$

$$
\begin{aligned}
&\leq \sum_{s=1}^{K} \sum_{i=1}^{k'} L \cdot \left(\|\alpha_i - \bar{\alpha_i}\| \cdot \|X\| + |\beta_i - \bar{\beta}_i| + |\gamma_i - \bar{\gamma}_i| \right) \\
&\leq K k' L (B\varepsilon + \varepsilon + \varepsilon) \lesssim \varepsilon.
\end{aligned}
$$

将这些界合并可得 $\|p_G - p_G\|_\infty \lesssim \varepsilon$. 因此， $\mathcal{H}$ 是度量空间 $(P_k(\Theta)\ \|\cdot\|_\infty)$ 的一个 ε–覆盖。根据覆盖数的定义，我们因此得到

$$
\begin{aligned}
N(\varepsilon, P_k(\Theta), \|\cdot\|_\infty) \leq |\mathcal{H}| = |\Omega_\varepsilon| \times |\Delta_\varepsilon| &= \mathcal{O}(\varepsilon^{-K(d+1)k}) \times \mathcal{O}(\varepsilon^{-(d+2)k}) \\
&= \mathcal{O}(\varepsilon^{-(Kd+K+d+2)k}),
\end{aligned}
$$

这等价于

$$
\log N(\varepsilon, P_k(\Theta), \|\cdot\|_\infty) \lesssim Kdk \log(1/\varepsilon) \lesssim Kdk \log(1/\varepsilon).
$$

第（ii）部分 固定 $\varepsilon > 0$，并选择 $\eta \leq \varepsilon$（稍后指定）。令 $p_1, p_2, \ldots, p_N$ 为在范数 $|\cdot|_\infty$ 下对 $P_k(\Theta)$ 的一个 η-覆盖，其中 $N := N(\eta, P_k(\Theta)\ |\cdot|_\infty)$。现在我们按如下方式构造对所有 $i \in [N]$ 均成立的括号：$[L_i(Y \mid X)\ U_i(Y \mid X)]$

$$
L_i(Y \mid X) := \max\{p_i(Y \mid X) - \eta,\, 0\}, \qquad U_i(Y \mid X) := \min\{p_i(Y \mid X) + \eta,\, 1\}.
$$

由此构造可知， $P_k(\Theta) \subset \bigcup_{i=1}^{N} [L_i(Y \mid X)\ U_i(Y \mid X)]$ 和 $U_i(Y \mid X) - L_i(Y \mid X) \leq \min\{2\eta, 1\}$。此外

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

回顾一下， $H_B(2K\eta,\ P_k(\Theta)\ |\cdot|_1)$ 定义为覆盖 $P_k(\Theta)$ 所需的括号的最小数量的对数，其中括号的 L_1-宽度不大于 $2K\eta$ 。对于任意 $p \in P_k(\Theta)$，在 η-覆盖中都存在某个 p_i，使得 $|p-p_i|\infty \le \eta$。相应的括号 $[L_i, U_i]$ 满足

$$p \in [L_i, U_i] \qquad \text{and} \qquad \|U_i - L_i\|_1 \le 2K\eta.$$

因此， η-覆盖中的每个 p_i 都会产生一个有效的 $2K\eta$-括号。因此，所需括号的数量不超过 $N(\eta,\ P_k(\Theta)\ |\cdot|\infty)$，由此得到

$$H_B(2K\eta, P_k(\Theta), \|\cdot\|_1) \ \le\ \log N(\eta, P_k(\Theta), \|\cdot\|_\infty) \ \lesssim\ Kdk\log(1/\eta),$$

其中，最后一个不等式使用了第（i）部分中的界。因此，取 $\eta = \varepsilon/(2K)$可得$H_B(\varepsilon,\ P_k(\Theta)\ |\cdot|_1)$ $\lesssim Kdk\log(1/\varepsilon)$。由于海林格距离满足 $h \le |\cdot|_1$，我们得到

$$H_B(\varepsilon, P_k(\Theta), d_H) \ \lesssim Kdk\ \log(1/\varepsilon).$$

□

给定一个混合集 $\mathcal{G}_k(\Theta)$，我们引入与之关联的条件密度族

$$\overline{P}_k(\Theta) \ :=\ \{\, p_{(G+G^*)/2}(\cdot \mid X) :\ G \in \mathcal{G}_k(\Theta)\,\}, \qquad \overline{P}_k^{1/2}(\Theta) \ :=\ \{\, p^{1/2}_{(G+G^*)/2}(\cdot \mid X) :\ G \in \mathcal{G}_k(\Theta)\,\}.$$

对于任意 $\xi > 0$，我们定义以真实条件密度 $p_{G^*}(\cdot \mid X)$ 为中心、并与集合 $P_k^{1/2}(\Theta)$ 相交的海林格球，如下所示

$$\overline{P}_k^{1/2}(\Theta, \xi) := \Big\{ p^{1/2} \in \overline{P}_k^{1/2}(\Theta) :\ \mathbb{E}_X\big[d_H\big(p(\cdot \mid X),\, p_{G^*}(\cdot \mid X)\big)\big] \le \xi \Big\}.$$

与前文相同， $H_B(\varepsilon,\ \mathcal{F},\ |\cdot|)$ 表示函数类 $\mathcal{F}$关于范数 $|\cdot|$的括号熵。参照 [36], 我们定义局部括号积分，以刻画 $\xi > 0$ 海林格球的大小：

$$J_B\big(\xi, \overline{P}_k^{1/2}(\Theta, \xi)\big) := \int_{\xi^2/2^{13}}^{\xi} \sqrt{H_B\big(t, \overline{P}_k^{1/2}(\Theta, \xi), \|\cdot\|\big)}\ dt \vee \xi.$$

其中 $t \vee \xi := \max\{t, \xi\}$。现在我们陈述 [36] 中关于密度估计速率的一个关键结果，并将其改写为适合我们记号的形式。

引理3. *(定理 7.4, [36])*
设 $\Psi(\xi)$ 满足 $\Psi(\xi) \ge J_B(\xi, P_k^{1/2}(\Theta, \xi))$，且 $\Psi(\xi)/\xi^2$ 关于 ξ 单调不增。那么，对于一个普适常数 c 以及满足 $\sqrt{n}\xi_n^2 \ge c\Psi(\xi_n)$ 的任意序列 (ξ_n)，我们有

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

对于任意 $\xi \ge \xi_n$。

该引理的证明见 [36]。

B.1.2 主证明

注意，$H_B\Big(t,\, P_k^{1/2}(\Theta,\ t)\ \|\cdot\|\Big)\ \le\ H_B(t,\ \ P_k(\Theta)\ \ d_H)\ \ \forall\, t>0$。因此，

$$
\begin{aligned}
J_B\Big(\xi,\, \overline{P}_k^{1/2}(\Theta,\xi)\Big) &= \int_{\xi^2/2^{13}}^{\xi} H_B^{1/2}\Big(t,\, \overline{P}_k^{1/2}(\Theta,t),\, \|\cdot\|\Big)\, dt \vee \xi \\
&\le \int_{\xi^2/2^{13}}^{\xi} H_B^{1/2}(t,\, P_k(\Theta),\, d_H)\, dt \vee \xi \qquad \text{(monotonicity)} \\
&\lesssim \int_{\xi^2/2^{13}}^{\xi} \sqrt{Kdk}\sqrt{\log(1/t)}\, dt \vee \xi \qquad \text{(using Lemma 2)} \\
&\lesssim \xi\, \sqrt{Kdk\log(1/\xi)}
\end{aligned}
$$

定义

$$\Psi(\xi)\ =\ \sqrt{Kdk}\cdot\xi\,\sqrt{\log(1/\xi)}\quad\text{and}\quad \xi_n=\sqrt{Kdk\log(n)/n}.$$

由于 $\Psi(\xi)\ge J_B(\xi, P_k^{1/2}(\Theta,\xi))$ 且 $\Psi(\xi)/\xi^2$ 是非递增的，引理 3 给出

$$\mathbb{P}\Big(\mathbb{E}_X\Big[d_H\Big(p_{\widehat{G}_n}(\cdot\mid X),\, p_{G^*}(\cdot\mid X)\Big)\Big] > C\sqrt{Kdk\log(n)/n}\Big)\ \le\ \exp(-c\log(n))\,,$$

其中 C 和 c 是仅依赖于 Θ的普适正常数。证明完毕。

B.2 命题1的证明

在该假设下，对于每个 $s\in[K]$，且 μ-几乎处处成立的$\mathbf{X}\ \in\mathcal{X}$，我们有

$$
\begin{aligned}
&\sum_{i=1}^{k}\frac{\exp(\gamma_i)\sigma((\alpha_i)^\top X+\beta_i)}{\sum_{j=1}^{k}\exp(\gamma_j)\sigma((\alpha_j)^\top X+\beta_j)}\cdot\frac{\exp((a_{is})^\top X+b_{is})}{\sum_{\ell=1}^{K}\exp((a_{i\ell})^\top X+b_{i\ell})} \\
&= \sum_{i=1}^{k'}\frac{\exp(\gamma_i')\sigma((\alpha_i')^\top X+\beta_i')}{\sum_{j=1}^{k'}\exp(\gamma_j')\sigma((\alpha_j')^\top X+\beta_j')}\cdot\frac{\exp((a_{is}')^\top X+b_{is}')}{\sum_{\ell=1}^{K'}\exp((a_{i\ell}')^\top X+b_{i\ell}')}
\end{aligned}
\tag{25}
$$

[17] 证明多项逻辑斯蒂混合模型是可辨识的，这意味着任意两个混合测度G 和 G' 必须具有相同数量的专家，且门控集合相同，即 $k=k'$ 和

$$\left\{\frac{\exp(\gamma_i)\sigma((\alpha_i)^\top X+\beta_i)}{\sum_{j=1}^{k}\exp(\gamma_j)\sigma((\alpha_j)^\top X+\beta_j)}: i\in[k]\right\}=\left\{\frac{\exp(\gamma_i')\sigma((\alpha_i')^\top X+\beta_i')}{\sum_{j=1}^{k'}\exp(\gamma_j')\sigma((\alpha_j')^\top X+\beta_j')}: i\in[k']\right\},$$

对几乎必然成立的$\mathbf{X}\ \in\mathcal{X}$。不失一般性，我们假设

$$\frac{\exp(\gamma_i)\sigma((\alpha_i)^\top X+\beta_i)}{\sum_{j=1}^{k}\exp(\gamma_j)\sigma((\alpha_j)^\top X+\beta_j)}=\frac{\exp(\gamma_i')\sigma((\alpha_i')^\top X+\beta_i')}{\sum_{j=1}^{k'}\exp(\gamma_j')\sigma((\alpha_j')^\top X+\beta_j')}\tag{26}$$

对于任意 $i \in [k]$。上述结果得出 $\gamma_i = \gamma_i'$， $\alpha_i = \alpha_i'$ 并且 $\beta_i = \beta_i'$ 对于所有 $i \in [k]$。因此，方程 26变为

$$\sum_{i=1}^{k} \exp(\gamma_i) u(Y = s|X; \alpha_i, \beta_i, a_i, b_i) = \sum_{i=1}^{k} \exp(\gamma_i) u(Y = s|X; \alpha_i', \beta_i', a_i', b_i'), \tag{27}$$

对于任意 $s \in [K]$，几乎必然有 $\mathbf{X} \in \mathcal{X}$，其中 $u(Y = s|X; \alpha, \beta, a, b) := \sigma(\alpha^\top X + \beta) f(Y = s|X;\ a, b)$ 接下来，我们将考虑集合 $[k]$ 的 r 子集，记为 $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_r$，它们满足以下性质： $\gamma_i = \gamma_{i'}$ 对于任意 i， $i' \in \mathcal{S}_t$ 对于 $t \in [r]$。因此，方程 27 可以改写为

$$\sum_{t=1}^{r} \sum_{i \in \mathcal{S}_t} \exp(\gamma_i) u(Y = s|X; \alpha_i, \beta_i, a_i, b_i) = \sum_{t=1}^{r} \sum_{i \in \mathcal{S}_t} \exp(\gamma_i) u(Y = s|X; \alpha_i', \beta_i', a_i', b_i').$$

对于任意 $s \in [K]$ ，几乎必然有 $\mathbf{X} \in \mathcal{X}$。
由上式可知，对于每个 $t \in [r]$，有 $\{(a_{il})^\top X + b_{il})_{l=1}^K : i \in \mathcal{S}_t\} = \{(a_{il}')^\top X + b_{il}')_{l=1}^K : i \in \mathcal{S}_t\}$ 几乎必然有 $\mathbf{X} \in \mathcal{X}$。因此

$$\{(a_{i1}, \ldots, a_{iK}, b_{i1}, \ldots, b_{iK}) : i \in \mathcal{S}_t\} = \{(a_{i1}', \ldots, a_{iK}', b_{i1}', \ldots, b_{iK}') : i \in \mathcal{S}_t\}.$$

同样，我们可以假设对于任意 $i \in \mathcal{S}_t$， $(a_{i1}, \ldots, a_{iK}, b_{i1}, \ldots, b_{iK}) = (a_{i1}', \ldots, a_{iK}', b_{i1}', \ldots, b_{iK}')$。因此，我们得到 $G \equiv G'$，从而完成证明。

C 实验细节

在本附录中，我们通过数值实验对标准 sigmoid 门控多项逻辑混合专家模型（不含温度参数）的最大似然估计收敛速率这一理论结果进行实证验证，相关实验见附录C.1；并将其与附录 C.2中的 softmax 门控多项逻辑混合专家模型进行比较。同时，在附录C.3中，我们针对温度增强型门控变体，通过实证展示：在 sigmoid 门控多项逻辑混合专家模型的参数估计中，使用欧几里得亲和度分数优于使用标准内积亲和度分数。

合成数据。 对于每个样本量 n，我们生成独立同分布样本 $(X_i, Y_i)_{i=1}^n$，具体做法是先从 $[0, 1]$ 上的均匀分布中抽取 X_i，再从与各定理配置中所指定混合专家模型相对应的真实条件密度 $p_{G^*}(Y = s|X)$、 $\widetilde{p}_{\widetilde{G}^*}(Y = s|X)$ 和 $p_{G^*}(Y = s|X)$ 中抽取 Y_i。

最大似然估计。 一种广泛用于计算最大似然 $\widehat{\ }$ $\widetilde{\ }$ 估计量 G （或 G， G）的方法是，对每组样本使用期望最大化（EM）算法[12]。由于在最大化步骤中无法得到门控参数的闭式更新，我们转而采用先前在 [4]中使用的基于 EM 的数值过程。我们将收敛容差设为 $\varepsilon = 10^{-6}$ ，并将算法最多运行 1000 次迭代。

实验设计。 我们的实证研究考察三种实验配置，分别与主论文中提出的理论设定直接对应。对于每种配置，我们针对 20 种不同的样本量（范围从 10^4 到 10^5）生成 20 个独立数据集。为确保实验间的一致性并便于比较，我们采用包含两名真实专家（$k^* = 2$）和两个类别（$K = 2$）的模型架构，同时在实验设定中拟合包含三名和四名专家的模型。

初始化。对于每个 $k \in \{k^* + 1, k^* + 2\}$，将集合 1、2、……、$k$ 中的索引随机分配到 k^* 个 Voronoi 单元中，记为 $\mathcal{A}_1$、……、$\mathcal{A}_{k^*}$，并确保每个单元至少包含一个元素。每次重复实验都会独立进行这一分配。随后，从以真实值为中心的小方差高斯分布中抽样，对模型参数进行初始化，以确保稳定收敛。

C.1 无温度的 Sigmoid 门控

表 2： sigmoid 门控函数与 softmax 门控函数比较分析所用的真实参数。

	门控参数	专家参数	
		类别 1 (a^*, b^*)	类别 2 (a^*, b^*)
Sigmoid 函数门控 (p_{G^*})			
专家 1	$(\gamma_0^*, \alpha_0^*, \beta_0^*) = (0。2, -1.0, -0.5)$	$(0.0, -1.0)$	$(-1.0, 1.0)$
专家 2	$(\gamma_0^*, \alpha_0^*, \beta_0^*) = (-0.3, 1.0, 0.5)$	$(0.0, 0.0)$	$(0.0, 2.0)$
softmax门控 (g_{G^*})			
专家 1	$(\beta_{01}^*\beta_{11}^*) = (0.2, -1.0)$	$(-1.0, 0。0)$	$(1。0, -1.0)$
专家 2	$(\beta_{02}^*\beta_{12}^*) = (-0.3, 1。0)$	$(0.0, 0.0)$	$(2.0, 0.0)$

问题设定与定理（1）中的定义相同。真实参数$\hat{}$对应的取值详见表2。如图（1）所示，最大似然估计量MLE G_n在Voronoi度量 $\mathcal{D}_1$ 下经验性地收敛至真实混合测度 G^*，收敛阶分别为 $\mathcal{O}$（$n^{-0.49}$）和 $\mathcal{O}$（$n^{-0.54}$），分别对应 $k = 3$ 和 $k = 4$。这一结果有力地支持了定理1中的理论结果。

C.2 CoSigmoid门控与Softmax门控之间的收敛性分析 g

模型定义。接下来，我们通过实证比较定理（1中的标准sigmoid门控MLMoE模型与softmax门控MLMoE模型的收敛行为，其条件密度 g_{G^*}（$Y = s|X$）定义见 [27]中的公式（1）。我们设置模型的真实参数，使其满足 [27], 中的情形2条件：至少有一个专家对协变量 X （$b_{i\ell}^* = 0$ 对所有 ℓ）的系数趋于零。已有证明表明，在标准softmax门控下，这种情形会导致参数估计的收敛速率较慢。

真实参数配置。两个模型的真实参数值详见表2。

实证收敛结果。图1展示了在将 $k = 3$ 和 $k = 4$ 个专家拟合到真实的 $k^* = 2$专家模型时，两种门控函数下最大似然估计量的实证收敛情况。sigmoid 门控模型（在 $\mathcal{D}_1$下测量）表现出更快的收敛速率： $k = 3$ 为 $\mathcal{O}(n^{-0.49})$， $k = 4$为 $\mathcal{O}(n^{-0.54})$；而 softmax 门控模型（根据 [27]中的 $\mathcal{D}_2$ 测量）收敛较慢，速率分别为 $\mathcal{O}(n^{-0.30})$ $(k = 3)$ 和 $\mathcal{O}(n^{-0.28})$ $(k = 4)$。

C.3 带温度的 Sigmoid 门控

采用内积亲和度分数的 Sigmoid 门控。

表3： 真　　温度和内积亲和度评分下 MLMoE 模型的参数　　re

	门控参数	专家参数	
	$(\gamma^*, \alpha^*, \beta^*, \tau^*)$	类别 1 $(a^*, \ b^*)$	类别 2 $(a^*, \ b^*)$
专家 1	(0，1。0，0，0。1)	(0，0)	(1.86, 0.963)
专家 2	(0，1。001，0，0。1)	(0，0)	(1.87, 0.964)

我们考虑与定理（2）中相同的实验设置。模型参数的真实值汇总于表 3 中。两个子图（1b）、（1e）展示了最大 ~似然估计量 G_n 在 Voronoi度量 $\mathcal{D}_2$ 下的估计误差，2。我们观察到，经验结果以 G^* 阶的速率 $\mathcal{O}(n^{-0.11})$ 和 $\mathcal{O}(n^{-0.15})$ 分别向真实混合测度 $k = 3$ 和 $k = 4$ 收敛。这些观察结果与定理 2 中的理论保证一致。**Sigmoid 函数门控与欧几里得亲和度分数。**

表 4： 真实　温度和欧几里得亲和度分数下MLMoE模型的参数

	门控参数	专家参数	
		类别 1	类别 2
专家 1	$(\gamma_0^*, \alpha_0^*, \beta_0^*, \tau^*) = (1, -5, -0.5, 2.0)$	$(a_{11}^* \, b_{11}^*) = (-1, \ 2)$	$(a_{12}^* \, b_{12}^*) = (0, 0)$
专家 2	$(\gamma_0^*, \alpha_0^*, \beta_0^*, \tau^*) = (-1, 5, 0.5, 2.0)$	$(a_{21}^* b_{21}^*) = (1, \ -1)$	$(a_{22}^* \, b_{22}^*) = (0, \ 0)$

实验配置遵循定理（3）。表 4 报告了该设置中使用的相应真实参数值。如两个子图（1b）、（1e）所示，估计量 G_n 在经验上收敛于 G^*，其度量为 Voronoi度量 $\mathcal{D}_3$。观测到的收敛速率分别为 $\mathcal{O}(n^{-0.52})$ 和 $\mathcal{O}(n^{-0.50})$，对应于$k = 3$ 和 $k = 4$，这表明收敛行为稳定。该实验证据与定理 3中的理论预测高度一致。

参考文献

[1] M. Abdin 等。Phi-3 技术报告：一款可在手机本地运行的高性能语言模型。*arXiv* 预印本 *arXiv:2404.14219*，2024年。（引自第1页。）

[2] X. Bing、F. Bunea、J. Niles-Weed 和 M. Wegkamp。使用预热期望最大化学习大型 Softmax 混合模型。*arXiv* 预印本 *arXiv:2409.09903*，2025年。（引自第3页。）

[3] E. Boix-Adsera 和 P. Rigollet。《细粒度专家的力量：粒度提升专家混合的表达能力》。*arXiv* 预印本 *arXiv:2505.06839*，2025年。（引用于第3页。）

[4] F. Chamroukhi、A. Samé、G. Govaert 和 P. Aknin。基于潜在过程的回归方法进行时间序列建模。神经网络，22(5-6):593–602，2009。（引用见第10页和第30页。）

[5] K. Chen、L. Xu 和 H. Chi。用于多类别分类的专家混合改进学习算法。《神经网络》，12(9):1229–1252，1999。（引用见第3页。）[6] Z. Chen、Y. Deng、Y. Wu、Q. Gu 和 Y. Li。探究深度学习中的专家混合层。载于 S. Koyejo、S. Mohamed、A. Agarwal、D. Belgrave、K. Cho 和 A. Oh（编），《神经信息处理系统进展》，第35卷，第23049–23062页。Curran Associates, Inc.，2022年。（引用见第1页和第3页。）[7] Z. Chi、L. Dong、S. Huang、D. Dai、S. Ma、B. Patra、S. Singhal、P. Bajaj、X. Song、X.-L. Mao、H. Huang 和 F. Wei。论稀疏专家混合中的表征坍塌。载于 A. H. Oh、A. Agarwal、D. Belgrave 和 K. Cho（编），《神经信息处理系统进展》，2022年。（引用见第2页。）[8] R. Csordás、K. Irie 和 J. Schmidhuber。近似两层前馈网络以实现高效 Transformer。载于 H. Bouamor、J. Pino 和 K. Bali（编），《计算语言学协会论文集：EMNLP 2023》，第674–692页，新加坡，2023年12月。计算语言学协会。（引用见第2页。）[9] D. Dai、C. Deng、C. Zhao、R. Xu、H. Gao、D. Chen、J. Li、W. Zeng、X. Yu、Y. Wu、Z. Xie、Y. Li、P. Huang、F. Luo、C. Ruan、Z. Sui 和 W. Liang。DeepSeekMoE：迈向专家混合语言模型中的极致专家专门化。载于《第62届计算语言学协会年会论文集》（第1卷：长篇论文），第1280–1297页。计算语言学协会，2024年8月。（引用见第3页。）[10] DeepSeek-AI 等。DeepSeek-v2：一种强大、经济且高效的专家混合语言模型。arXiv 预印本 arXiv:2405.04434，2024年。（引用见第2页。）[11] DeepSeek-AI 等。DeepSeek-v3 技术报告。arXiv 预印本 arXiv:2412.19437，2024年。（引用见第2页。）[12] A. P. Dempster、N. M. Laird 和 D. B. Rubin。通过期望最大化算法从不完整数据中进行最大似然估计。《英国皇家统计学会杂志：B辑（方法论）》，39(1):1–22，1977年。（引用见第10页和第30页。）[13] N. T. Diep、H. Nguyen、C. Nguyen、M. Le、D. M. H. Nguyen、D. Sonntag、M. Niepert 和 N. Ho。论零初始化注意力：最优提示与门控因子估计。载于《第42届国际机器学习大会》，2025年。（引用见第2页。）[14] S. Faria 和 G. Soromenho。拟合线性回归混合模型。《统计计算与模拟杂志》，80(2):201–225，2010年。（引用见第1页。）[15] GeminiTeam。Gemini 2.5：以高级推理、多模态、长上下文和新一代智能体能力拓展前沿。arXiv 预印本 arXiv:2507.06261，2025年。（引用见第1页。）[16] A. Grattafiori、A. Dubey、A. Jauhri、A. Pandey、A. Kadian、A. Al-Dahle、A. Letman、A. Mathur 等。llama 3 模型家族。arXiv 预印本 arXiv:2407.21783，2024年。（引用见第1页。）[17] B. Grün 和 F. Leisch。具有可变效应和固定效应的多项Logit模型有限混合模型的可识别性。《分类学杂志》，25(2):225–247，2008年。（引用见第29页。）

[18] X. Han、H. Nguyen、C. Harris、N. Ho 和 S. Saria。Fusemoe：用于灵活模态融合的混合专家Transformer。载于《神经信息处理系统进展》，2024年。（引自第2页。）[19] R. A. Jacobs、M. I. Jordan、S. J. Nowlan 和 G. E. Hinton。局部专家自适应混合。《神经计算》，第3卷，1991年。（引自第1页。）[20] A. Q. Jiang、A. Sablayrolles、A. Roux、A. Mensch、B. Savary、C. Bamford、D. S. Chaplot、D. de las Casas、E. B. Hanna、F. Bressand、G. Lengyel、G. Bour、G. Lample、L. R. Lavaud、L. Saulnier、M.-A. Lachaux、P. Stock、S. Subramanian、S. Yang、S. Antoniak、T. L. Scao、T. Gervet、T. Lavril、T. Wang、T. Lacroix 和 W. E. Sayed。Mixtral专家混合模型。arXiv预印本 arXiv 2401.04088，2024年。（引自第1页。）[21] J. Kwon 和 C. Caramanis。多个线性回归混合模型的期望最大化算法会收敛。载于 S. Chiappa 和 R. Calandra 编，《第二十三届国际人工智能与统计会议论文集》，《机器学习研究论文集》第108卷，第1727–1736页。PMLR，2020年8月。（引自第1页。）[22] M. Le、C. Nguyen、H. Nguyen、Q. Tran、T. Le 和 N. Ho。重新审视前缀微调：提示之间重参数化的统计优势。载于第十三届国际学习表征会议，2025年。（引自第2页。）[23] D. Lepikhin、H. Lee、Y. Xu、D. Chen、O. Firat、Y. Huang、M. Krikun、N. Shazeer 和 Z. Chen。GShard：利用条件计算和自动分片扩展巨型模型。载于国际学习表征会议，2021年。（引自第1页。）[24] B. Li、Y. Shen、J. Yang、Y. Wang、J. Ren、T. Che、J. Zhang 和 Z. Liu。稀疏专家混合是领域可泛化学习器。载于第十一届国际学习表征会议，2023年。（引自第2页。）[25] H. Liang、Z. Fan、R. Sarkar、Z. Jiang、T. Chen、K. Zou、Y. Cheng、C. Hao 和 Z. Wang。M^{3}ViT：通过模型-加速器协同设计实现高效多任务学习的混合专家视觉Transformer。载于 NeurIPS，2022年。（引自第1页。）[26] T. Manole 和 N. Ho。有限混合模型下最大似然估计的精细收敛速率。载于第39届国际机器学习会议论文集，《机器学习研究论文集》第162卷，第14979–15006页。PMLR，2022年7月17–23日。（引自第5页。）[27] H. Nguyen、P. Akbarian、T. Nguyen 和 N. Ho。softmax门控多项逻辑专家混合的一般理论。载于第41届国际机器学习会议论文集，第37617–37648页，2024年。（引自第2、3、6、11和31页。）[28] H. Nguyen、P. Akbarian、T. Pham、T. Nguyen、S. Zhang 和 N. Ho。在专家混合模型中扰动余弦路由器的统计优势。载于国际学习表征会议，2025年。（引自第2页。）[29] H. Nguyen、T. T. Doan、Q. Pham、N. D. Q. Bui、N. Ho 和 A. Rinaldo。论 DeepSeek-MoE：共享专家与归一化sigmoid门控的统计优势。arXiv预印本 arXiv:2505.10860，2025年。（引自第3页。）

[30] H. Nguyen、N. Ho 和 A. Rinaldo。在专家混合模型中，sigmoid 门控比 softmax门控更具样本效率。载于《神经信息处理系统进展》，2024年。（引自第2页。）[31] N. T. Pham 和 F. Chamroukhi。用于分类的函数型专家混合。arXiv预印本 arXiv:2202.13934，2022年。（引自第3页。）[32] Qwen 等。Qwen2.5技术报告。arXiv预印本 arXiv:2412.15115，2025年。（引自第1页。）[33] C. Riquelme、J. Puigcerver、B. Mustafa、M. Neumann、R. Jenatton、A. S. Pint、D. Keysers 和 N. Houlsby。利用稀疏专家混合扩展视觉能力。载于《神经信息处理系统进展》第34卷，第8583–8595页。Curran Associates, Inc.，2021年。（引自第1页。）[34] N. Shazeer、A. Mirhoseini、K. Maziarz、A. Davis、Q. Le、G. Hinton 和 J. Dean。超大型神经网络：稀疏门控专家混合层。载于国际学习表征会议，2017年。（引自第1页。）[35] T. Truong、C. Nguyen、H. Nguyen、M. Le、T. Le 和 N. Ho。RepLoRA：从专家混合的视角重参数化低秩适配。载于第42届国际机器学习大会，2025年。（引自第2页。）[36] S. van de Geer。《M估计中的经验过程》。剑桥大学出版社，2000年。（引自第25和28页。）[37] M. Wang 和 W. E. 结构化复杂任务中专家混合的表达能力。载于《第38届神经信息处理系统进展会议》（NeurIPS），2025年。亮点论文。（引自第3页。）[38] S. E. Yuksel、J. N. Wilson 和 P. D. Gader。专家混合的二十年。《IEEE神经网络与学习系统汇刊》，23(8):1177–1193，2012年。（引自第1页。）[39] S. Yun、I. Choi、J. Peng、Y. Wu、J. Bao、Q. Zhang、J. Xin、Q. Long 和 T. Chen。Flex-moe：通过灵活专家混合建模任意模态组合。载于第三十八届神经信息处理系统年会，2024年。（引自第2页。）