arXiv:2602.07997v1 [stat.ML] 8 Feb 2026

Softmax门控多项逻辑专家混合模型的快速模型选择与稳定优化

.

TrungKhang Tran [*1] **TrungTin Nguyen** [*2 3] **Md Abul Bashar** [2 3] **Nhat Ho** [4] **Richi Nayak** [2 3] **Christopher Drovandi** [2 3]

摘要

专家混合（MoE）架构通过学习得到的门控网络组合专门化预测器，在回归和分类任务中表现出色；然而，对于采用 softmax 多项逻辑门控的分类任务，关于稳定最大似然训练的严格保证和原则性模型选择的研究仍然有限。我们在全数据（批量）模式下同时解决这两个问题。首先，我们针对 softmax 门控多项逻辑混合专家模型，利用显式二次下界函数推导出一种批量极小化-极大化（MM）算法，得到逐坐标闭式更新，可保证目标函数单调上升，并全局收敛至驻点（遵循标准 MM 意义），避免了 EM 类实现中常见的近似 M 步。其次，我们证明了条件密度估计和参数恢复的有限样本速率，并将混合测度树状图改造用于分类问题，从而得到一种无需扫描的专家数量选择器；合并冗余拟合原子后，该选择器可达到近参数最优速率。在生物蛋白质-蛋白质相互作用预测任务上的实验验证了完整流程，与强大的统计和机器学习基线方法相比，该流程提高了准确率，并使概率校准效果更佳。

1. 引言

1.1. 专家混合模型

专家混合（MoE）模型（Jacobs等，1991年；Jordan与Jacobs，1994年）通过将门控模型与多个专家模型相结合来表示异质的预测变量—响应关系，其中二者都依赖于输入。门控网络为专家分配依赖于输入的权重，从而实现条件计算（Bengio，2013年；Chen等，2022年），并诱导出输入域的数据自适应划分。在分类任务中，Softmax门控多项逻辑专家混合（SGMLMoE）（Chen等，1999年；Yuksel与Gader，2010年；Huynh与Chamroukhi，2019年；Pham与Chamroukhi，2022年；Nguyen et al., 2024a）能够提供校准良好的多类别概率，同时保留门控机制的条件计算优势。MoE理论在多个情形下均得到近似与估计保证的支持，包括经典混合模型（Genovese与Wasserman，2000年；Rakhlin等，2005年；Nguyen，2013；Ho与Nguyen，2016a；b；Nguyen等，2020b；2022b；Chong等，2024年）、回归混合模型（Do等，2025年；Ho等，2022年）以及更广义的混合专家模型框架（Jiang与Tanner，1999b；Nguyen等，2024c；Norets，2010年；Nguyen等，2016；2019；2021a；2023a；2024a；b）；另请参阅相关综述（Yuksel等，2012年；Masoudnia与Ebrahim-pour，2014年；Nguyen与Chamroukhi，2018年；Nguyen，2021年；Cai等，2025年）。尽管取得了这些进展，对于SGMLMoE而言，仍有两个核心问题尚未得到充分研究：具有可证明收敛性的稳定最大似然训练，以及具有严格保证和计算效率的模型选择（尤其是专家数量的选择）。

1.2. 全数据情形下通过极小化-极大化实现稳定的最大似然训练

SGMLMoE 的基于似然的估计颇具挑战，因为观测数据目标函数以高度非凸的方式耦合了 Softmax 门控和多项逻辑专家。期望最大化（Dempster 等人，1977；Meilijson，1989；McLachlan 和 Krishnan，1997）是潜变量模型的经典范式，但在多项逻辑混合专家模型中，M 步通常没有闭式解，因此实际应用中的 EM 需要内循环或近似最大化，这可能使单调性和全局收敛保证变得复杂

[*]共同第一作者。 [1]新加坡国立大学计算机学院，新加坡。 [2]细胞系统数学分析卓越研究中心。 [3]昆士兰科技大学数学科学学院，布里斯班，澳大利亚。 [4]德克萨斯大学奥斯汀分校统计与数据科学系，奥斯汀，美国。通信作者：TrungTin Nguyen <t600.nguyen@qut.edu.au>。

预印本。*2026*年*2*月*10*日。

极小化-极大化算法（Ortega & Rheinboldt，1970年；De Leeuw，1977；De Leeuw & Heiser，1977；Lange，2016；Sun等人，2017年；Nguyen，2017；Lange等人，2021年；Mairal，2015；2013年）则优化显式替代函数，这些函数对目标函数进行上界或下界逼近，并可在保证稳定性的同时产生简单的更新。我们聚焦于批处理机制，避免 EM/MM 的增量式和小批量变体中固有的随机逼近噪声（Borkar，2008；Kushner & Yin，2003；Dieuleveut等人，2023年；Fort & Moulines，2023年；Wang等人，2019年；Capp´e & Moulines，2009；Capp´e，2011；Fort等人，2021a；2020；Nguyen等人，2020a；Kuhn等人，2020；Karimi等人，2019b；Corff & Fort，2013；Karimi等人，2019a；Oudoumanessah等人，2024年；2025年；Chen等人，2018年；Fort等人，2021b；Fort & Moulines，2021）以及相关的变分/自由能视角（Neal & Hinton，1998）。

1.3. SGMLMoE 模型中的模型选择

选择专家数量仍是核心问题。经典准则要求拟合多个模型规模，然后应用赤池信息准则（Akaike，1974；Fr¨uhwirth-Schnatter等人，2018）、BIC（Schwarz，1978；Khalili等人，2024；Forbes等人，2022a；Berrettini等人，2024；Forbes等人，2022b；Nguyen & Nguyen，2025；Ho等人，2025）、ICL（Biernacki等人，2000；Fr¨uhwirth-Schnatter等人，2012）、扩展贝叶斯信息准则（Foygel & Drton，2010；Nguyen& Li，2024）或SWIC（Sin & White，1996；Westerhout等人，2024）。对于混合专家模型，过度指定可能导致不可辨识性和奇异性，从而削弱通常的渐近理论依据。替代方法包括非渐近惩罚（Nguyen等人，2021b；2022a;c；2023b；Montuelle & Le Pennec，2014；Nguyen等人，2023c）以及贝叶斯/后处理策略，例如合并-截断-合并（Fr¨uhwirth-Schnatter，2019；Zens，2019；Guha等人，2021；Nguyen等人，2024d），但许多方法仍依赖多轮扫描。近期一种无需扫描的方法采用混合测度树状图（Do等人，2024；Thai等人，2025；Hai等人，2026），其主要针对高斯专家设定；我们将这一思路应用于多项逻辑分类，并将其与稳定的批量MM训练相结合。

1.4. 贡献与论文结构

我们为 SGMLMoE 开发了一个统一框架，将稳定的批量最大似然训练与轮次-自由模型选择。在优化方面，我们构造了一个显式的二次 MM 替代函数，从而得到逐坐标闭式更新；每次迭代都能精确地最小化该替代函数，因此无需内循环 M 步即可保证单调上升，并全局收敛到一个驻点。在统计方面，我们建立了条件密度估计和参数恢复的有限样本速率，并提出了一种基于树状图的混合测度聚合路径，可恢复近乎

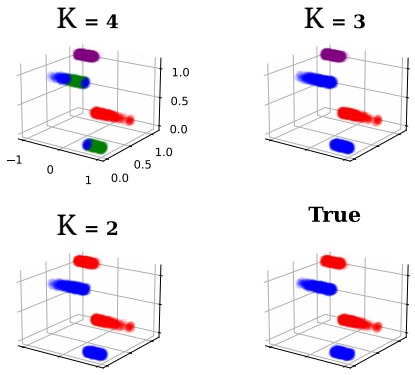


图 1. 过度指定拟合（$K = 4, M = 2$）上的 DSC 合并路径示意图：连续合并可移除近乎重复的专家，并恢复真实专家结构（$K_0 = 2, M = 2$）。

参数化速率（合并冗余拟合原子后；见表 1），并且无需多次K 训练即可得到一致的阶数选择器（见图 1）。

文章结构。 第 2 节 介绍 SGMLMoE 与似然目标函数。第 3 节给出 Batch MM 算法及其单调性。 第 4 节 建立 Voronoi 损失、聚合路径、收敛速率以及树状图选择准则。实验报告于 第 5 节。所有其余技术证明与实现细节均推迟至附录。

2. 预备知识

SGMLMoE 模型。 我们将输出 $\mathrm{y} \in \mathbb{Y}$ 称为响应变量，将输入 $\mathbf{x} \in \mathbb{X} \subset \mathbb{R}^P, P \in \mathbb{N}$称为预测变量。我们考虑一个规模为 N 的批量数据集，记为 $\{(\mathbf{x}_n, \mathrm{y}_n) : n \in [N]\}$，由 $(\mathbf{x},\mathrm{y})$ 的独立同分布副本组成，其中给定 $\mathbf{x} = \boldsymbol{x}$ 时 y 的条件分布由未知的条件概率质量函数（或密度） $\pi(\cdot \mid \mathbf{x} = \boldsymbol{x})$刻画。相应的观测值记为 $(\boldsymbol{x}, y)$。受专家混合模型通用逼近结果的启发，我们将$\pi(\cdot \mid \mathbf{x} = \boldsymbol{x})$逼近为如下形式的 SGMLMoE

$$s_{\boldsymbol{\theta}}(y \mid \boldsymbol{x}) = \sum_{k=1}^{K} \mathrm{g}_k\big(\boldsymbol{w}(\boldsymbol{x})\big)\, \mathrm{e}_k\big(y; \boldsymbol{v}(\boldsymbol{x})\big), \qquad (1)$$

其中， $\boldsymbol{\theta}$ 汇集了所有未知参数， K表示专家数量。门控概率由 softmax映射 $\mathrm{g}_k(\boldsymbol{w}(\boldsymbol{x})) = \exp\big(w_k(\boldsymbol{x})\big)/(\sum_{l=1}^{K} \exp\big(w_l(\boldsymbol{x})\big))$ $\forall k \in [K]$定义，其中门控评分函数为 $\boldsymbol{w}(\boldsymbol{x}) = (w_1(\boldsymbol{x}) \ldots, w_K(\boldsymbol{x}))$。在离散输出设定下，我们取 $\mathbb{Y} = [M], M \in \mathbb{N}$，并将每个专家指定为多项逻辑回归模型 $\mathrm{e}_k\big(y = m; \boldsymbol{v}(\boldsymbol{x}; \boldsymbol{v})\big) = \exp\big(v_{m,k}(\boldsymbol{x})\big) / \sum_{l=1}^{M} \exp\big(v_{l,k}(\boldsymbol{x})\big), \forall k \in [K], \forall m \in [M]$，其中专家评分 $\boldsymbol{v}(\boldsymbol{x}) =$

$(\upsilon_{m,k}(\boldsymbol{x}))_{m\in[M],k\in[K]}$ 被建模为关于 $\boldsymbol{x}$ 的多项式函数。具体而言，对于次数 $D_W, D_V \in \mathbb{N}$, $w_k(\boldsymbol{x}) = \sum_{d=0}^{D_W} \boldsymbol{\omega}_k^\top, {}_d\boldsymbol{x}^d = \sum_{d=0}^{D_W} \left(\sum_{p=1}^{P} \omega_{k,d,p} x_p^d\right)$, $\upsilon_{m,k}(\boldsymbol{x}) = \sum_{d=0}^{D_V} \boldsymbol{\upsilon}_{m,k}^\top, {}_d\boldsymbol{x}^d = \sum_{d=0}^{D_V} \left(\sum_{p=1}^{P} \upsilon_{m,k,d,p} x_p^d\right)$，其系数集合分别为 $\boldsymbol{\omega} = (\omega_{k,d,p})_{k\in[K], d\in\{0,\dots,D_W\}, p\in[P]}$ 和 $\boldsymbol{\upsilon} = (\upsilon_{m,k,d,p})_{m\in[M], k\in[K], d\in\{0,\dots,D_V\}, p\in[P]}$。完整参数向量记为 $\boldsymbol{\theta} = (\boldsymbol{\omega}, \boldsymbol{\upsilon})$。为简化起见，我们取 $D_W = D_V = D$，将一般情形 $D_W \neq D_V$ 留待未来研究。

最大对数似然估计量。我们将总体目标函数的一个最大化点记为 $\boldsymbol{\theta}^0 \in \log s_{\boldsymbol{\theta}}(\mathrm{y} \mid \mathbf{x})]\boldsymbol{\theta} \mapsto \mathbb{E}_{\mathrm{y}|\mathbf{x}\sim\mathbb{P}}[$。给定批量数据集 $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$，我们通过最大似然估计 $\boldsymbol{\theta}$，$\widehat{\boldsymbol{\theta}}_N \in \operatorname{argmax}_{\boldsymbol{\theta}\in\mathbb{T}} \frac{1}{N}\mathcal{L}(\boldsymbol{\theta})$，其中观测数据对数似然为

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log\left[\sum_{k=1}^{K} \mathsf{g}_k\big(\boldsymbol{w}(\boldsymbol{x}_n)\big)\, \mathsf{e}_k\big(y_n; \boldsymbol{v}(\boldsymbol{x}_n)\big)\right]. \quad (2)$$

本文采用全数据（批量）模式，重点关注公式（1）–公式（2）的稳定最大似然拟合，不使用增量或小批量变体中采用的随机近似方法。

SGMLMoE 模型的可辨识性。由于 softmax 门控对加性平移保持不变，门控参数并非完全可辨识。具体而言，若将每个系数向量平移 $\boldsymbol{\omega}_{k,d} \mapsto \boldsymbol{\omega}_{k,d} + \boldsymbol{t}$，其中 $\boldsymbol{t} \in \mathbb{R}^P$ 为某个值，则得到的门控概率 $\mathsf{g}_k(\boldsymbol{w}(\boldsymbol{x}))$ 保持不变，因此 $\{\boldsymbol{\omega}_{k,d}\}$ 仅在共同平移意义下可辨识。关于这一现象及其标准处理方法，(Nguyen et al., 2023a) 在讨论 softmax 门控高斯混合专家模型的最大似然估计收敛速率时进行了阐述。遵循 (Hennig, 2000; Jiang & Tanner, 1999a)，我们通过固定一个参考专家并施加以下约束来确保可辨识性，且不失一般性：$\{\boldsymbol{\omega}_{K,d,p}\}_{d\in\{0,\dots,D\}, p\in[P]} = \mathbf{0}$。由此得到等价参数化 $\mathsf{g}_K(\boldsymbol{x};\, \boldsymbol{\omega}) = 1 - \sum_{k=1}^{K-1} \mathsf{g}_k(\boldsymbol{x}; \boldsymbol{\omega})$，且 $\mathsf{g}_k(\boldsymbol{x};\, \boldsymbol{\omega}) = \exp(w_k(\boldsymbol{x};\, \boldsymbol{\omega}_k)) \Big/ \left(1 + \sum_{l=1}^{K-1} \exp(w_l(\boldsymbol{x}; \boldsymbol{\omega}_l))\right)$ 对所有 $k \in [K-1]$ 均成立。除门控网络的可辨识性约束外，我们还对多项逻辑回归专家网络采用一种可辨识性约定；(Nguyen et al., 2024a) 在分析 softmax 门控多项逻辑回归混合专家模型的收敛速率时也考虑了这一约定。具体而言，我们固定一个参考类别，并在不失一般性的前提下设定 $\{\upsilon_{M,k,d,p}\}_{d\in\{0,\dots,D\}, p\in[P]} = \mathbf{0}$ 对所有 $k \in [K]$ 均成立，因此最后一个类别的概率由其余 $M-1$ 个 logit 决定。等价地，对于每个专家 k，有 $\mathsf{e}_k(\boldsymbol{\upsilon}_{M,k}(\boldsymbol{x})) = 1 - \sum_{m=1}^{M-1} \mathsf{e}_k(\boldsymbol{\upsilon}_{m,k}(\boldsymbol{x}))$，且 $\mathsf{e}_k(\boldsymbol{\upsilon}_{m,k}(\boldsymbol{x})) = \exp(\upsilon_{m,k}(\boldsymbol{x})) \Big/ \left(1 + \sum_{l=1}^{M-1} \exp(\upsilon_{l,k}(\boldsymbol{x}))\right)$ 对所有 $m \in [M-1]$ 均成立。

3. SGMLMoE模型的极小化-极大化算法

现在，我们将极小化-极大化框架专门应用于公式（1）中的 SGMLMoE模型在离散输出 $\mathrm{y} \in \mathbb{Y} = [M]$下的全数据（批量）最大似然问题。令 $\{(\boldsymbol{x}_n, \mathrm{y}_n)\}_{n\in[N]}$ 为$(\mathbf{x}, \mathrm{y})$ 的独立同分布副本，并回顾 $\mathcal{L}(\boldsymbol{\theta})$ 中公式（2）的观测数据对数似然。对于每个数据点 $n \in [N]$ 和专家 $k \in [K]$，定义当前迭代点处的常见后验责任度 $\boldsymbol{\theta}^{(t)}$ 如下：

$$\tau_{n,k}^{(t)} = \frac{\mathsf{g}_k\big(\boldsymbol{w}^{(t)}(\boldsymbol{x}_n)\big)\ \mathsf{e}_k\Big(\boldsymbol{v}_{\mathrm{y}_n}^{(t)}(\boldsymbol{x}_n)\Big)}{\sum_{l=1}^{K} \mathsf{g}_l\big(\boldsymbol{w}^{(t)}(\boldsymbol{x}_n)\big)\ \mathsf{e}_l\Big(\boldsymbol{v}_{\mathrm{y}_n}^{(t)}(\boldsymbol{x}_n)\Big)}. \quad (3)$$

为简化多项式特征的记号，令

$$\hat{\boldsymbol{x}}_n = \left[x_{n,1}^0, \dots, x_{n,1}^D, \dots, x_{n,P}^0, \dots, x_{n,P}^D\right]^\top \in \mathbb{R}^{P(D+1)}.$$

接下来，我们使用指示多项式 $\mathbf{1}(z, l)$ 来编码类别标签：$\mathbf{1}(z, l) = \prod_{q=1}^{M} (z-q)\,(z-l)!\,(M-z)!\,(-1)^{M-z}, \forall z, l \in [M]$，使得 $\mathbf{1}(z, l) \in \{0, 1\}$。这种表示便于编码，并满足恒等式 $\upsilon_{m,k}(\boldsymbol{x}) = \sum_{l=1}^{M} \mathbb{1}(m, l)\upsilon_{l,k}(\boldsymbol{x})$，据此可从 logit 向量恢复任意类别对数几率。

批量充分统计量记号。对于每个 $n \in [N]$，在 $\mathbb{R}^{(K-1)P(D+1)}$ 中定义门控设计向量：

$$\boldsymbol{s}_n^{(t)} = \left[\tau_{n,1}^{(t)} x_{n,1}^0, \dots, \tau_{n,1}^{(t)} x_{n,1}^D, \dots, \tau_{n,K-1}^{(t)} x_{n,P}^D\right]^\top, \quad (4)$$

并通过堆叠 $r_{n\,d',l,p,k}^{(t)} = \mathbb{1}(\mathrm{y}_n, l)\, \tau_{n\,,k}^{(t)} x_{n,p}^d$，$d \in \{0, \dots, D\}$，$l \in [M]$，$p \in [P]$，$k \in [K]$，到 $\boldsymbol{r}_n^{(t)} = \left[r_{n,0,1,1,1}^{(t)}, \dots, r_{n,D,1,1,1}^{(t)}, \dots, r_{n,D,M,1,1}^{(t)}, \dots, r_{n,D,M,P,K}^{(t)}\right]^\top$ 中，来定义专家设计向量$\boldsymbol{r}_n^{(t)}$

我们还定义每个专家的类别参数块k 为 $\boldsymbol{c}_k = \left[\upsilon_{1,k,0,1}, \dots, \upsilon_{1,k,D,1}, \dots, \boldsymbol{\upsilon}_{M,k,D,1}, \dots, \upsilon_{M,k,D,P}\right]^\top$，$\boldsymbol{\upsilon} = \operatorname{vec}([\boldsymbol{c}_1, \dots, \boldsymbol{c}_K])$. 最后，我们将批量统计量汇总为 $\boldsymbol{s}^{(t)} = \sum_{n=1}^{N} \boldsymbol{s}_n^{(t)}, \boldsymbol{r}^{(t)} = \sum_{n=1}^{N} \boldsymbol{r}_n^{(t)}$.

门控与专家的二次主化函数。对于 $k \in [K-1]$，令 $w_k(\boldsymbol{x}) = \sum_{d=0}^{D} \boldsymbol{\omega}_{k,\ d}^\top {}_d\boldsymbol{x}^d$ 并定义门控 log-sum-exp 项 $g_n(\boldsymbol{w}) = \log\left(1 + \sum_{k=1}^{K-1} \exp(w_k(\boldsymbol{x}_n))\right)$. 对于专家，对每个 $k \in [K]$定义多项式 log-sum-exp 项 $e_n(\boldsymbol{c}_k) = \log\left(1 + \sum_{m=1}^{M-1} \exp(\upsilon_{m,\ k}(\boldsymbol{x}_n))\right)$. 对于任意 $K \in \mathbb{N}$，我们使用基于以下矩阵的二次主化函数

$$\boldsymbol{B}_{n,K} = \left(\frac{3}{4}\boldsymbol{I}_{K-1} - \frac{\mathbf{1}_{K-1}\mathbf{1}_{K-1}^\top}{2(K-1)}\right) \otimes \hat{\boldsymbol{x}}_n \hat{\boldsymbol{x}}_n^\top. \quad (5)$$

算法1 SGMLMoE 批量 MM 算法 要求：数据 $\{(x_n, y_n)\}_{n=1}^N$；(K, M)；次数 D；初始值 $\theta^{(0)} = (w^{(0)}, v^{(0)})$；容差 ε；最大迭代次数 T。确保：$\{\theta^{(t)}\}_{t\geq 0}$。1：对于 $t = 0, 1, \ldots, T-1$，执行 2：E 步。对于每个 $n \in [N]$，根据公式（3）计算责任度$\{\tau_{n,k}^{(t)}\}_{k\in[K]}$。3：使用公式（4）和 $r_n^{(t)}$ 的定义（其中 $\mathbf{1}(y_n, l)$），构造充分统计量 $s^{(t)} = \sum_n s_n^{(t)}$ 和 $r^{(t)} = \sum_n r_n^{(t)}$。4：根据公式（5），利用分块 $B_{n,K}, B_{n,M}$ 构造曲率矩阵 $B_{n,K-1} = \sum_n B_{n,K}$ 和 $B_{n,M-1} = \sum_n B_{n,M}$。5：计算 $\nabla g(w^{(t)}) = \sum_n \nabla g_n(w^{(t)})$ 以及专家 log-sum-exp 梯度 $\{\nabla e_n(v^{(t)})\}_{n\in[N]}$（定义见公式（6））。6：M 步。通过公式（6）中替代函数 $\mathcal{S}(\theta, \theta^{(t)})$ 的闭式极小化解更新 (w, v)：$w^{(t+1)} = w^{(t)} + B_{n,K-1}^{-1}\left(s^{(t)} - \nabla g(w^{(t)})\right)$，$v^{(t+1)} = \arg\min_v \mathcal{S}((w^{(t+1)}, v), \theta^{(t)})$，其中 $v^{(t+1)}$ 的显式分块线性求解见附录B。7：设置 $\theta^{(t+1)} = (w^{(t+1)}, v^{(t+1)})$。8：停止。若 $|\mathcal{L}(\theta^{(t+1)}) - \mathcal{L}(\theta^{(t)})| \leq \varepsilon$，则退出。9：循环结束

为简洁起见，引入中心化增量 $\bar{w}_t = w - w^{(t)}$ 和 $\bar{c}_{k,t} = c_k - c_k^{(t)}$。

定理1（批量负对数似然的替代函数。）。以下界定义了 $-\mathcal{L}(\theta)$ 的批量 *MM* 替代函数，证明推迟至*D* 节.2：

$$-\mathcal{L}(\theta) \leq C^{(t)} + \mathcal{S}_1(\theta, \theta^{(t)}) + \mathcal{S}_2(\theta, \theta^{(t)}) \qquad (6)$$

Here $C^{(t)}$ *is an independent constant w.r.t* θ, $\mathcal{S}_1(\theta, \theta^{(t)}) = \sum_{n=1}^N \{g_n(w^{(t)}) + \bar{w}_t^\top \nabla g_n(w^{(t)}) + \frac{1}{2}\bar{w}_t^\top B_{n,K}\bar{w}_t\}$, *and* $\mathcal{S}_1(\theta, \theta^{(t)}) = \sum_{n=1}^N \sum_{k=1}^K \tau_{n,k}^{(t)}\{e_n(c_k^{(t)}) + \bar{c}_{k,t}^\top \nabla e_n(c_k^{(t)}) + \frac{1}{2}\bar{c}_{k,t}^\top B_{n,M}\bar{c}_{k,t}\}$.

算法2中的这些更新算法2精确地极小化替代函数公式（6），因此依据定理2，继承了观测数据似然的标准 MM 单调上升性质。该定理的证明基于 $\mathcal{S}$ 的 MM 优超-相切性质以及精确极小化；参见附录中的D 节.3 。

定理*2*（批量 *SGMLMoE* 的 *MM* 单调性）。设$\mathcal{S}(\theta, \theta^{(t)})$为方程（6）中 $-\mathcal{L}(\theta)$的批量 *MM* 替代函数，且 $\theta^{(t+1)}$ 为算法*1*的输出。则$\theta^{(t+1)} \in \arg\min_\theta \mathcal{S}(\theta, \theta^{(t)})$，且 $\mathcal{L}(\theta^{(t+1)}) \geq \mathcal{L}(\theta^{(t)})$对所有 t成立。

4. 面向快速率的聚合与 SGMLMoE 的免扫描模型选择

4.1. 过拟合下的“速率差距”

在SGMLMoE 分类模型方程（1）–方程（2）中，基于似然的估计具有挑战性，其原因与softmax 门控高斯专家中出现的结构性原因相同（Hai等人，2026年），但此处的专家为多项逻辑专家：(i)*softmax* 平移不变性使得门控参数仅在相差一个公共平移的意义下可识别；(ii)门控—专家耦合在 $s_\theta(\cdot \mid x)$ 的局部展开中产生非平凡的相互作用；(iii)*softmax* 分子—分母耦合会产生抵消现象，从而即使各个分量本身并不接近，密度也可能彼此接近。

当拟合模型设定过度 $(K > K_0)$时，这些效应最为明显：多个拟合原子可能聚集在同一个真实专家周围。在此情况下，条件密度 ${}^{s}\hat{G}_N(\cdot \mid x)$ 的全变差收敛速度可以达到近参数速率，而分量级参数沿近不可辨识方向的收敛速度则较慢。本节旨在：(a) 引入一种与门控诱导的划分几何相匹配的沃罗诺伊型损失，以解决这些收敛较慢的方向；(b) 利用该损失构建一条层次聚合路径（即混合测度的树状图），从而得到一种无需扫描且一致的 专家数量选择器。

实践意义（单次拟合工作流）。拟合一个设定过度的 SGMLMoE（中等 $K \geq K_0$），使用算法2中的批量 MM 程序，然后通过迭代合并构建其树状图 $\widehat{\ }$，并通过高度—似然树状图选择准则（DSC）选择 K。这样就无需针对多个候选规模训练多个模型。

4.2. SGMLMoE 的混合测度表示

提升特征与截距/斜率分解。固定多项式次数 D，定义非恒定提升特征映射 $\phi_D(x) := \mathrm{vec}([x^1, x^2, \ldots, x^D]) \in \mathbb{R}^{PD}$, $\widehat{\mathbb{X}} := \phi_D(\mathbb{X}) \subset \mathbb{R}^{PD}$.（由于 $\mathbb{X}$ 是紧集且 ϕ_D 连续，根据海涅-博雷尔定理，$\mathbb{X}$ 是紧集）。对于每个门控网络 $k \in [K-1]$，将 $w_k(x) = \bar{\omega}_k + \hat{\omega}_k^\top \hat{x}$, $\hat{x} := \phi_D(x)$ 分解，其中 $\bar{\omega}_k := \sum_{p=1}^P \omega_{k,0,p} \in \mathbb{R}$ 且 $\hat{\omega}_k := \mathrm{vec}([\omega_{k,1}, \ldots, \omega_{k,D}]) \in \mathbb{R}^{PD}$。对于每个专家 $k \in [K]$ 和类别 $m \in [M-1]$，同样写作 $v_{m,k}(x) = \bar{v}_{m,k} + \hat{v}_{m,k}^\top \hat{x}$，其中 $\bar{v}_{m,k} := \sum_{p=1}^P v_{m,k,0,p} \in \mathbb{R}$ 且 $\hat{v}_{m,k} := \mathrm{vec}([v_{m,k,1}, \ldots, v_{m,k,D}]) \in \mathbb{R}^{PD}$。

可识别性约定（如第 2 节所述）。我们固定参考专家和参考类别：$(\bar{\omega}_K, \hat{\omega}_K) = (0, 0)$，$(\bar{v}_{M,k}, \hat{v}_{M,k}) = (0, 0)$ $\forall k \in [K]$.

混合测度。令 $\mathbb{T}$ 表示可辨识的

单个 SGMLMoE 原子的参数空间：$\eta = \big(\hat{\boldsymbol{\omega}}, \bar{\omega}, \{(\hat{\boldsymbol{v}}_m, \bar{v}_m)\}_{m=1}^{M-1}\big) \in \mathbb{T}$，并令 $\mathcal{O}_K(\mathbb{T})$ 表示至多包含 K 个原子的有限混合测度集合：$\mathcal{O}_K(\mathbb{T}) = \Big\{G = \sum_{i=1}^{K} \pi_i \delta_{\eta_i} : \ 1 \le K \le K, \ \pi_i > 0, \ \eta_i \in \mathbb{T}\Big\}$. 我们将权重参数化为 $\pi_i = \exp(\bar{\omega}_i)$（因此 $\bar{\omega}_i = \log \pi_i$）。

Conditional Probability Mass Function Induced by G. Given $G = \sum_{i=1}^{K} \pi_i \delta_{\eta_i}$, define for $\hat{\boldsymbol{x}} = \phi_D(\boldsymbol{x})$ $\mathsf{g}_i(\boldsymbol{x}; G) = \frac{\pi_i \exp(\hat{\boldsymbol{\omega}}_i^\top \hat{\boldsymbol{x}})}{\sum_{j=1}^{K} \pi_j \exp(\hat{\boldsymbol{\omega}}_j^\top \hat{\boldsymbol{x}})}$, $\mathsf{e}_i(y = m \mid \boldsymbol{x}; G) = \frac{\exp(\bar{v}_{m,i} + \hat{\boldsymbol{v}}_{m,i}^\top \hat{\boldsymbol{x}})}{1 + \sum_{\ell=1}^{M-1} \exp(\bar{v}_{\ell,i} + \hat{\boldsymbol{v}}_{\ell,i}^\top \hat{\boldsymbol{x}})}$, with $\mathsf{e}_i(y = M \mid \boldsymbol{x}; G) = 1 - \sum_{m=1}^{M-1} \mathsf{e}_i(y = m \mid \boldsymbol{x}; G)$, and $s_G(y \mid \boldsymbol{x}) = \sum_{i=1}^{K} \mathsf{g}_i(\boldsymbol{x}; G)\, \mathsf{e}_i(y \mid \boldsymbol{x}; G)$. For any identifiable $\boldsymbol{\theta}$, there exists $G(\boldsymbol{\theta}) \in \mathcal{O}_K(\mathbb{T})$ such that $s_{G(\boldsymbol{\theta})} \equiv s_{\boldsymbol{\theta}}$.

至多包含 K 个专家的最大似然估计。 令 $G_0 \in \mathcal{O}_{K_0}(\mathbb{T})$ 为真实混合测度（$2 \le K_0 \le K$）。定义（混合测度）最大似然估计 $G_N \in \operatorname{argmax}_{G \in \mathcal{O}_K(\mathbb{T})} \frac{1}{N} \sum_{n=1}^{N} \log s_G(y_n \mid \boldsymbol{x}_n)$。

4.3. SGMLMoE的沃罗诺伊单元与损失

从“密度接近”到“参数接近”。 过拟合混合模型/MoE分析的一个关键步骤是建立逆不等式：s_G 与 s_{G_0} 之间的差异较小，应意味着在适应几何结构的损失下，G与 G_0 之间的差异也较小。对于Softmax门控混合专家模型，适用的几何结构是沃罗诺伊结构：将拟合原子归入每个真实原子周围的单元中。

沃罗诺伊划分。 记 $G = \sum_{i=1}^{K} \pi_i \delta_{\eta_i}$ 和 $G_0 = \sum_{k=1}^{K_0} \pi_k^0 \delta_{\eta_k^0}$。当 G 与 G_0 足够接近时（例如满足下文所述的条件），定义一个匹配 $i \mapsto k(i) \in [K_0]$，将每个拟合原子 η_i 按照 $\mathbb{T}$上的局部度量，分配给距离最近的真实原子 $\eta_{k(i)}^0$。定义沃罗诺伊单元 $\mathbb{V}_k := \{ i \in [K] : \ k(i) = k \}, k \in [K_0]$.

积分全变差差异。 对于 $\boldsymbol{x} \in \mathbb{X}$，定义 $\mathcal{D}_{\mathrm{TV}}(s_G(\cdot \mid \boldsymbol{x}) s_{G_0}(\cdot \mid \boldsymbol{x}))$ 为 $\frac{1}{2} \sum_{m=1}^{M} |s_G(m \mid \boldsymbol{x}) - s_{G_0}(m \mid \boldsymbol{x})|$，其协变量平均形式$\mathcal{D}_{\mathrm{TV}}(s_G, s_{G_0})$ 定义为 $\mathbb{E}_{\mathbf{x}}[\mathcal{D}_{\mathrm{TV}}(s_G(\cdot \mid \mathbf{x}) \ s_{G_0}(\cdot \mid \mathbf{x}))]$.

SGMLMoE 沃罗诺伊损失。 对于 $i \in \mathbb{V}_k$，定义参数差异 $\Delta\hat{\boldsymbol{\omega}}_{ik} := \hat{\boldsymbol{\omega}}_i - \hat{\boldsymbol{\omega}}_k^0$，$\Delta\bar{v}_{ikm} := \bar{v}_{m,\,i} - \bar{v}_{m,\,k}^0$，$\Delta\hat{\boldsymbol{v}}_{ikm} := \hat{\boldsymbol{v}}_{m,\,i} - \hat{\boldsymbol{v}}_{m,\,k}^0$，$m \in [M-1]$，以及损失

$$\mathcal{D}_{\mathrm{V}}(G, G_0) = \mathcal{D}_{\mathrm{E}}(G, G_0) + \sum_{\substack{k \in [K_0]: |\mathbb{V}_k| > 1 \\ i \in \mathbb{V}_k, m \in [M-1]}} \pi_i [\|\Delta\hat{\boldsymbol{\omega}}_{ik}\|^2 + (|\Delta\bar{v}_{ikm}|^2 + \|\Delta\hat{\boldsymbol{v}}_{ikm}\|^2)], \tag{7}$$

where $\mathcal{D}_{\mathrm{E}}(G, G_0) := \sum_{k=1}^{K_0} \Big| \sum_{i \in \mathbb{V}_k} \pi_i - \pi_k^0 \Big| + \sum_{\substack{k \in [K_0]: |\mathbb{V}_k| = 1 \\ i \in \mathbb{V}_k, m \in [M-1]}} \pi_i [\|\Delta\hat{\boldsymbol{\omega}}_{ik}\| + (|\Delta\bar{v}_{ikm}| + \|\Delta\hat{\boldsymbol{v}}_{ikm}\|)]$.

局部结构不等式（逆界）。 下一个结果是为 SGMLMoE 建立的“密度 $\Rightarrow$ 沃罗诺伊”逆界。

定理 3。 对于 $G \in \mathcal{O}_K(\mathbb{T})$，当 $\mathcal{D}_{\mathrm{TV}}(s_G, s_{G_0}) \to 0$，

$$\mathcal{D}_{\mathrm{TV}}(s_G, s_{G_0}) \gtrsim \mathcal{D}_{\mathrm{V}}(G, G_0). \tag{8}$$

4.4. 为什么合并专家？树状图视角

过拟合图景。 当 $K > K_0$ 时，某些沃罗诺伊单元$\mathbb{V}_k$ 可能包含多个拟合原子。从条件密度的角度来看，这些原子是“近乎重复”的，并会在似然几何中造成缓慢或平坦的方向。树状图机制通过反复合并来折叠每个这样的簇，从而生成一系列混合测度的层级结构 $\widehat{G}_N^{(K)} = \widehat{G}_N \rightsquigarrow \widehat{G}_N^{(K-1)} \rightsquigarrow \cdots \rightsquigarrow \widehat{G}_N^{(2)}$. 我们利用的关键事实是：(i) 每次合并都会减少冗余；(ii) 沿着合并链，Voronoi 损失会变得越来越容易（单调地）。

原子差异度（合并准则）。 对于两个原子$\pi_i \delta_{\eta_i}$ 和 $\pi_j \delta_{\eta_j}$，定义

$$\mathrm{d}(\pi_i \delta_{\eta_i}, \pi_j \delta_{\eta_j}) := \frac{\pi_i \pi_j}{\pi_i + \pi_j} \Big(\|\hat{\boldsymbol{\omega}}_i - \hat{\boldsymbol{\omega}}_j\|^2 + \sum_{m=1}^{M-1} \big(\|\hat{\boldsymbol{v}}_{m,i} - \hat{\boldsymbol{v}}_{m,j}\|^2 + |\bar{v}_{m,i} - \bar{v}_{m,j}|^2 \big) \Big). \tag{9}$$

合并算子（权重平均原子）。 合并原子 i 和 j 会通过参数重心产生一个新原子 $(\pi_*, \boldsymbol{\eta}_*)$，对于 $\boldsymbol{\eta}_*$，具体如下：给定任意 $m \in [M-1]$,

$$\pi_* = \pi_i + \pi_j, \ \bar{\omega}_* = \log \pi_*, \hat{\boldsymbol{v}}_{m,*} = \frac{\pi_i}{\pi_*} \hat{\boldsymbol{v}}_{m,i} + \frac{\pi_j}{\pi_*} \hat{\boldsymbol{v}}_{m,j},$$
$$\bar{v}_{m,*} = \frac{\pi_i}{\pi_*} \bar{v}_{m,i} + \frac{\pi_j}{\pi_*} \bar{v}_{m,j}, \ \hat{\boldsymbol{\omega}}_* = \frac{\pi_i}{\pi_*} \hat{\boldsymbol{\omega}}_i + \frac{\pi_j}{\pi_*} \hat{\boldsymbol{\omega}}_j. \tag{10}$$

定理 4（沿合并链的沃罗诺伊单调性）。设 $G^{(K)}$, $G^{(K-1)}$, ..., $G^{(K_0)}$ 通过根据式 (9)反复合并最近的一对原子而得到，从 $G^{(K)} \in \mathcal{O}_K(\mathbb{T})$ 开始。对于 $G^{(K)}$ 充分接近 G_0，$\mathcal{D}_{\mathrm{V}}(G^{(K)}, G_0) \gtrsim \mathcal{D}_{\mathrm{V}}(G^{(K-1)}, G_0) \gtrsim \cdots \gtrsim \mathcal{D}_{\mathrm{V}}(G^{(K_0)} \ G_0)$.

4.5. 沿树状图的有限样本速率

高度（结构信号）。 在层级 κ，定义树状图高度 $h_N^{(\kappa)} := \min_{i \neq j \text{ 的原子 } \widehat{G}_N^{(\kappa)}} \mathrm{d}(\pi_i \delta_{\eta_i}, \pi_j \delta_{\eta_j})$，$\kappa \in [K]$，并令 $h^{(\kappa')}$ 0 为真实链 $G^{(\kappa')}$ 0 上的类似高度，$\kappa' \in [K_0]$。

最大似然估计的密度速率。 我们采用已知的近参数界（至对数因子 $\tilde{\mathcal{O}}$（对数因子 N/N）$^{1/2}$)），该界针对多项分类情形下的全变差最大似然估计已由（Nguyen等人，2024年a）确立：

事实 1(密度收敛速率)。存在通用常数 $C_1, C_2 > 0$，使得 $\mathbb{P}(\mathcal{D}_{\mathrm{TV}}({}^{S}\widehat{G}_N, s_{G_0}) > C_1\sqrt{\text{对数 } N/N}) \lesssim N^{-C_2}$.

从密度速率到沃罗诺伊速率和高度速率。 结合 定理 3 与 事实 1 可得层级K上的沃罗诺伊控制，而 定理 4 中的单调性则沿路径向下传递：

定理 5（沿路径的沃罗诺伊距离与高度速率）。以至少 $1 - c_1 N^{-c_2}$ 的概率，对每个 $\kappa \in \{K_0 + 1, \ldots, K\}$ 和 $\kappa' \in [K_0]$, $\max\{\mathcal{D}_{\mathrm{V}}(\widehat{G}_N^{(\kappa)}, G_0)\ h_N^{(\kappa)},\ |h^{(\kappa')}_N - h^{(\kappa')0}|\} \lesssim \sqrt{\text{对数 } N/N}$.

4.6. 沿路径的似然与免扫描选择

经验与总体对数似然。对于 $G \in \mathcal{O}_K(\mathbb{T})$，定义 $\mathcal{L}(G) := \mathbb{E}_{(\mathbf{x},\mathrm{y})\sim G_0}[\text{对数 } s_G(\mathbf{y} \mid \mathbf{x})]$，以及 $\bar{\ell}_N(G) := \frac{1}{N}\sum_{n=1}^{N} \text{对数 } s_G(y_n \mid \boldsymbol{x}_n)$。记 $\bar{\ell}_N^{(\kappa)} := \bar{\ell}_N(\widehat{G}_N^{(\kappa)})$。

条件 K（局部下界）。 存在常数c_α, $c_\beta > 0$，使得对所有充分小的 $\epsilon > 0$ 以及所有满足 $\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\| \le \epsilon$的 $\boldsymbol{\theta}$, $\boldsymbol{\theta}^0 \in \mathbb{T}$，对所有$(\mathbf{x}, \mathrm{y}) \in \mathbb{X} \times \mathbb{Y}$, $\log s_{\boldsymbol{\theta}}(\mathrm{y} \mid \mathbf{x}) \ge (1 + c_\beta\epsilon) \log s_{\boldsymbol{\theta}^0}(\mathrm{y} \mid \mathbf{x}) - c_\alpha\epsilon$.

定理 6（沿树状图的似然控制）。假设定理 *5* 与条件 *K* 成立。则对任意 $\kappa \in \{K_0 + 1, \ldots, K\}$, $\bar{\ell}_N(\widehat{G}_N^{(\kappa)}) - \mathcal{L}(G_0) \lesssim (\log N/N)^{1/4}$. 此外，对任意 $\kappa' \in [K_0]$, $\bar{\ell}_N(\widehat{G}^{(\kappa')}{}_N) \to \mathcal{L}(G^{(\kappa')}{}_0)$ 当 $N \to \infty$ 时依 $\mathbb{P}_{G_0}$- 概率。

DSC（高度 +似然）。给定 $1 \ll \omega_N \ll (N/\log N)^{1/4}$，定义树状图选择准则 $\mathrm{DSC}_N^{(\kappa)} := -\left(h_N^{(\kappa)} + \omega_N\, \bar{\ell}_N^{(\kappa)}\right)$。一个实用的选择是$\omega_N = \log N$。选择 $\widehat{K}_N := \arg\min_{\kappa \in \{2,\ldots,K\}} \mathrm{DSC}_N^{(\kappa)}$.

定理 7（DSC 对 SGMLMoE 的一致性）。假设$K_0 \ge 2$且定理 *6* 的条件 $K_N \to K_0$ 成立，且 $\mathbb{P}_{G_0}$-概率收敛，当 $N \to \infty$ 时。

为什么 DSC 在过拟合时优于纯似然惩罚的解释。高度能够检测结构冗余：较小的 $h_N^{(\kappa)}$ 表示 κ-原子模型包含一对几乎无法区分的原子（通常由过度设定造成），而标准准则（AIC/BIC/ICL）不会直接对其进行惩罚。DSC 将这一结构信号与拟合效果结合起来，从而得到一个与沃罗诺伊几何相契合、无需扫描的选择器。

5. 数值实验

本节在合成数据集和真实世界数据集上开展实验，系统评估所提出方法的有效性、稳定性和实用价值。

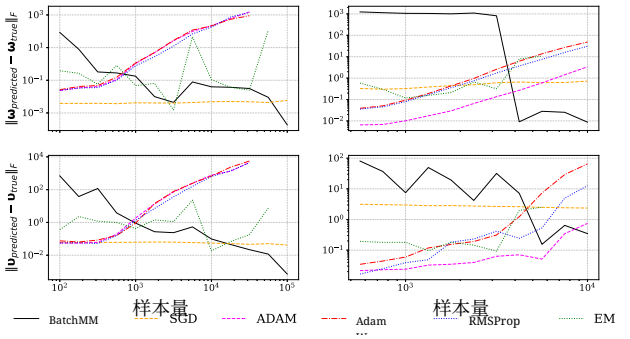


图 *2.* 批量MM算法的收敛行为（参数误差与样本量的关系）。

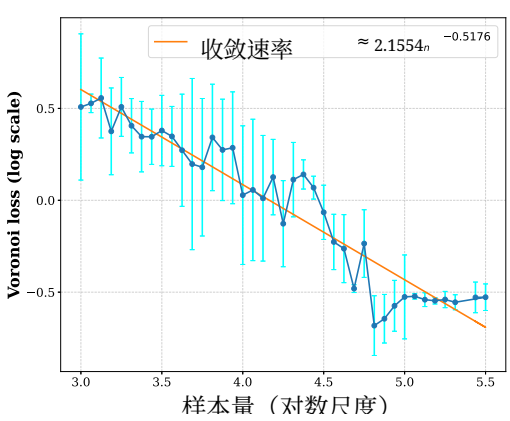
流程：通过批量MM进行训练、使用Voronoi损失进行诊断，并基于树状图通过DSC进行模型选择与合并。在合成数据集上，我们首先证明，所提出的批量MM算法能够可靠地收敛至全局最优；而包括期望最大化和常用的基于梯度的优化器在内的经典优化方法，在正确设定和过度参数化两种情形下，往往会陷入次优驻点。随后，我们通过实证验证，在精确拟合和过拟合设定下，Voronoi损失的衰减均与定理 5中确立的有限样本收敛速率一致。最后，我们展示了所提出的基于DSC的合并策略在恢复真实专家数量方面始终优于标准模型选择准则（赤池信息准则、BIC和ICL），尤其是在小样本情形下；同时，随着样本量增大，该策略仍保持一致性。在真实世界数据集上，我们展示了通过批量MM训练的SGMLMoE，其预测性能优于支持向量机和随机森林等强基线方法。此外，DSC能够选出恰当的专家数量，在预测准确率与内存效率之间实现理想的权衡；所得树状图则通过图5，以可解释的方式对拟合的混合测度进行层次化概括。

5.1. 模拟研究

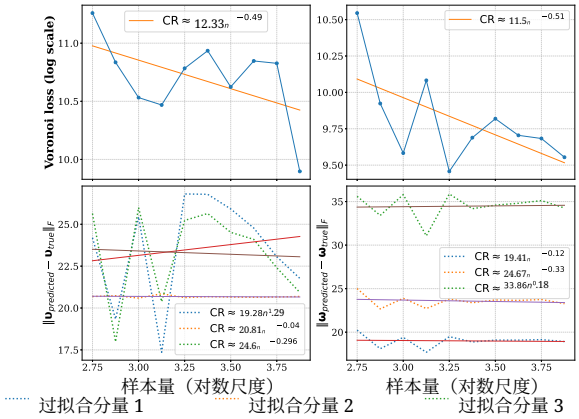
数据集。我们考虑一个受控的合成设定，其中多项式次数为 $D = 1$，专家数量为 $K_0 = 2$，类别数量为 $M = 2$，输入维度为 $P = 1$。真实门控参数为$\boldsymbol{\omega}_{0,0,0} = \boldsymbol{\omega}_{1,0,0} = \boldsymbol{\omega}_{1,1,0} = 0$ 和$\boldsymbol{\omega}_{0,0,1} = 8$，专家参数为$\boldsymbol{\upsilon}_{0,0,0,0} = \boldsymbol{\upsilon}_{1,0,1,0} = \boldsymbol{\upsilon}_{0,1,0,0} = \boldsymbol{\upsilon}_{1,1,1,0} = 0\boldsymbol{\upsilon}_{0,0,1,0} = \boldsymbol{\upsilon}_{0,1,1,0} = 20$以及$\boldsymbol{\upsilon}_{1,0,0,0} = -\boldsymbol{\upsilon}_{1,1,0,0} = 10$。为生成数据集 $(\mathbb{X}, \mathbb{Y})$，我们首先固定样本量 N（该值因实验而异）。随后，从标准正态分布$\mathcal{N}(0, 1)$中独立同分布地抽取输入 $\mathbb{X} = \{x_n\}_{n=1}^N$。给定 $\mathbb{X}$，我们根据由 $(\boldsymbol{\omega}_0, \boldsymbol{\upsilon}_0)$ 指定的SGMLMoE模型生成离散输出$\mathbb{Y} = \{y_n\}_{n=1}^N$。

表 *1.* 采用标准 Softmax 门控的 SGMLMoE 的（对数调整）密度和参数速率汇总。

设定	密度	精确指定参数 ($\lvert\mathbb{V}_k\rvert = 1$)	过度指定参数 ($\lvert\mathbb{V}_k\rvert > 1$)
精确拟合	$\widetilde{\mathcal{O}}((\text{对数 } N/N)^{1/2})$	$\widetilde{\mathcal{O}}((\text{对数 } N/N)^{1/2})$	—
过拟合	$\widetilde{\mathcal{O}}((\text{对数 } N/N)^{1/2})$	$\widetilde{\mathcal{O}}((\text{对数 } N/N)^{1/2})$	$\widetilde{\mathcal{O}}((\text{对数 } N/N)^{1/4})$
合并	$\widetilde{\mathcal{O}}((\text{对数 } N/N)^{1/2})$	$\widetilde{\mathcal{O}}((\text{对数 } N/N)^{1/2})$	$\widetilde{\mathcal{O}}((\text{对数 } N/N)^{1/2})$



(a) 精确拟合：沃罗诺伊损失的经验收敛。



*(b)*过拟合：沃罗诺伊损失的经验收敛。

图*3*。 精确设定和过度指定条件下沃罗诺伊损失的经验收敛。

初始化。在良好设定情形（$K = K_0$）下，通过对真实值 $\boldsymbol{\omega}_0$ 和 $\boldsymbol{\upsilon}_0$ 添加加性高斯噪声来初始化参数：每个坐标均设为对应真实值加上一个独立的 $\mathcal{N}(0, 1)$扰动，并按预设噪声水平缩放。在过参数化情形（$K > K_0$）下，我们在真实模型之外引入额外的专家组件。对于每个额外的簇 $k \in \{K_0 + 1, \ldots, K\}$，我们设定$\boldsymbol{\omega}_k$,0, $_0 = 0$, $\boldsymbol{\omega}_k$,0,$_1 = 8$, $\boldsymbol{\upsilon}_0$,$_k$,0,$_0 = \boldsymbol{\upsilon}_1$,$_k$,1,$_0 = 0$,$\boldsymbol{\upsilon}_0$,$_k$,1, $_0 = 20$以及$\boldsymbol{\upsilon}_1$,$_k$,0,$_0 = 10$。原有的 K_0 个专家按良好设定情形中的方式初始化。此设计确保过参数化确实产生近重复原子，这正是我们的Voronoi分析和合并策略所针对的情形。

BatchMM的收敛性。我们研究了使用所提出的批量MM算法训练的SGMLMoE的收敛行为，并将其与经典优化方法进行比较，包括随机梯度下降（SGD）、Adam、AdamW、RMSprop和期望最大化。在良好设定情形下，我们将样本量从 $N = 10^2$ 变为 $N = 10^5$。在过参数化情形下，我们固定 $K = 3$，并将N 从 $10^{2.75}$ 变为 10^4。我们通过估计参数$(\boldsymbol{\omega},\ \boldsymbol{\upsilon})$与真实参数$(\boldsymbol{\omega}_0,\ \boldsymbol{\upsilon}_0)$之间的Frobenius范数来评估收敛性。如图2所示，批量MM始终收敛到全局最小值，而其他方法通常收敛速度明显较慢；在过参数化情形下，随着 N 增大，它们还经常收敛到次优解。

Voronoi 损失的收敛。我们实证验证了由定理 5所预测的收敛速率。在精确拟合设定（图3a）中，将样本量从 10^3 变化至$10^{5.5}$，所得对数-对数斜率与参数化速率 $N^{-1/2}$一致。在过度设定情形下，当 $K \in \{3, 4\}$ 与 $N \in [10^{2.5}, 10^4]$时，图3b中拟合的回归线与预测速率相符。对于 $K = 4$，各组件的Frobenius误差表明，一个伪组件以较慢的次参数化速率收敛，经验上接近 $N^{-0.25}$，这与定理 5针对过覆盖的Voronoi单元所预测的速率一致。相比之下，另一个过拟合组件的误差随 N 增大而上升，反映出过参数化模型中轻微的排斥效应：冗余组件会偏离真实参数，以解决不可辨识性（此处相关单元为 $\mathbb{V}_k$）。

使用DSC进行合并。最后，我们评估基于DSC的合并策略在恢复正确专家数量方面的有效性。我们固定 $K = 4$，并将样本量从 $N = 10^2$ 变化到 10^4。对于每个 N，我们使用不同的随机种子重复实验20次。我们通过记录以下两项指标，将DSC与AIC、BIC和ICL进行比较：(i) 各准则选出正确专家数量的频率；(ii) 选出的专家数量的平均值。结果如 图4所示。在不同样本量下，DSC选出正确数量的频率均高于基准准则。尤其是在小样本区间，DSC明显更加稳健；而当 N 足够大时，DSC能够始终恢复真实的专家数量，与我们的理论所确立的一致性保证相符。

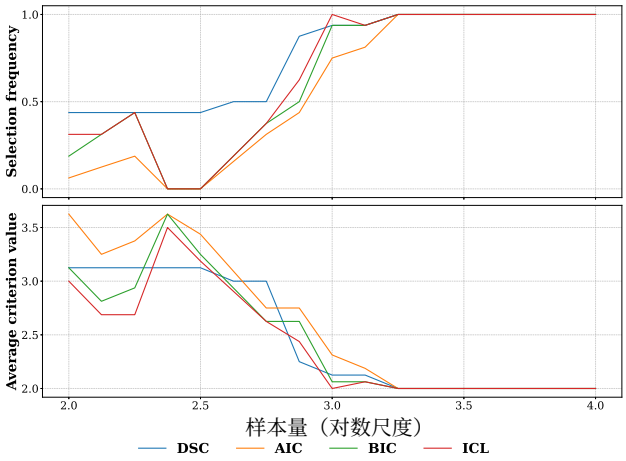


图4. DSC模型选择性能。

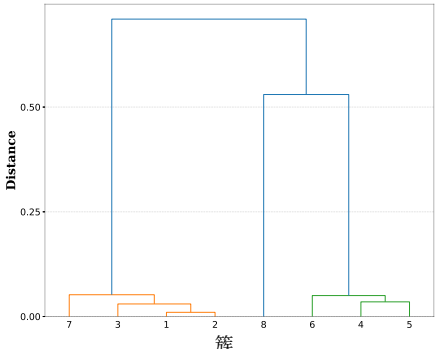


图5. 混合测度的树状图。

K	召回率	精确率	F1	# P	CoMs
1	0.54 ±0.04	0.43 ±0.03	0.48 ±0.003	6	
2	0.72 ±0.02	0.61 ±0.02	0.66 ±0.003	12	ICL
3	0.88 ±0.01	0.73 ±0.02	0.80 ±0.003	18	DSC
4	0.88 ±0.01	0.74 ±0.01	0.80 ±0.002	24	AIC, BIC
5	0.89 ±0.01	0.75 ±0.01	0.81 ±0.002	30	
6	0.91 ±0.01	0.76 ±0.01	0.83 ±0.001	36	
7	0.92 ±0.01	0.76 ±0.01	0.83 ±0.001	42	
8	0.92 ±0.01	0.76 ±0.01	0.83 ±0.001	48	

表2. 不同专家数量下的性能 K。

5.2. 真实数据集

初始化。由于真实数据中的潜在专家指派未知，我们采用一种基于无监督学习、切实可行的两阶段初始化策略。首先，我们应用经典聚类方法（例如K均值或谱聚类）获得初步分组。随后，我们根据这些指派构建初始SGMLMoE参数，为Batch MM提供结构化且由数据驱动的初始化。

专家估计。我们从蛋白质-蛋白质相互作用预测数据集中选取5,000个样本的一个子集（Tang等人，2023b；Qi等人，2007）来训练所提出的DSC方法，以及包括AIC、BIC和ICL在内的基准准则，以选择专家数量，专家数量上限设为 $K = 8$。随后，使用另一组5,000个样本进行模型训练，并针对候选 K值进行交叉验证。表2报告了交叉验证性能随 K变化的情况，以及各准则选出的专家数量。DSC选择了 $K = 3$，在保持模型紧凑的同时取得了较强的预测性能。相比之下，ICL倾向于选择较小的模型（$K = 2$），导致精确率和F1分数降低；而AIC和BIC选择了较大的模型（$K = 4$），增加了模型复杂度，却只带来微乎其微的性能提升。总体而言，这些结果表明，与经典准则相比，DSC在准确性和模型复杂度之间实现了更好的权衡。

准确率比较。我们将专家数量固定为$K = 3$，并使用批量MM算法在5,000个样本上训练SGMLMoE。我们使用召回率、精确率和F1分数，将其性能与标准基线方法——朴素贝叶斯、随机森林、支持向量机和逻辑回归进行比较。如表3所示，SGMLMoE在所有方法中整体表现最佳，取得了最高的精确率（0.74）和F1分数（0.80），同时保持了有竞争力的召回率（0.88）。朴素贝叶斯的精确率和F1分数较低，表明其判别能力有限；而随机森林、支持向量机、逻辑回归和三隐藏层多层感知机虽优于朴素贝叶斯，但在精确率和整体F1分数方面始终逊于SGMLMoE。这些结果凸显了SGMLMoE在捕捉异质性方面的优势，并从实证上验证了相较于标准训练基线，批量MM具有良好的实用稳定性。最后，图5展示了通过算法2得到的拟合混合测度树状图，直观呈现了推断出的专家结构层次。

结论与展望。总之，我们的实验证实了本文的核心观点：在全数据情形下，Batch MM 可为 SGMLMoE 实现稳定的最大似然训练，而 Voronoi损失/树状图视角则为过参数化提供了有原则的补救方案。经-

表3. 与基线方法的比较。

方法	召回率	精确率	F1
SGMLMoE	**0.88** ±0.012	**0。74** ±0.010	**0.80** ±0.002
NB	0.83 ±0.011	0.64 ±0.009	0.72 ±0.001
RF	0.85 ±0.012	0.68 ±0.008	0.76 ±0.001
SVM	0.85 ±0.010	0.69 ±0.009	0.76 ±0.002
LR	0.84 ±0.014	0.65 ±0.009	0.73 ±0.003
三隐藏层多层感知机	0.87 ±0.008	0.71 ±0.009	0.77 ±0.001

验上，Batch MM 能够可靠收敛，Voronoi 损失的衰减与预测的有限样本速率一致，并且 DSC 始终选择紧凑的专家数量，从而在蛋白质–蛋白质相互作用预测中保持准确率并改善校准。未来工作包括将 MM 替代函数扩展到增量/小批量设置，将树状图理论推广到更丰富的专家参数化与结构化正则化，以及在污染和协变量偏移条件下建立鲁棒性保证。

影响声明

本文介绍的工作旨在推动机器学习领域的发展。我们的工作可能产生许多社会影响，但我们认为没有哪项影响需要在此特别强调。

参考文献

Akaike，H. 关于统计模型识别的新见解。*IEEE* 自动控制汇刊，19(6):716–723，1974年。

Bengio，Y. 表征的深度学习：展望未来。载于统计语言与语音处理，第1–37页，柏林，海德堡，2013年。ISBN 978-3-642- 39593-2。

Berrettini, M.、Galimberti, G.、Ranciati, S. 和 Murphy, T. B. 识别英国下议院的英国脱欧投票模式：基于具有灵活伴随协变量效应的贝叶斯专家混合模型的分析。皇家统计学会会刊 *C* 辑：应用统计，73(3):621–638，2024年6月。ISSN 0035-9254。

Biernacki, C.、Celeux, G. 和 Govaert, G. 使用积分完备似然评估用于聚类的专家混合模型。*IEEE* 模式分析与机器智能汇刊，22(7):719–725，2000年。

Borkar, V. S. 随机逼近：动力系统视角，第 9 卷。Springer，2008 年。

Cai, W.、Jiang, J.、Wang, F.、Tang, J.、Kim, S. 和 Huang, J. 大语言模型中的专家混合综述。*IEEE* 知识与数据工程汇刊，37(7):3896–3915，2025年。

Capp´e, O. 隐马尔可夫模型的在线 EM 算法。计算与图形统计学杂志，20(3):728–749，2011年1月。ISSN 1061-8600。

Capp´e, O. 和 Moulines, E. 潜在数据模型的在线期望最大化算法。《皇家统计学会杂志》*B*辑：统计方法，71(3):593–613，2009年2月。ISSN 1369-7412。

Chen, J.、Zhu, J.、Teh, Y. W. 和 Zhang, T. 采用方差缩减的随机期望最大化。载于 Bengio, S.、Wallach, H.、Larochelle, H.、Grauman, K.、Cesa-Bianchi, N. 和 Garnett, R.（编），神经信息处理系统进展，第31卷。Curran Associates, Inc.，2018年。

Chen, K.、Xu, L. 和 Chi, H. 用于多类分类的专家混合改进学习算法。《神经网络》，12(9):1229–1252，1999年。ISSN 0893-6080。

Chen, Z.、Deng, Y.、Wu, Y.、Gu, Q. 和 Li, Y. 探究深度学习中的专家混合层。载于神经信息处理系统进展，2022年。

Chong, M. C.、Nguyen, H. D. 和 Nguyen, T. 通过 h 提升的 Kullback–Leibler 散度研究紧致域上混合密度估计的风险界。机器学习研究汇刊，2024年。ISSN 2835-8856。

Consortium, U. UniProt：蛋白质信息中心。《核酸研究》，43(D1)：D204–D212，2015年。

Corff, S. L. 和 Fort, G. 基于在线期望最大化的隐马尔可夫模型推断算法。《电子统计学杂志》，7(none)：763 – 792，2013年。

De Leeuw, J. 凸分析在多维尺度分析中的应用。*Recentdevelopmentsin statistics*，第133–145页，1977年。

De Leeuw, J. 和 Heiser, W. J. 多维尺度分析中校正矩阵算法的收敛性。*Geo-metricrepresentationsofrelationaldata*，36：735–752，1977年。

Dempster, A. P., Laird, N. M., 和 Rubin, D. B. 通过EM算法从不完整数据中进行最大似然估计。《皇家统计学会杂志》。*B*辑（方法论），39(1):1–38，1977年3月。ISSN 00359246。

Dieuleveut, A., Fort, G., Moulines, E., 和 Wai, H.-T. 超越梯度的随机逼近：用于信号处理和机器学习。*IEEE*信号处理汇刊，71:3117–3148，2023年。

Do, D., Do, L., McKinley, S. A., Terhorst, J., 和 Nguyen, X. 混合测度树状图：有限混合模型的潜在层级学习与模型选择。*arXiv*预印本 *arXiv:2403.01684*，2024年。

Do, D., Do, L., 和 Nguyen, X. 异质响应回归中的强可辨识性与参数学习。《电子统计学杂志》，19(1):131–203，2025年。

Forbes, F., Nguyen, H. D., Nguyen, T., 和 Arbel, J. 用于近似贝叶斯计算的专家混合后验替代模型。载于 *JDS2022 - 53`法国统计学会年´会（SFdS）统计日´，法´国،法国统计学会 (SFdS)*，法国里昂，2022年6月a。

Forbes, F.、Nguyen, H. D.、Nguyen, T. 与 Arbel, J. 基于替代后验分布的近似贝叶斯计算的汇总统计量与差异度量。*Statistics and Computing*，32(5):85，2022年10月b。ISSN 1573-1375。

Fort, G. 与 Moulines, E. 扰动近端预条件 Spider 算法：非渐近收敛界。载于*2021*年 *IEEE* 统计信号处理研讨会（*SSP*），第 96–100 页，2021年。Fort, G. 与 Moulines, E. 面向非凸复合优化的带方差缩减的随机变度量近端梯度。*Statistics and Computing*，33(3)，2023年4月。ISSN 1573-1375。

Fort, G., Moulines, E., 和 Wai, H.-T. 一种随机路径积分微分估计期望最大化算法。载于 Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., 和 Lin, H.（编），神经信息处理系统进展，第33卷，第16972–16982页。Curran Associates公司，2020年。

Fort, G., Gach, P., 和 Moulines, E. 面向有限和优化的快速增量期望最大化：非渐近收敛。*Statisticsand Computing*，31(4):48，2021年6月a。ISSN1573-1375。

Fort, G., Moulines, E. 和 Wai, H.-T. Geom-Spider-EM：用于非凸有限和优化的更快速方差缩减随机期望最大化。载于*ICASSP2021 -2021 IEEE International Conference onAcoustics, Speech and Signal Processing (ICASSP)*，第 3135–3139 页，2021b。

Foygel, R. 和 Drton, M. 高斯图模型的扩展贝叶斯信息准则。载于 Lafferty, J.、Williams, C.、Shawe-Taylor, J.、Zemel, R. 和 Culotta, A.（编），神经信息处理系统进展，第 23 卷。Curran Associates公司，2010年。

Fr¨uhwirth-Schnatter, S. 保持平衡——用于有限混合、专家混合和马尔可夫混合模型边际似然估计的桥采样。*Brazil-ian Journal of Probability and Statistics*，33(4)：706 – 733，2019年。

Fr¨uhwirth-Schnatter, S.、Pamminger, C.、Weber, A. 和 Winter-Ebmer, R. 劳动力市场进入与收入动态：使用专家混合马尔可夫链聚类进行贝叶斯推断。《应用计量经济学杂志》，27(7):1116–1137，2012年。

Fr¨uhwirth-Schnatter, S.、Pittner, S.、Weber, A. 和 Winter-Ebmer, R. 使用时变专家混合马尔可夫链聚类分析工厂关闭效应。《应用统计年鉴》，12(3):1796–1830，2018年。出版商：数理统计学会。

Genovese, C. R. 和 Wasserman, L. 高斯混合筛的收敛速率。《统计学年鉴》，28(4):1105 – 1127，2000年。

Guha, A.、Ho, N. 和 Nguyen, X. 关于贝叶斯混合建模中参数的后验收缩与可解释性。伯努利，27(4):2159 – 2188，2021年。

Hai, D. T., Mai, T. N., Nguyen, T., Ho, N., Nguyen, B. T. 和 Drovandi, C. Softmax门控高斯专家混合模型的混合测度树状图：无需模型遍历的一致性。载于第*29*届人工智能与统计国际会议，《机器学习研究论文集》。PMLR，2026年。

Hennig, C. 聚类线性回归模型的可识别性。《分类学期刊》，17(2):273–296，2000年。ISSN 1432-1343。

Ho, B. H., Chi, L. N., Nguyen, T., Hoang, V. H., Nguyen, B. T. 和 Drovandi, C. 基于模型的聚类中非随机缺失情形下变量选择的统一框架。载于第三十九届神经信息处理系统年会，2025年。

Ho, N. 和 Nguyen, X. 某些弱可辨识有限混合模型的参数估计收敛速率。《统计年鉴》，44(6):2726–2755，2016a。

Ho, N. 和 Nguyen, X. 有限混合模型的强可辨识性与参数估计收敛速率。《电子统计学杂志》，10(1):271–307，2016b。

Ho, N. 和 Nguyen, X. 分布有限混合模型中的奇异结构及其对参数估计的影响。《*SIAM*数据科学数学期刊》，1 (4):730–758，2019年1月。

Ho, N.、Yang, C.-Y. 和 Jordan, M. I. 高斯专家混合模型的收敛速率。《机器学习研究期刊》，2022年。

Huynh, B. T. 和 Chamroukhi, F. 广义线性专家混合模型中的估计与特征选择。*arXiv*预印本 *arXiv:1907.06994*，2019年。

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. 和 Hinton, G. E. 自适应局部专家混合模型。《神经计算》，3(1):79–87，1991年。出版商：麻省理工学院出版社。

Jiang, W. 和 Tanner, M. A. 关于专家混合模型的可辨识性。神经网络，12(9):1253–1258，1999a。ISSN 0893-6080。

Jiang, W. 和 Tanner, M. A. 指数族回归模型的层次专家混合模型：近似与最大似然估计。《统计年鉴》，第 987–1011 页，1999b。

Jordan, M. I. 和 Jacobs, R. A. 专家层次混合模型与期望最大化算法。《神经计算》，6(2): 181–214，1994年。出版方：麻省理工学院出版社。

Karimi, B.、Miasojedow, B.、Moulines, E. 和 Wai, H.- T. 有偏随机逼近方案的非渐近分析。收录于 Beygelzimer, A. 和 Hsu, D.（编），第三十二届学习理论会议论文集，机器学习研究论文集第 99 卷，第 1944–1974 页。PMLR，2019年6月。Karimi, B.、Wai, H.-T.、Moulines, E. 和 Lavielle, M. 论（快速）增量期望最大化方法的全局收敛。收录于神经信息处理系统进展，第 32 卷，2019b。

Keener, R. W. 理论统计学：核心课程主题。施普林格科学与商业媒体，2010年。

Khalili, A., Yang, A. Y. 和 Da, X. 稀疏专家混合模型中参数数量发散时的估计与组特征选择。统计规划与推断期刊，第 106250 页，2024年。ISSN 0378-3758。

Kuhn, E.、Matias, C. 和 Rebafka, T. 采用小批量抽样的随机逼近EM算法的性质。*Statisticsand Computing*，30(6):1725–1739，2020年11月。ISSN1573-1375。

Kushner, H. 和 Yin, G. G. 随机逼近、递归算法及其应用，第35卷。Springer Science & Business Media，2003年。

Lange, K. *MM*优化算法。工业与应用数学学会，宾夕法尼亚州费城，2016年。

Lange, K.、Won, J.-H.、Landeros, A. 和 Zhou, H. 通过MM算法进行非凸优化：收敛理论。载于*Wiley StatsRef*：在线统计学参考资料，第1–22页。约翰·威立父子有限公司，2021年。ISBN 978-1- 118-44511-2。

Mairal, J. 面向大规模优化的随机主化最小化算法。神经信息处理系统进展，第26卷，2013年。

Mairal, J. 增量式优超-最小化优化及其在大规模机器学习中的应用。《*SIAM*优化期刊》，25(2)：829–855，2015。

Masoudnia, S. 和 Ebrahimpour, R. 专家混合：文献综述。《人工智能评论》，42(2)：275–293，2014。ISSN 1573-7462。

McLachlan, G. J. 和 Krishnan, T. 《期望最大化算法及其扩展》。威利，1997。

Meilijson, I. 对EM算法本身的快速改进。英国皇家统计学会杂志：*B*辑（方法论），51(1):127–138，1989年9月。ISSN 0035-9246。

Montuelle, L. 和 LePennec, E. 带逻辑斯蒂权重的高斯混合回归模型：一种惩罚最大似然方法。《电子统计学杂志》，8 (1):1661–1695，2014年。

Neal, R. M. 和 Hinton, G. E. 论EM算法：为增量式、稀疏式及其他变体提供理论依据。载于 Jordan, M. I.（编），图模型学习，第355–368页。施普林格荷兰，多德雷赫特，1998年。ISBN 978-94-011-5014-9。

Nguyen, D. N. 和 Li, Z. 高维转录组数据异质依赖关系中高斯图模型的联合学习。载于第*16*届亚洲机器学习会议（会议论文轨道），2024年。

Nguyen, H., Nguyen, T. 和 Ho, N. 揭开高斯专家混合模型中 Softmax 门控的奥秘。载于神经信息处理系统进展，2023年12月a。Nguyen, H., Akbarian, P., Nguyen, T. 和 Ho, N. 关于 Softmax 门控多项逻辑斯蒂专家混合模型的一般理论。载于第 *41* 届国际机器学习会议论文集，2024年a。Nguyen, H., Nguyen, T., Nguyen, K. 和 Ho, N. 高斯门控专家混合模型参数估计收敛速率的研究。载于第 *27* 届人工智能与统计学会议论文集，第 238 卷，第 2683–2691 页，2024年5月b。Nguyen, H. D. 面向机器学习与统计估计的优超-最小化算法导论。*WIREs* 数据挖掘与知识发现，7(2):e1198，2017年。Nguyen, H. D. 和 Chamroukhi, F. 专家混合建模的实践与理论方面：综述。*Wiley* 跨学科评论：数据挖掘与知识发现，8(4):e1246，2018年。Nguyen, H. D. 和 Nguyen, T. 有限混合模型阶数选择中 BIC 的修正。*2506.20124*，2025年。Nguyen, H. D., Lloyd-Jones, L. R. 和 McLachlan, G. J. 混合专家模型的通用逼近定理。《神经计算》，28(12):2585–2593，2016年。出版方：麻省理工学院出版社。Nguyen, H. D., Chamroukhi, F. 和 Forbes, F. 关于多输出高斯门控线性专家混合模型的逼近结果。《神经计算》，366:208–214，2019年。ISSN 0925-2312。

Nguyen, H. D., Forbes, F. 和 McLachlan, G. J. 指数族有限混合模型的小批量学习。*Statistics and Computing*，30(4):731–748，2020年7月a。ISSN 1573-1375。

Nguyen, H. D.、Nguyen, T.、Chamroukhi, F. 和 McLachlan, G. J. 通过专家混合模型逼近Lebesgue空间中的条件概率密度函数。*Journalof StatisticalDistributions andApplications*，8(1)：13，2021年8月a。ISSN 2195-5832。

Nguyen, H. D.、Nguyen, T. 和 Forbes, F. 使用专家混合模型进行贝叶斯无似然推断。载于*2024*年国际神经网络联合会议（*IJCNN*），第 1–8 页，2024c。

Nguyen, T. 高维专家混合模型中的模型选择与近似：从理论到实践。博士论文，Normandie Universit´e，2021年12月。Nguyen, T., Nguyen, H. D., Chamroukhi, F., and McLach- lan, G. J. 通过在无穷远处趋于零的连续密度函数的有限混合进行近似。*Cogent Mathe-matics &Statistics*，7(1):1750861，2020b。Nguyen, T., Chamroukhi, F., Nguyen, H. D., and Forbes, F. 多项式专家混合模型的块对角形式中的非渐近模型选择。预印本。*arXiv:2104.08959*，2021年5月。Nguyen, T., Chamroukhi, F., Nguyen, H. D., and Forbes, F. 采用非渐近方法对专家混合模型进行惩罚式模型选择。载于 *JDS 2022 -53`emes Journ´eesde Statistique de la Soci´et´eFranc¸aisede Statistique(SFdS)*，法国里昂，2022年6月。Nguyen, T., Chamroukhi, F., Nguyen, H. D., and McLach- lan, G. J. 在Lebesgue空间中通过位置-尺度有限混合模型逼近概率密度函数。*CommunicationsinStatistics -Theory andMethods*，第1–12页，2022年5月。ISSN 0361-0926。Nguyen, T., Nguyen, H. D., Chamroukhi, F., and Forbes, F. 一种采用惩罚式方法在高维专家混合模型中进行模型选择的非渐近方法。《电子统计学杂志》，16(2):4742 – 4822，2022c。Nguyen, T., Nguyen, D. N., Nguyen, H. D., and Cham-roukhi, F. 基于联合秩选择与变量选择的高维专家混合模型选择非渐近理论。载于*2023*年澳大利亚联合人工智能会议，2023年12月。Nguyen, T., Nguyen, H. D., Chamroukhi, F., and McLach- lan, G. J. 高维专家混合模型中Lasso的非渐近预言不等式。*arXiv:2009.10622*，2023年2月。

Nguyen, T., Forbes, F., Arbel, J., and Duy Nguyen, H. 用于逆问题的贝叶斯非参数专家混合模型。《非参数统计学杂志》，第1–60页，2024年5月。ISSN 1048-5252。

Nguyen, X. 有限与无限混合模型中潜在混合测度的收敛性。《统计学年鉴》，41 (1):370–400，2013年。

Norets, A. 通过平滑回归混合模型逼近条件密度。《统计学年鉴》，38(3):1733– 1766，2010年。

Ortega, J. 和 Rheinboldt, W. 多变量非线性方程的迭代求解，第 30 卷。SIAM，1970年。

Oudoumanessah, G., Coudert, T., Lartizien, C., Dojat, M., Christen, T., 和 Forbes, F. 可扩展的磁共振指纹成像：从大规模数据中对高维椭圆混合模型进行增量推断，2024年。

Oudoumanessah, G., Coudert, T., Meyer, L., Delphin, A., Dojat, M., Lartizien, C., 和 Forbes, F. 全局聚类，局部简化：用于磁共振指纹成像的可扩展高效字典压缩。载于*IEEE*国际生物成像研讨会（*ISBI*），2025年。

Pham, N. T. 和 Chamroukhi, F. 用于分类的函数型混合专家模型。载于 *53`*法国统´计学会（*SFdS*）统计学年会´´，2022年。

Qi, Y., Klein-Seetharaman, J. 和 Bar-Joseph, Z. 一种用于蛋白质-蛋白质相互作用预测的特征专家混合方法。《*BMC*生物信息学》，8(10):S6，2007年12月。ISSN 1471-2105。

Rakhlin, A., Panchenko, D. 和 Mukherjee, S. 混合密度估计的风险界。*ESAIM*：概率与统计，9:220–229，2005年。

Schwarz, G. 模型维数的估计。《统计年刊》，6(2):461–464，1978年。

Sin, C.-Y. 和 White, H. 用于选择可能设定错误的参数模型的信息准则。《计量经济学杂志》，71(1):207–225，1996年。ISSN 0304-4076。

Sun, Y., Babu, P., 和 Palomar, D. P.《信号处理、通信与机器学习中的优超-极小化算法》。《IEEE信号处理汇刊》，65(3):794–816，2017年。

Tang, T., Zhang, X., Liu, Y., Peng, H., Zheng, B., Yin, Y., 和 Zeng, X. 机器学习在蛋白质-蛋白质相互作用预测中的应用：模型、挑战与趋势。*Briefings in Bioinformatics*, 24(2):bbad076, 2023a。

Tang, T., Zhang, X., Liu, Y., Peng, H., Zheng, B., Yin, Y., 和 Zeng, X. 机器学习在蛋白质-蛋白质相互作用预测中的应用：模型、挑战与趋势。*Briefings inBioinformatics*, 24(2):bbad076, 2023年3月b。ISSN 1477-4054。

Thai, T., Nguyen, T., Do, D., Ho, N., 和 Drovandi, C.《使用混合测度树状图对高斯门控高斯专家混合模型进行模型选择》。《arXiv预印本 arXiv:2505.13052》，2025年。

Van de Geer, S.《M估计中的经验过程》，第6卷。剑桥大学出版社，2000年。

Wang, Z., Ji, K., Zhou, Y., Liang, Y., 和 Tarokh, V.《SpiderBoost与动量：更快的方差缩减算法》。载于 Wallach, H., Larochelle, H., Beygelzimer, A., Alch´e-Buc, F. d., Fox, E., 和 Garnett, R.（编），《神经信息处理系统进展》，第32卷。Curran Associates公司，2019年。

Westerhout, J.、Nguyen, T.、Guo, X. 和 Nguyen, H. D.《最小经验风险的渐近分布》。载于第四十一届国际机器学习会议，2024年。

Yuksel, S. E. 和 Gader, P.《用于分类的变分专家混合模型及其在地雷探测中的应用》。载于*2010*年第*20*届国际模式识别会议，第2981–2984页，2010年。

Yuksel, S. E.、Wilson, J. N. 和 Gader, P. D.《专家混合模型二十年》。*IEEE*神经网络与学习系统汇刊，23(8):1177–1193，2012年。ISSN 2162-2388，卷号 23。

Zens, G. 混合专家模型中的贝叶斯收缩：识别类别归属的稳健决定因素。数据分析与分类进展，13(4):1019–1051，2019年12月。ISSN 1862-5355。

《Softmax门控多项逻辑专家混合模型的快速模型选择与稳定优化》补充材料

附录组织结构。本附录从三个互补层面汇集了支撑主论文的材料。A 节补充讨论了经验启示和具体的未来研究方向。B 节提供了 SGMLMoE 批量 MM 例程中可直接用于实现的细节，包括线性代数恒等式、完整的算法步骤，以及有关条件数和停止准则的实用说明。C 节记录了实验方案（数据预处理、评估和可复现性细节）。最后，D 节包含完整证明。

A. 扩展结论与展望

总体而言，数值实验支持了本文的核心观点：在 SGMLMoE 中，所提出的批量 MM 更新为全数据情形下的最大似然训练提供了稳定可靠的方案，而沃罗诺伊损失／树状图视角则为处理过参数化提供了切实有效的方法。实验结果显示：(i) 在正确设定和过度设定情形下，批量 MM 均能带来单调且可重复的似然改进；(ii) 在精确拟合和过拟合情形下，沃罗诺伊损失均会下降，与有限样本理论的结论一致；(iii) DSC 能够稳健地恢复有效专家数量，从而得到紧凑的模型，并在蛋白质–蛋白质相互作用数据上实现具有竞争力或更优的预测性能和概率校准效果。

展望未来，有几个自然的拓展方向。首先，值得将本文的批量 MM 稳定性保证与增量式或小批量变体结合起来，以处理规模更大的数据集，同时保留单调替代函数结构。其次，将树状图／DSC 分析拓展到更丰富的专家参数化形式（例如更深层特征映射、结构化稀疏性或依赖上下文的专家共享），有望扩大其适用范围，同时通过合并层级保持可解释性。第三，在轻微模型错设情况下观察到的稳健性表明，值得从理论上研究污染和协变量偏移情形，并将 DSC 式结构惩罚项作为纯似然标准的一种原则性替代方案。最后，开发开源的生产级实现，并采用更广泛的生物学基准数据集（不限于 PPI），将有助于进一步明确层次聚合在异质分类问题中何时以及如何权衡准确性、校准效果和内存／计算开销。

B. 批量 MM 算法详解

求逆恒等式。 在实现中，我们利用公式（5）的分块结构。具体而言，

$$\boldsymbol{B}_{n,K-1}^{-1} = \left(\frac{4}{3}\boldsymbol{I}_{K-1} + \frac{8}{3(K-1)}\mathbf{1}_{K-1}\mathbf{1}_{K-1}^{\top}\right) \otimes \left(\sum_{n=1}^{N} \hat{\boldsymbol{x}}_n \hat{\boldsymbol{x}}_n^{\top}\right)^{-1}, \tag{11}$$

$$\left[\boldsymbol{B}_{n,M-1}^{K}\right]^{-1} = \boldsymbol{I}_K \otimes \boldsymbol{B}_{n,M-1}^{-1}. \tag{12}$$

此外，对于任意元素为正的对角矩阵 $\mathrm{diag}(\boldsymbol{a})$，我们有 $\mathrm{diag}(\boldsymbol{a})^{-1} = \mathrm{diag}(1/\boldsymbol{a})$，并在需要时将其应用于$\boldsymbol{Z}_n^{(t)}$ 。

SGMLMoE 批量 MM 算法。 算法2 实现了用于拟合 SGMLMoE 的批量 MM 程序。在每次迭代中，它首先计算责任度 $\tau_{n\,k}^{(t)}$，针对所有样本和专家，使用当前门控网络与专家概率，然后构造批量类充分统计量项 $(\boldsymbol{s}_n^{(t)},\ \boldsymbol{r}_n^{(t)})$ 以及相应的曲率分块$(\boldsymbol{B}_{n,\ K},\ \boldsymbol{B}_{n,\ M})$用以定义二次 MM 替代函数，见 公式（6）。随后将这些逐样本量在批次上聚合，以形成全局更新，包括对角加权矩阵 $\boldsymbol{Z}_n^{(t)}$以及对数求和指数项的梯度g_n （门控网络）和 e_n （专家）。最后，该方法通过求解两个预条件线性方程组，对门控网络和专家参数执行闭式 MM 更新，从而得到 $(\boldsymbol{w}^{(t+1)},\ \boldsymbol{v}^{(t+1)})$。根据其构造，所得迭代单调提升对数似然，并且当似然增量低于容差 ε （或当

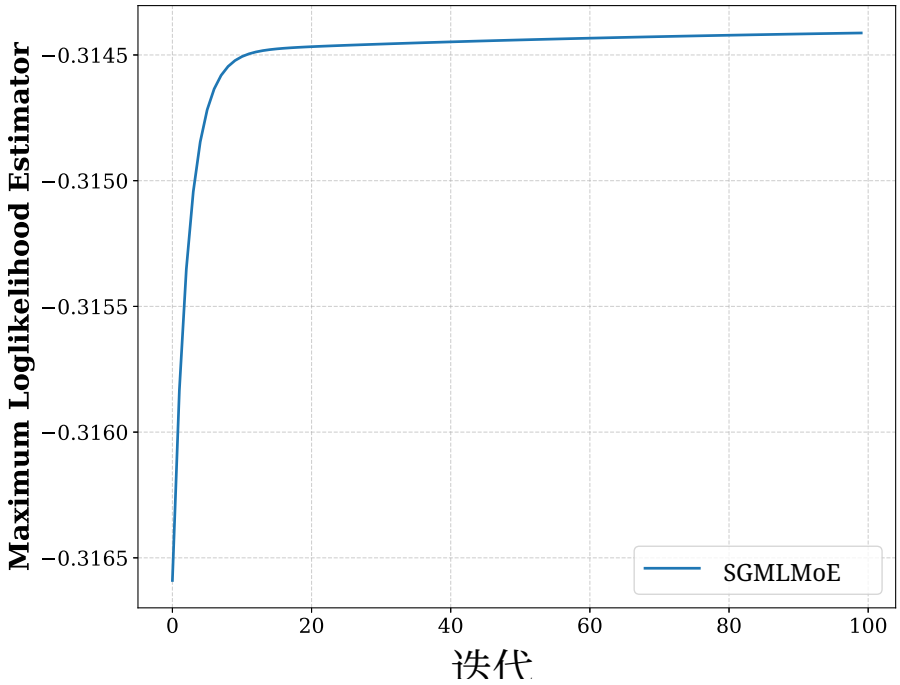


图*6.* 在 $N = 1000$下，SGMLMoE前100次优化迭代中最大对数似然的变化。

达到最大迭代次数时）。

C. 实验细节

真实数据集预处理。 我们首先使用（Tang等人，2023b；Qi等人，2007）提供的数据集来获取正相互作用标签。遵循Tang等人 （2023a），我们采用从Consortium（2015）收集的蛋白质序列数据作为输入，得到9,760条蛋白质序列。每条序列使用ProtLearn库编码为20维氨基酸组成向量。随后，我们将数据集限制为出现在DIP数据库中的蛋白质，从而得到2,321个保留标识符及相应序列编码的蛋白质。为构建带标签的相互作用样本，我们选取 n个在DIP中具有经验证的相互作用的蛋白质对，并计算其序列编码之间的逐元素绝对差，生成标记为1的20维特征向量。同样，我们采样 n 个未出现在DIP中的无相互作用蛋白质对，并以相同方式构建其特征向量，赋予标签0。正样本与负样本被拼接，形成一个 $2n \times 20$ 特征矩阵及相应的二元标签向量，所得数据集在训练前被随机打乱。

最大对数似然估计量的单调性。 经典优化方法（如随机梯度下降和Adam）在混合模型或混合专家模型中常常导致不稳定的最大似然估计行为。这种不稳定性源于高度非凸的似然函数景观，其中不恰当的步长可能导致振荡或发散，使性能对细致的学习率调优十分敏感。此类调优依赖于具体问题，且可能计算代价高昂，从而限制稳健性与可复现性。相比之下，我们提出的方法通过采用基于MM的优化方案避免了这些问题，该方案保证最大似然估计目标函数的单调改进，如定理2所确立。这种单调性确保了稳定且可预测的优化动态。我们通过在图6中绘制跨迭代次数的最大似然估计轨迹，从经验上说明了这一行为。如图所示，我们的方法呈现出平滑、非递减的似然进展，与基于梯度的优化器下观察到的波动行为形成对比。该实验在第5.2节所述的合成数据集上进行，样本数为 $N = 1000$ 个。

替代函数构造在MM算法等优化技术中起着关键作用。借助数学不等式，可以用更简单、更易处理的形式来逼近或界定复杂的目标函数。以下是构建替代函数时常用的若干基本不等式：

算法2 SGMLMoE 批量 MM 算法

要求：批量数据 $\{(\boldsymbol{x}_n, \mathsf{y}_n)\}_{n=1}^N$，专家/类别数量 (K, M)，多项式次数 D，初始化 $\boldsymbol{\theta}^{(0)} = (\boldsymbol{w}^{(0)}, \boldsymbol{v}^{(0)})$，容差 $\varepsilon > 0$，最大迭代次数 T。确保：MM 迭代序列 $\{\boldsymbol{\theta}^{(t)}\}_{t\geq 0}$。1：对于 $t = 0,\ 1, \ldots, T-1$ 执行 2：E 步（责任度）。3：对于 $n = 1, \ldots, N$ 执行 4：构造特征向量 $\hat{\boldsymbol{x}}_n = \left[x_{n,1}^0, \ldots, x_{n,1}^D, \ldots, x_{n,P}^0, \ldots, x_{n,P}^D\right]^\top$。5：计算门控概率 $\{\mathsf{g}_k(\boldsymbol{w}^{(t)}(\boldsymbol{x}_n))\}_{k=1}^K$。6：对每个 $k \in [K]$ 计算专家概率 $\{\mathsf{e}_k(\mathsf{y}_n = m; \boldsymbol{v}^{(t)}(\boldsymbol{x}_n))\}_{m=1}^M$。7：计算后验概率 8：构建 $\boldsymbol{s}_n^{(t)}$和 $\boldsymbol{r}_n^{(t)}$（利用 $\mathbf{1}(\mathsf{y}_n, l)$ 以及公式（6）及其上下文中的定义）。9：构建曲率分块$\boldsymbol{B}_{n,K}$ 和$\boldsymbol{B}_{n,M}$（如公式（5）所示）。10：结束循环 11：汇总批量统计量。12：$\boldsymbol{s}^{(t)} = \sum_{n=1}^N \boldsymbol{s}_n^{(t)}$，$\boldsymbol{r}^{(t)} = \sum_{n=1}^N \boldsymbol{r}_n^{(t)}$。13：$\boldsymbol{B}_{n,K-1} = \sum_{n=1}^N \boldsymbol{B}_{n,K}$，$\boldsymbol{B}_{n,M-1} = \sum_{n=1}^N \boldsymbol{B}_{n,M}$，$\boldsymbol{B}_{n,M-1}^K = \boldsymbol{I}_K \otimes \boldsymbol{B}_{n,M-1}$。14：对于 $n = 1, \ldots, N$ 执行 15：设置 $\boldsymbol{\tau}_{(t)n,:} = [\tau_{(t)\,n}, 1, \ldots, \tau_{(t)\,n,K}]^\top$ 和 $\boldsymbol{Z}_n^{(t)} = \mathrm{diag}\left(\boldsymbol{\tau}_{(t)n,:} \otimes \mathbf{1}_{P(D+1)}\right)$。16：结束循环 17：计算对数求和指数项的梯度。18：$\nabla g(\boldsymbol{w}^{(t)}) = \sum_{n=1}^N \nabla g_n(\boldsymbol{w}^{(t)})$，其中 $g_n(\boldsymbol{w}) = \log\left(1 + \sum_{k=1}^{K-1} \exp(w_k(\boldsymbol{x}_n))\right)$。19：对每个 $n \in [N]$，构造 $\nabla e_n(\boldsymbol{v}^{(t)}) = \mathrm{vec}\left([\nabla e_n(\boldsymbol{c}_1^{(t)}) \ldots, \nabla e_n(\boldsymbol{c}_K^{(t)})]\right)$，其中 $\dfrac{\mathsf{g}_k(\boldsymbol{w}^{(t)}(\boldsymbol{x}_n))\,\mathsf{e}_k(\mathsf{y}_n; \boldsymbol{v}^{(t)}(\boldsymbol{x}_n))}{\sum_{\ell=1}^K \mathsf{g}_\ell(\boldsymbol{w}^{(t)}(\boldsymbol{x}_n))\,\mathsf{e}_\ell(\mathsf{y}_n; \boldsymbol{v}^{(t)}(\boldsymbol{x}_n))}$, $k \in [K]$.

$= \log\left(1 + \sum_{m=1}^{M-1} \exp(v_{m,k}(\boldsymbol{x}_n))\right)$.20:

M 步（闭式 MM 更新）。

21: 更新门控参数:

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} + \boldsymbol{B}_{n,K-1}^{-1}\left(\boldsymbol{s}^{(t)} - \nabla g(\boldsymbol{w}^{(t)})\right).$$

22: 更新专家参数:

$$\boldsymbol{v}^{(t+1)} = \boldsymbol{v}^{(t)} + \left(\sum_{n=1}^N \boldsymbol{Z}_n^{(t)}\, \boldsymbol{B}_{n,M-1}^K\right)^{-1}\left(\boldsymbol{r}^{(t)} - \sum_{n=1}^N \boldsymbol{Z}_n^{(t)}\, \nabla e_n(\boldsymbol{v}^{(t)})\right).$$

23: 设定 $\boldsymbol{\theta}^{(t+1)} = (\boldsymbol{w}^{(t+1)},\ \boldsymbol{v}^{(t+1)})$。24: **停止准则**。如果 $|\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(t)})| \leq \varepsilon$，则 **跳出**。25: **结束循环**

D. 主要结果的证明

D.1. 技术贡献

遵循 MM 范式，我们首先在 定理1 中为 SGMLMoE 提出一种新颖的代理目标函数，并在 定理2中严格证明其满足 MM 算法所要求的单调性。基于此，我们转而研究模型合并，并对 Nguyen等人 (2024年a) 中定理 3.1 的证明作了大幅修改，以适配我们提出的 Voronoi 度量和 SGMLMoE 框架，从而得到 定理 3中的关键不等式。利用这一结果，我们进而沿用 Do等人 (2024年) 的引理 2 论证（另见 Thai等人 (2025年)；Hai等人 (2026年)），在 定理 4中为我们的模型推导出类似的单调链性质。最后，我们将 Do等人(2024年) 针对混合模型的定理 1–3 的证明改编到我们的 Voronoi 损失和 SGMLMoE 设定中，从而在 定理 5、定理 6和定理 7中分别给出关于收敛速率、模型选择一致性和合并收敛性的理论保证。

D.2. 定理1的证明

我们通过结合以下两点推导公式（6）中的代理界：(i) 基于责任度的变分上界（E 步）；(ii) 对基线 log-sum-exp 项进行一致二次主化（MM 步）。我们首先回顾本证明中使用的记号模块。

批量充分统计量记号。对于每个 $n \in [N]$，定义 $\mathbb{R}^{(K-1)P(D+1)}$ 中的门控设计向量：

$$\boldsymbol{s}_n^{(t)} = \left[\tau_{n,1}^{(t)} x_{n,1}^0, \ldots, \tau_{n,1}^{(t)} x_{n,1}^D, \ldots, \tau_{n,K-1}^{(t)} x_{n,P}^D\right]^\top. \tag{13}$$

将,,,,(y,) () .,,,,,,依次堆叠，定义专家设计向量

$$\boldsymbol{r}_n^{(t)} \qquad r_{n\ d\ l\ p\ k}^{(t)} = \mathbb{1}_{\ \ n\ \ l}\ \tau_{n\ k}^{\ t}\ x_{n\ p}^{d}\quad d \in \{0,\ldots,D\}\quad l \in [M]\quad p \in [P]\quad k \in [K]$$

$$\boldsymbol{r}_n^{(t)} = \left[r_{n,0,1,1,1}^{(t)}, \ldots, r_{n,D,1,1,1}^{(t)}, \ldots, r_{n,D,M,1,1}^{(t)}, \ldots, r_{n,D,M,P,K}^{(t)}\right]^\top.$$

我们还将每个专家 k 的类别参数块定义为

$$\boldsymbol{c}_k = \left[\boldsymbol{v}_{1,k,0,1}, \ldots, \boldsymbol{v}_{1,k,D,1}, \ldots, \boldsymbol{v}_{M,k,D,1}, \ldots, \boldsymbol{v}_{M,k,D,P}\right]^\top, \qquad \boldsymbol{v} = \mathrm{vec}([\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K]).$$

最后，我们将批量统计量汇总为

$$\boldsymbol{s}^{(t)} = \sum_{n=1}^{N} \boldsymbol{s}_n^{(t)}, \qquad \boldsymbol{r}^{(t)} = \sum_{n=1}^{N} \boldsymbol{r}_n^{(t)}.$$

门控与专家的二次主化函数。对于 $k \in [K-1]$，令 $w_k(\boldsymbol{x}) = \sum_{d=0}^{D} \boldsymbol{\omega}_{k,d}^\top \boldsymbol{x}^d$，并定义门控 log-sum-exp 项

$$g_n(\boldsymbol{w}) = \log\left(1 + \sum_{k=1}^{K-1} \exp(w_k(\boldsymbol{x}_n))\right).$$

对于专家，对每个 $k \in [K]$ 定义多项式 log-sum-exp 项

$$e_n(\boldsymbol{c}_k) = \log\left(1 + \sum_{m=1}^{M-1} \exp(v_{m,k}(\boldsymbol{x}_n))\right).$$

对于任意 $K \in \mathbb{N}$，我们使用基于以下矩阵的二次主化函数

$$\boldsymbol{B}_{n,K} = \left(\frac{3}{4}\boldsymbol{I}_{K-1} - \frac{\mathbf{1}_{K-1}\mathbf{1}_{K-1}^\top}{2(K-1)}\right) \otimes \hat{\boldsymbol{x}}_n \hat{\boldsymbol{x}}_n^\top, \tag{14}$$

类似地，$\boldsymbol{B}_{n,\ M}$ 通过在公式 (14)中用 M 替换 K 来定义。为简洁起见，引入中心化增量$\bar{\boldsymbol{w}}_t = \boldsymbol{w} - \boldsymbol{w}^{(t)}$ 和 $\bar{\boldsymbol{c}}_{k,\ t} = \boldsymbol{c}_k - \boldsymbol{c}_k^{(t)}$。（这里 $\hat{\boldsymbol{x}}_n = [x_{n,1}^0, \ldots, x_{n,1}^D, \ldots, x_{n,P}^0, \ldots, x_{n,P}^D]^\top$。）

步骤1：利用责任度构造变分上界。对于每个 $n \in [N]$，定义

$$a_{n,k}(\boldsymbol{\theta}) := \mathsf{g}_k(\boldsymbol{w}(\boldsymbol{x}_n))\, \mathsf{e}_k(\mathrm{y}_n; \boldsymbol{v}(\boldsymbol{x}_n)), \qquad s_{\boldsymbol{\theta}}(\mathrm{y}_n \mid \boldsymbol{x}_n) = \sum_{k=1}^{K} a_{n,k}(\boldsymbol{\theta}).$$

根据对数和不等式（等价地，詹森不等式 / KL散度非负性），对于任意 $\boldsymbol{\tau}_{n,:} \in \Delta_K$，

$$-\log s_{\boldsymbol{\theta}}(\mathrm{y}_n \mid \boldsymbol{x}_n) \le -\sum_{k=1}^{K} \tau_{n,k} \log a_{n,k}(\boldsymbol{\theta}) + \sum_{k=1}^{K} \tau_{n,k} \log \tau_{n,k}. \tag{15}$$

选择 $\tau_{n,k} = \tau_n^{(t)}{}_{,k} := a_{n,k}(\boldsymbol{\theta}^{(t)}) / \sum_\ell a_{n,\ell}(\boldsymbol{\theta}^{(t)})$ 使得 公式（15）在 $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$处取等号。

展开 $\log a_{n,k}(\boldsymbol{\theta}) = \log \mathsf{g}_k(\boldsymbol{w}(\boldsymbol{x}_n)) + \log \mathsf{e}_k(\mathrm{y}_n; \boldsymbol{v}(\boldsymbol{x}_n))$，可得到两个凸的对数求和指数项以及关于得分的线性项：

$$-\sum_{k=1}^{K} \tau_{n,k}^{(t)} \log \mathsf{g}_k(\boldsymbol{w}(\boldsymbol{x}_n)) = g_n(\boldsymbol{w}) - \sum_{k=1}^{K-1} \tau_{n,k}^{(t)}\, w_k(\boldsymbol{x}_n), \tag{16}$$

$$-\sum_{k=1}^{K} \tau_{n,k}^{(t)} \log \mathsf{e}_k(\mathrm{y}_n; \boldsymbol{v}(\boldsymbol{x}_n)) = \sum_{k=1}^{K} \tau_{n,k}^{(t)}\, e_n(\boldsymbol{c}_k) - \sum_{k=1}^{K} \tau_{n,k}^{(t)} \sum_{l=1}^{M} \mathbb{1}(\mathrm{y}_n, l)\, v_{l,k}(\boldsymbol{x}_n). \tag{17}$$

公式（16）和（17）中的线性项与上述堆叠约定下的充分统计量内积 $\boldsymbol{w}^\top \boldsymbol{s}_n^{(t)}$和 $\boldsymbol{v}^\top \boldsymbol{r}_n^{(t)}$相匹配（因此 $w_k(\boldsymbol{x}_n) = \boldsymbol{\omega}_k^\top \hat{\boldsymbol{x}}_n$ 和 $v_{l,k}(\boldsymbol{x}_n) = \boldsymbol{v}_l^\top{}_{,k} \hat{\boldsymbol{x}}_n$）。
汇总 公式（15）至（17）并对 n求和可得

$$-\log s_{\boldsymbol{\theta}}(\mathrm{y} \mid \mathbf{x}) \le C^{(t)} - \sum_{n=1}^{N} \boldsymbol{w}^\top \boldsymbol{s}_n^{(t)} - \sum_{n=1}^{N} \boldsymbol{v}^\top \boldsymbol{r}_n^{(t)} + \sum_{n=1}^{N} g_n(\boldsymbol{w}) + \sum_{n=1}^{N} \sum_{k=1}^{K} \tau_{n,k}^{(t)}\, e_n(\boldsymbol{c}_k), \tag{18}$$

其中 $C^{(t)} := \sum_{n=1}^{N} \sum_{k=1}^{K} \tau^{(t)}{}_{n,k} \log \tau_n^{(t)}{}_{,k}$不依赖于 $\boldsymbol{\theta}$。

步骤 2：对 g_n 和 e_n进行二次主化。 g_n 和 e_n 均为由仿射得分映射复合而成的基线对数和指数函数。经过标准计算可得

$$\nabla^2 g_n(\boldsymbol{w}) = \left(\operatorname{diag}(\hat{\boldsymbol{p}}_n) - \hat{\boldsymbol{p}}_n \hat{\boldsymbol{p}}_n^\top\right) \otimes \hat{\boldsymbol{x}}_n \hat{\boldsymbol{x}}_n^\top,$$

其中 $\hat{\boldsymbol{p}}_n$ 是关于 $k \in [K-1]$的基线 softmax 概率。

将 命题1 应用于 $q = K-1$（门控网络）和 $q = M-1$（专家），再通过 引理1沿线性得分映射传递所得的海森矩阵上界，便可得到 公式（5）中的二次主化函数。事实上，根据 引理1（其中 $q = K-1$）和 命题1，

$$\nabla^2 g_n(\boldsymbol{w}) \preceq \boldsymbol{B}_{n,K},$$

其中 $\boldsymbol{B}_{n,K}$ 定义见 公式（14）。因此，沿着 $\boldsymbol{w}^{(t)}$ 与 $\boldsymbol{w}$之间线段的二阶泰勒公式给出全局二次上界

$$g_n(\boldsymbol{w}) \le g_n(\boldsymbol{w}^{(t)}) + \bar{\boldsymbol{w}}_t^\top \nabla g_n(\boldsymbol{w}^{(t)}) + \frac{1}{2} \bar{\boldsymbol{w}}_t^\top \boldsymbol{B}_{n,K} \bar{\boldsymbol{w}}_t, \qquad \forall \boldsymbol{w}. \tag{19}$$

同样的论证也适用于 $e_n(\boldsymbol{c}_k)$（其中 $q = M-1$），对每个 $k \in [K]$，可得：

$$e_n(\boldsymbol{c}_k) \le e_n(\boldsymbol{c}_k^{(t)}) + \bar{\boldsymbol{c}}_{k,t}^\top \nabla e_n(\boldsymbol{c}_k^{(t)}) + \frac{1}{2} \bar{\boldsymbol{c}}_{k,t}^\top \boldsymbol{B}_{n,M} \bar{\boldsymbol{c}}_{k,t}, \qquad \forall \boldsymbol{c}_k. \tag{20}$$

步骤 3：代理界。 将 公式（19） 和 （20） 代入 公式（18），并将所有不依赖于 $\boldsymbol{\theta}$的项吸收进 $C^{(t)}$中。这恰好得到 公式（6）中所述的替代函数不等式，即

$$-\log s_{\boldsymbol{\theta}}(\mathrm{y} \mid \mathbf{x}) \leq -\log\big[g(\boldsymbol{\theta}, \boldsymbol{x}_{[N]}, \mathrm{y}_{[N]}; \boldsymbol{\theta}^{(t)})\big],$$

具有公式（6）中的显式二次型。这就完成了证明。

Lem引理1（克罗内克积保持Loewner序）。若 $\boldsymbol{A} \preceq \boldsymbol{B}$ 和 $\boldsymbol{C} \succeq \mathbf{0}$ 均为对称矩阵，则 n

$$\boldsymbol{A} \otimes \boldsymbol{C} \preceq \boldsymbol{B} \otimes \boldsymbol{C}.$$

证明引理*1*。由于 $\boldsymbol{A} \preceq \boldsymbol{B}$，我们有 $\boldsymbol{B} - \boldsymbol{A} \succeq \mathbf{0}$。对于任意向量 $\boldsymbol{x}$,

$$\boldsymbol{x}^{\top}\big((\boldsymbol{B} - \boldsymbol{A}) \otimes \boldsymbol{C}\big)\boldsymbol{x} = \big\|\big((\boldsymbol{B} - \boldsymbol{A})^{1/2} \otimes \boldsymbol{C}^{1/2}\big)\boldsymbol{x}\big\|^2 \geq 0,$$

因此 $(\boldsymbol{B} - \boldsymbol{A}) \otimes \boldsymbol{C} \succeq \mathbf{0}$，这等价于 $\boldsymbol{A} \otimes \boldsymbol{C} \preceq \boldsymbol{B} \otimes \boldsymbol{C}$。 □

命题1（基线 softmax Hessian 的一致界）。设 $q \in \mathbb{N}$ 且令$\tilde{\boldsymbol{p}} \in (0,1)^{q+1}$ 满足 $\sum_{j=1}^{q+1} \tilde{p}_j = 1$。记$\hat{}$ $\boldsymbol{p} = [\tilde{p}_1, \ldots, \tilde{p}_q]^{\top}$ 和 $\boldsymbol{\Lambda} = \mathrm{diag}(\hat{}\boldsymbol{p})$。则

$$\boldsymbol{\Lambda} - \hat{\boldsymbol{p}}\hat{\boldsymbol{p}}^{\top} \preceq \boldsymbol{A}_q := \frac{3}{4}\boldsymbol{I}_q - \frac{\mathbf{1}_q\mathbf{1}_q^{\top}}{2q}.$$

证明命题*1*。在此证明中，我们使用以下定义： $\mathbf{1}_{q+1} := (1, \ldots, 1)^{\top} \in \mathbb{R}^{q+1}$，以及 $\mathbf{1}_{q+1}^{\perp} := \big\{\boldsymbol{u} \in \mathbb{R}^{q+1} : \mathbf{1}_{q+1}^{\top}\boldsymbol{u} = 0\big\}$.

步骤1：完整单纯形协方差界。定义$(q+1) \times (q+1)$ 协方差型矩阵

$$\boldsymbol{W}(\tilde{\boldsymbol{p}}) = \mathrm{diag}(\tilde{\boldsymbol{p}}) - \tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}^{\top}.$$

令$\mathbf{z} \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{q+1}\}$ 是一个独热随机向量，且 $\mathbb{E}[\mathbf{z}] = \tilde{\boldsymbol{p}}$。那么 $\boldsymbol{W}(\tilde{\boldsymbol{p}}) = \mathrm{Cov}(\mathbf{z})$，对于任意 $\boldsymbol{u}$ ，

$$\boldsymbol{u}^{\top}\boldsymbol{W}(\tilde{\boldsymbol{p}})\boldsymbol{u} = \mathrm{Var}(\boldsymbol{u}^{\top}\mathbf{z}).$$

由于 $\boldsymbol{u}^{\top}\mathbf{z}$ 取值于 $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{q+1}\}$，其极差至多为

$$\max_{i,j} |\boldsymbol{u}_i - \boldsymbol{u}_j| = \max_{i,j} |\boldsymbol{u}^{\top}(\boldsymbol{e}_i - \boldsymbol{e}_j)| \leq \|\boldsymbol{u}\| \max_{i,j} \|\boldsymbol{e}_i - \boldsymbol{e}_j\| = \sqrt{2}\,\|\boldsymbol{u}\|.$$

根据Popoviciu不等式，$\mathrm{Var}(\boldsymbol{u}^{\top}\mathbf{z}) \leq \frac{1}{4}$(极差)$^2 \leq \frac{1}{2}\|\boldsymbol{u}\|^2$。因此 $\boldsymbol{W}(\tilde{\boldsymbol{p}}) \preceq \frac{1}{2}\boldsymbol{I}_{q+1}$。此外， $\boldsymbol{W}(\tilde{\boldsymbol{p}})\mathbf{1}_{q+1} = \mathbf{0}$，所以 $\boldsymbol{W}(\tilde{\boldsymbol{p}})$的取值范围是 $\mathbf{1}_{q+1}^{\perp}$，因此

$$\boldsymbol{W}(\tilde{\boldsymbol{p}}) \preceq \frac{1}{2}\Big(\boldsymbol{I}_{q+1} - \frac{\mathbf{1}_{q+1}\mathbf{1}_{q+1}^{\top}}{q+1}\Big).$$

步骤 2：取主子矩阵。 $\boldsymbol{W}(\tilde{\boldsymbol{p}})$ 的左上角 $q \times q$主子矩阵恰好是 $\boldsymbol{\Lambda} - \hat{}\boldsymbol{p}\hat{}\boldsymbol{p}^{\top}$。对右端取相应的主子矩阵，得到

$$\boldsymbol{\Lambda} - \hat{\boldsymbol{p}}\hat{\boldsymbol{p}}^{\top} \preceq \frac{1}{2}\Big(\boldsymbol{I}_q - \frac{\mathbf{1}_q\mathbf{1}_q^{\top}}{q+1}\Big).$$

步骤 3：放宽至 $\boldsymbol{A}_q$ 用于 公式（5）。 尚需证明 $\frac{1}{2}\big(\boldsymbol{I}_q - \frac{\mathbf{1}\mathbf{1}^{\top}}{q+1}\big) \preceq \frac{3}{4}\boldsymbol{I}_q - \frac{\mathbf{1}\mathbf{1}^{\top}}{2q}$。它们的差等于 s

$$\Big(\frac{3}{4} - \frac{1}{2}\Big)\boldsymbol{I}_q - \Big(\frac{1}{2q} - \frac{1}{2(q+1)}\Big)\mathbf{1}_q\mathbf{1}_q^{\top} = \frac{1}{4}\boldsymbol{I}_q - \frac{1}{2q(q+1)}\mathbf{1}_q\mathbf{1}_q^{\top},$$

其特征值为1_4 在 $\mathbf{1}_q^{\perp}$和 $\frac{1}{4} - \frac{q}{2q(q+1)} = \frac{q-1}{4(q+1)} \geq 0$ 沿 $\mathbf{1}_q$方向。因此，该差矩阵是半正定的，从而证明了这一结论。 □

D.3. 定理2的证明：验证算法1的极小化-极大化性质

在整个证明过程中，我们使用第 D.2 节中引入的记号和构件，尤其是批量充分统计量 $\boldsymbol{s}_n^{(t)}$， $\boldsymbol{r}_n^{(t)}$见公式（4）以及二次曲率矩阵 $\boldsymbol{B}_{n,K}$、 $\boldsymbol{B}_{n,M}$ 见公式（5）。令 $\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N}$ 对数 $s_{\boldsymbol{\theta}}(\mathrm{y}_n \mid \boldsymbol{x}_n)$ 表示观测数据对数似然。

步骤1：极小化-极大化上界与相切（根据定理1）。根据公式（6）的右侧定义 $\mathcal{S}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ 关于 $-\mathcal{L}(\boldsymbol{\theta})$ 的批次替代函数（将所有与 $\boldsymbol{\theta}$ 无关的项并入常数）。根据定理1，对于每个迭代点 $\boldsymbol{\theta}^{(t)}$，该替代函数都满足极小化-极大化条件

$$-\mathcal{L}(\boldsymbol{\theta}) \le \mathcal{S}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) \ \forall \boldsymbol{\theta}, \qquad -\mathcal{L}(\boldsymbol{\theta}^{(t)}) = \mathcal{S}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)}). \tag{21}$$

步骤2：通过算法1精确最小化替代函数。固定 $\boldsymbol{\theta}^{(t)}$，并使用中心化增量 $\bar{\boldsymbol{w}}_t = \boldsymbol{w} - \boldsymbol{w}^{(t)}$和$\bar{\boldsymbol{c}}_{k,\ t} = \boldsymbol{c}_k - \boldsymbol{c}_k^{(t)}$ 如D.2节所述。由于该替代函数由以下两部分之和构成：（i）关于 $\boldsymbol{w}$ 的门控二次项；（ii） K 个相互独立的专家二次项，分别关于 $\{\boldsymbol{c}_k\}_{k=1}^{K}$，因此它可以分解为

$$\mathcal{S}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = C^{(t)} + \mathcal{S}_g(\boldsymbol{w}; \boldsymbol{\theta}^{(t)}) + \sum_{k=1}^{K} \mathcal{S}_{e,k}(\boldsymbol{c}_k; \boldsymbol{\theta}^{(t)}),$$

其中， $C^{(t)}$ 不依赖于 $\boldsymbol{\theta}$，且（忽略加法常数）

$$\mathcal{S}_g(\boldsymbol{w}; \boldsymbol{\theta}^{(t)}) = -\boldsymbol{w}^\top \boldsymbol{s}^{(t)} + \sum_{n=1}^{N} \Big\{ \nabla g_n(\boldsymbol{w}^{(t)})^\top \bar{\boldsymbol{w}}_t + \frac{1}{2} \bar{\boldsymbol{w}}_t^\top \boldsymbol{B}_{n,K} \bar{\boldsymbol{w}}_t \Big\}, \tag{22}$$

$$\mathcal{S}_{e,k}(\boldsymbol{c}_k; \boldsymbol{\theta}^{(t)}) = -\boldsymbol{c}_k^\top \boldsymbol{r}_k^{(t)} + \sum_{n=1}^{N} \tau_{n,k}^{(t)} \Big\{ \nabla e_n(\boldsymbol{c}_k^{(t)})^\top \bar{\boldsymbol{c}}_{k,t} + \frac{1}{2} \bar{\boldsymbol{c}}_{k,t}^\top \boldsymbol{B}_{n,M} \bar{\boldsymbol{c}}_{k,t} \Big\}. \tag{23}$$

这里， $\boldsymbol{s}^{(t)} = \sum_{n=1}^{N} \boldsymbol{s}_n^{(t)}$、 $\boldsymbol{r}^{(t)} = \sum_{n=1}^{N} \boldsymbol{r}_n^{(t)}$和 $\boldsymbol{r}_k^{(t)}$ 是与专家 k 对应的 $\boldsymbol{r}^{(t)}$ 的块（根据定义 $\boldsymbol{\upsilon} = \mathrm{vec}([\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K])$ 的堆叠约定）。

引入聚合曲率和梯度

$$\boldsymbol{B}_{K-1} := \sum_{n=1}^{N} \boldsymbol{B}_{n,K}, \qquad \boldsymbol{B}_{M-1,k}^{(t)} := \sum_{n=1}^{N} \tau_{n,k}^{(t)} \boldsymbol{B}_{n,M},$$

$$\nabla g(\boldsymbol{w}^{(t)}) := \sum_{n=1}^{N} \nabla g_n(\boldsymbol{w}^{(t)}), \qquad \nabla e_k^{(t)} := \sum_{n=1}^{N} \tau_{n,k}^{(t)} \nabla e_n(\boldsymbol{c}_k^{(t)}).$$

然后E 式（22） 是关于 $\boldsymbol{w}$ 的（严格）凸二次函数，当 $\boldsymbol{B}_{K-1} \succ \mathbf{0}$时，对其求导可得 ds

$$\nabla_{\boldsymbol{w}} \mathcal{S}_g(\boldsymbol{w}; \boldsymbol{\theta}^{(t)}) = -\boldsymbol{s}^{(t)} + \nabla g(\boldsymbol{w}^{(t)}) + \boldsymbol{B}_{K-1}(\boldsymbol{w} - \boldsymbol{w}^{(t)}).$$

将其设为$\mathbf{0}$ 可得唯一极小值点

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} + \boldsymbol{B}_{K-1}^{-1}\big(\boldsymbol{s}^{(t)} - \nabla g(\boldsymbol{w}^{(t)})\big), \tag{24}$$

这与 算法1 中的门控更新一致（聚合记号相同）。

同样，对于每个 $k \in [K]$，公式（23）关于 $\boldsymbol{c}_k$ 是一个凸二次函数，并且

$$\nabla_{\boldsymbol{c}_k} \mathcal{S}_{e,k}(\boldsymbol{c}_k; \boldsymbol{\theta}^{(t)}) = -\boldsymbol{r}_k^{(t)} + \nabla e_k^{(t)} + \boldsymbol{B}_{M-1,k}^{(t)}(\boldsymbol{c}_k - \boldsymbol{c}_k^{(t)}).$$

令其等于$\mathbf{0}$ 可得

$$\boldsymbol{c}_k^{(t+1)} = \boldsymbol{c}_k^{(t)} + \big(\boldsymbol{B}_{M-1,k}^{(t)}\big)^{-1}\big(\boldsymbol{r}_k^{(t)} - \nabla e_k^{(t)}\big), \qquad k \in [K], \tag{25}$$

等价于算法1 中的专家更新（该算法以向量化形式表示）。由于 $\mathcal{S}$ 可分解为门控块和专家块，因此拼接后的 $\boldsymbol{\theta}^{(t+1)} = (\boldsymbol{w}^{(t+1)}, \boldsymbol{\upsilon}^{(t+1)})$ 满足

$$\boldsymbol{\theta}^{(t+1)} \in \arg\min_{\boldsymbol{\theta}} \ \mathcal{S}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}). \tag{26}$$

如果某个曲率矩阵是奇异的，则使用 Moore–Penrose伪逆，或添加一个较小的岭正则化项，仍可满足相同的一阶条件 $\lambda \boldsymbol{I}$；无论采用哪种方式，公式（21）中的 MM有效性都不受影响。

步骤 3：单调性（MM上升）。通过精确最小化公式（26），

$$\mathcal{S}(\boldsymbol{\theta}^{(t+1)},\boldsymbol{\theta}^{(t)}) \le \mathcal{S}(\boldsymbol{\theta}^{(t)},\boldsymbol{\theta}^{(t)}).$$

根据公式（21）中的相切关系，$\mathcal{S}(\boldsymbol{\theta}^{(t)},\ \boldsymbol{\theta}^{(t)}) = -\mathcal{L}(\boldsymbol{\theta}^{(t)})$。根据在公式（21）中取 $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t+1)}$，$-\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) \le \mathcal{S}(\boldsymbol{\theta}^{(t+1)},\ \boldsymbol{\theta}^{(t)})$ 时的优超关系，结合这三个显示式可得

$$-\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) \le -\mathcal{L}(\boldsymbol{\theta}^{(t)}) \qquad \Longleftrightarrow \qquad \mathcal{L}(\boldsymbol{\theta}^{(t+1)}) \ge \mathcal{L}(\boldsymbol{\theta}^{(t)}),$$

这就证明了定理2。

D.4. 定理 3的证明

该结果将Nguyen等人（2024a）所发展的逆不等式技术扩展到了具有多项式得分函数的 SGMLMoE 参数化。全文中，我们使用$\hat{\boldsymbol{x}} = \phi_D(\boldsymbol{x}) \in \widehat{\mathbb{X}} \subset \mathbb{R}^{PD}$ 以及在第4.2节–第4.3节中引入的可识别混合测度表示。

回顾一下，对于 $G = \sum_{l=1}^{K} \pi_l \delta_{\eta_l} \in \mathcal{O}_K(\mathbb{T})$ 和 $G_0 = \sum_{k=1}^{K_0} \pi_k^0 \delta_{\eta_k^0} \in \mathcal{O}_{K_0}(\mathbb{T})$，记作 $s_G(y \mid \boldsymbol{x}) = \sum_{l=1}^{K} \mathsf{g}_l(\boldsymbol{x};\ G)\, \mathsf{e}_l(y \mid \boldsymbol{x};\ G)$，并使用协变量平均总变差差异 $\mathcal{D}_{\mathrm{TV}}(s_G,\ s_{G_0}) = \mathbb{E}_{\mathbf{x}}\big[\mathcal{D}_{\mathrm{TV}}(s_G(\cdot \mid \mathbf{x})\ s_{G_0}(\cdot \mid \mathbf{x}))\big]$。

我们分别证明局部部分和整体部分。

局部结构。 我们证明以下局部逆不等式：

$$\liminf_{\varepsilon \to 0} \inf_{\substack{G \in \mathcal{O}_K(\mathbb{T}):\\ \mathcal{D}_{\mathrm{V}}(G,G_0) \le \varepsilon}} \frac{\mathcal{D}_{\mathrm{TV}}(s_G, s_{G_0})}{\mathcal{D}_{\mathrm{V}}(G, G_0)} > 0. \tag{27}$$

反设 t hat 公式（27） 不成立。那么存在一个序列 $G^{(r)} = \sum_{l=1}^{K} \pi_l^{(r)} \delta_{\eta_l^{(r)}} \in \mathcal{O}_K(\mathbb{T})$ 使得 t

$$\mathcal{D}_{\mathrm{V}}(G^{(r)}, G_0) \to 0 \qquad \text{and} \qquad \frac{\mathcal{D}_{\mathrm{TV}}(s_{G^{(r)}}, s_{G_0})}{\mathcal{D}_{\mathrm{V}}(G^{(r)}, G_0)} \to 0 \qquad (r \to \infty). \tag{28}$$

当 r 足够大时，Voronoi划分 $\{\mathbb{V}_k\}_{k=1}^{K_0}$ 如第4.3节所述定义良好。对于 $l \in \mathbb{V}_k$，将局部差$\Delta\hat{\boldsymbol{\omega}}_{lk}^{(r)} = \hat{\boldsymbol{\omega}}_l^{(r)} - \hat{\boldsymbol{\omega}}_k^0$, $\Delta\bar{v}_{lkm}^{(r)} = \bar{v}_m^{(r)},_l - \bar{v}_{m,k}^0$, $\Delta\hat{\boldsymbol{v}}_{lkm}^{(r)} = \hat{\boldsymbol{v}}_m^{(r)},_l - \hat{\boldsymbol{v}}_{m,k}^0$ 记作 $m \in [M-1]$，并记作 $\Delta\pi_k^{(r)} = \sum_{l \in \mathbb{V}_k} \pi_l^{(r)} - \pi_k^0$。根据公式（7）和 $\mathcal{D}_{\mathrm{V}}(G^{(r)}, G_0) \to 0$，所有这些量都趋于零。

步骤1：有限泰勒表示。定义真实门归一化因子 $Z_0(\hat{\boldsymbol{x}}) := \sum_{k=1}^{K_0} \pi_k^0 \exp\big((\hat{\boldsymbol{\omega}}_k^0)^\top \hat{\boldsymbol{x}}\big)$. 对于每个类别$m \in [M]$，考虑重缩放后的差

$$T_r(m, \hat{\boldsymbol{x}}) := Z_0(\hat{\boldsymbol{x}}) \Big(s_{G^{(r)}}(m \mid \boldsymbol{x}) - s_{G_0}(m \mid \boldsymbol{x}) \Big). \tag{29}$$

利用 $\mathsf{g}_l(\cdot;\ G)$ 的 softmax 形式和 $\mathsf{e}_l(\cdot \mid \cdot; G)$ 的多项逻辑形式，对每个真实原子 η_k^0 进行泰勒展开，得到分解

$$T_r(m, \hat{\boldsymbol{x}}) = \sum_{k=1}^{K_0} \sum_{\xi \in \Xi_k} A_{k,\xi}^{(r)}\, \Psi_{k,\xi}(m, \hat{\boldsymbol{x}}) + R_r(m, \hat{\boldsymbol{x}}), \tag{30}$$

其中：(i) $\{\Psi_{k,\xi}(m,\ \hat{\boldsymbol{x}})\}$ 是由解析基函数构成的有限集合，其形式为

$$\hat{\boldsymbol{x}}^\alpha \exp\big((\hat{\boldsymbol{\omega}}_k^0)^\top \hat{\boldsymbol{x}}\big)\, \partial^\beta \mathsf{e}(\, m \mid \hat{\boldsymbol{x}}; \bar{v}_k^0, \hat{\boldsymbol{v}}_k^0),$$

其中多重指标 α，β 的总阶数有界（当 $|\mathbb{V}_k| = 1$ 时为一阶，当 $|\mathbb{V}_k| > 1$ 时至多为二阶）；(ii) 系数 $A_{k,\xi}^{(r)}$ 是关于 $l \in \mathbb{V}_k$ 的显式有限和，其项为 $\Delta\hat{\boldsymbol{\omega}}_{lk}^{(r)}$、$\Delta\bar{v}_{lkm}^{(r)}$、$\Delta\hat{\boldsymbol{v}}_{lkm}^{(r)}$ 和 $\Delta\pi_k^{(r)}$ 的单项式，并以 $\pi_l^{(r)}$ 加权；(iii) 余项满足

$$\sup_{m \in [M]} \mathbb{E}_{\mathbf{x}}\big[\, |R_r(m, \phi_D(\mathbf{x}))| \,\big] = o\big(\mathcal{D}_{\mathrm{V}}(G^{(r)}, G_0)\big), \qquad (r \to \infty). \tag{31}$$

步骤 2：某个归一化系数必须非零。令

$$m_r := \max_{k,\xi} \frac{|A_{k,\xi}^{(r)}|}{\mathcal{D}_{\mathrm{V}}(G^{(r)}, G_0)}.$$

我们断言 $\liminf_{r\to\infty} m_r > 0$。否则，$m_r \to 0$，那么公式（30）中的所有系数都将满足 $|A_{k,\xi}^{(r)}| = o(\mathcal{D}_{\mathrm{V}}(G^{(r)}, G_0))$。但是，系数集合 $\{A_{k,\xi}^{(r)}\}$ 按构造包含了公式（7）中出现的相同的一阶和二阶加权增量：对于单元 $|\mathbb{V}_k| = 1$，它包含加权线性项$\sum_{l\in\mathbb{V}_k} \pi_l^{(r)} \|\Delta \hat{\boldsymbol{\omega}}_{lk}^{(r)}\|, \sum_{l\in\mathbb{V}_k} \pi_l^{(r)} |\Delta \bar{v}_{lkm}^{(r)}|, \sum_{l\in\mathbb{V}_k} \pi_l^{(r)} \|\Delta \hat{\boldsymbol{v}}_{lkm}^{(r)}\|$；对于单元 $|\mathbb{V}_k| > 1$，它包含加权二次项 $\sum_{l\in\mathbb{V}_k} \pi_l^{(r)} \|\Delta \hat{\boldsymbol{\omega}}_{lk}^{(r)}\|^2, \sum_{l\in\mathbb{V}_k} \pi_l^{(r)} |\Delta \bar{v}_{lkm}^{(r)}|^2, \sum_{l\in\mathbb{V}_k} \pi_l^{(r)} \|\Delta \hat{\boldsymbol{v}}_{lkm}^{(r)}\|^2$，以及质量差异项 $|\Delta \pi_k^{(r)}|$。因此，$m_r \to 0$ 将迫使 $\mathcal{D}_{\mathrm{V}}(G^{(r)}, G_0) = o(\mathcal{D}_{\mathrm{V}}(G^{(r)}, G_0))$，这与假设矛盾。因此 $\liminf_{r\to\infty} m_r > 0$。

步骤 3：Fatou 引理和线性无关性导出矛盾。将公式（30）除以 $m_r \mathcal{D}_{\mathrm{V}}(G^{(r)}, G_0)$ 并取协变量期望。利用公式（28）、定义公式（29）以及公式（31），可得

$$\mathbb{E}_{\mathbf{x}}\left[\sum_{m=1}^{M} \frac{|T_r(m, \phi_D(\mathbf{x}))|}{m_r \mathcal{D}_{\mathrm{V}}(G^{(r)}, G_0)}\right] \longrightarrow 0.$$

根据 Fatou 引理，对每个 $m \in [M]$，我们都可以提取一个子序列（不重新标记），使得

$$\frac{T_r(m, \hat{\boldsymbol{x}})}{m_r \mathcal{D}_{\mathrm{V}}(G^{(r)}, G_0)} \longrightarrow 0 \qquad \text{for a.e. } \hat{\boldsymbol{x}} \in \widehat{\mathbb{X}}.$$

沿着这个子序列，归一化系数（由有界性）收敛至极限$a_{k,\xi} = \lim_{r\to\infty} A_{k,\xi}^{(r)} / (m_r \mathcal{D}_{\mathrm{V}}(G^{(r)}, G_0))$，且至少有一个$a_{k,\xi} \neq 0$，因为$\liminf m_r > 0$。对公式（30）取极限可得，对每个 $m \in [M]$ 以及几乎处处的$\hat{\boldsymbol{x}}$,

$$\sum_{k=1}^{K_0} \sum_{\xi\in\Xi_k} a_{k,\xi}\, \Psi_{k,\xi}(m, \hat{\boldsymbol{x}}) = 0. \tag{32}$$

最后，根据 SGMLMoE 的可辨识性以及$\hat{\boldsymbol{\omega}}_1^0, \ldots, \hat{\boldsymbol{\omega}}_{K_0}^0$互不相同这一事实，解析函数族$\sim\{\Psi_{k,\xi}(m, \cdot)\}_{k,\xi}$ 在 $\mathbb{X}$上线性无关。因此公式（32）迫使所有$a_{k,\xi} = 0$均为零，这与至少一个非零相矛盾。这个矛盾证明了局部界公式（27）。

全局结构。根据局部结果，存在 $\varepsilon' > 0$ 和 $c' > 0$ 使得

$$\mathcal{D}_{\mathrm{TV}}(s_G, s_{G_0}) \geq c'\, \mathcal{D}_{\mathrm{V}}(G, G_0) \qquad \text{whenever } \mathcal{D}_{\mathrm{V}}(G, G_0) \leq \varepsilon'.$$

剩下的工作是在集合 $\{G \in \mathcal{O}_K(\mathbb{T}) : \mathcal{D}_{\mathrm{V}}(G, G_0) \geq \varepsilon'\}$上证明互补界：

$$\inf_{\substack{G\in\mathcal{O}_K(\mathbb{T}):\\ \mathcal{D}_{\mathrm{V}}(G,G_0)\geq\varepsilon'}} \frac{\mathcal{D}_{\mathrm{TV}}(s_G, s_{G_0})}{\mathcal{D}_{\mathrm{V}}(G, G_0)} > 0. \tag{33}$$

假设方程（33）不成立。则存在 $G'^{(r)} \in \mathcal{O}_K(\mathbb{T})$，使得 $\mathcal{D}_{\mathrm{V}}(G'^{(r)}, G_0) \geq \varepsilon'$ 且 $\mathcal{D}_{\mathrm{TV}}(s_{G'^{(r)}}, s_{G_0}) \to 0$。根据对 $\mathbb{T}$ 的既定紧致性假设（因而在弱拓扑下对 $\mathcal{O}_K(\mathbb{T})$ 也成立），取一个子序列，使得$G'^{(r)} \Rightarrow G'$，其中 $G' \in \mathcal{O}_K(\mathbb{T})$。由 $G \mapsto s_G(\cdot \mid \boldsymbol{x})$ 的连续性和控制收敛，$\mathcal{D}_{\mathrm{TV}}(s_{G'}, s_{G_0}) = 0$，因此对几乎处处的 $\boldsymbol{x}$，有 $s_{G'}(\cdot \mid \boldsymbol{x}) = s_{G_0}(\cdot \mid \boldsymbol{x})$。根据 SGMLMoE 的可辨识性，这意味着 $G' \equiv G_0$，与$\mathcal{D}_{\mathrm{V}}(G'^{(r)}, G_0) \geq \varepsilon'$ 矛盾。因此方程（33）成立；结合局部部分和整体部分可得方程（8）。证明完毕。

D.5. 定理 4 的证明

我们证明链条中的第一个关系，$\mathcal{D}_{\mathrm{V}}\big(G^{(K)},\ G_0\big) \gtrsim \mathcal{D}_{\mathrm{V}}\big(G^{(K-1)},\ G_0\big)$，，因为其余不等式均可通过在每次合并步骤中重复同样的论证得出。

全文中，我们将一般的 K-原子测度写作

$$G = \sum_{l=1}^{K} \pi_l\, \delta_{\eta_l}, \qquad \pi_l = \exp(\bar{\omega}_l), \qquad \eta_l = \Big(\hat{\boldsymbol{\omega}}_l, \bar{\omega}_l, \{(\hat{\boldsymbol{\upsilon}}_{m,l}, \bar{\upsilon}_{m,l})\}_{m=1}^{M-1}\Big) \in \mathbb{T},$$

并使用由 G_0 诱导的沃罗诺伊划分 $\{\mathbb{V}_k\}_{k=1}^{K_0}$，如第4.3节所示。对于 $l \in \mathbb{V}_k$，回顾参数间隔$\Delta\hat{\boldsymbol{\omega}}_{lk} = \hat{\boldsymbol{\omega}}_l - \hat{\boldsymbol{\omega}}_k^0$, $\Delta\bar{\upsilon}_{lkm} = \bar{\upsilon}_{m,l} - \bar{\upsilon}_{m,k}^0$, $\Delta\hat{\boldsymbol{\upsilon}}_{lkm} = \hat{\boldsymbol{\upsilon}}_{m,l} - \hat{\boldsymbol{\upsilon}}_{m,k}^0$，其中 $m \in [M-1]$。

步骤1：最近的一对位于同一个沃罗诺伊单元中。设 $G^{(K)}$ 充分接近 G_0，使得沃罗诺伊单元 $\mathbb{V}_k$ 定义良好且稳定。由于公式（9）中的合并规则使用 $\hat{\boldsymbol{\omega}}$和 $(\bar{\upsilon}_m, \hat{\boldsymbol{\upsilon}}_m)_{m\in[M-1]}$中的平方欧氏不相似度，属于不同沃罗诺伊单元的任意一对原子之间都保持固定的正间隔（该间隔仅取决于 G_0 以及 $\mathbb{T}$上的局部度量）；而发生过拟合时，同一单元内的原子对则可以任意接近。因此，当 $G^{(K)}$ 足够接近 G_0时，$h^{(K)} = \min_{l\neq l'} \mathrm{d}(\pi_l\delta_{\eta_l}, \pi_{l'}\delta_{\eta_{l'}})$ 中使目标最小的原子对必定属于同一个单元 $\mathbb{V}_k$。固定这样一个单元并将其记为 $\mathbb{V}_{k^\star}$，并假设合并索引为 $(l_1, l_2) \subset \mathbb{V}_{k^\star}$。

步骤 2：合并的记号。 令 $G^{(K-1)}$ 由 $G^{(K)}$ 通过合并原子 l_1 和 l_2 得到，合并方式遵循方程 (10)。将合并后的原子记为(π_*, η_*)，其中

$$\pi_* = \pi_{l_1} + \pi_{l_2}, \qquad \hat{\boldsymbol{\omega}}_* = \frac{\pi_{l_1}}{\pi_*}\hat{\boldsymbol{\omega}}_{l_1} + \frac{\pi_{l_2}}{\pi_*}\hat{\boldsymbol{\omega}}_{l_2},$$

并且，对每个 $m \in [M-1]$,

$$\bar{\upsilon}_{m,*} = \frac{\pi_{l_1}}{\pi_*}\bar{\upsilon}_{m,l_1} + \frac{\pi_{l_2}}{\pi_*}\bar{\upsilon}_{m,l_2}, \qquad \hat{\boldsymbol{\upsilon}}_{m,*} = \frac{\pi_{l_1}}{\pi_*}\hat{\boldsymbol{\upsilon}}_{m,l_1} + \frac{\pi_{l_2}}{\pi_*}\hat{\boldsymbol{\upsilon}}_{m,l_2}.$$

所有其他原子均保持不变。由于 $l_1,\ l_2 \in \mathbb{V}_{k^\star}$ 且 $G^{(K)}$ 接近 G_0，合并后的原子也仍被分配到同一个单元 $\mathbb{V}_{k^\star}$。

步骤 3：沃罗诺伊权重失配项保持不变。 对于每个 $k \in [K_0]$，在 $\mathbb{V}_k$ 内进行合并不会改变权重失配贡献 $\big|\sum_{l\in\mathbb{V}_k} \pi_l - \pi_k^0\big|$：事实上，对于 $k \neq k^\star$，没有任何变化；而对于 $k = k^\star$,

$$\sum_{l\in\mathbb{V}_{k^\star}} \pi_l = \sum_{l\in\mathbb{V}_{k^\star}\setminus\{l_1,l_2\}} \pi_l + \pi_{l_1} + \pi_{l_2} = \sum_{l\in\mathbb{V}_{k^\star}\setminus\{l_1,l_2\}} \pi_l + \pi_*,$$

因此，相对于 $\pi_{k^\star}^0$ 的绝对偏差在合并前后完全相同。

步骤 4：单元内二次项在重心合并下减小。首先考虑过度覆盖单元上的门控斜率二次项 $|\mathbb{V}_{k^\star}| > 1$:

$$\sum_{l\in\mathbb{V}_{k^\star}} \pi_l \|\Delta\hat{\boldsymbol{\omega}}_{lk^\star}\|^2 = \sum_{l\in\mathbb{V}_{k^\star}\setminus\{l_1,l_2\}} \pi_l \|\Delta\hat{\boldsymbol{\omega}}_{lk^\star}\|^2 + \pi_{l_1}\|\Delta\hat{\boldsymbol{\omega}}_{l_1k^\star}\|^2 + \pi_{l_2}\|\Delta\hat{\boldsymbol{\omega}}_{l_2k^\star}\|^2.$$

由 $u \mapsto \|u\|^2$的凸性以及定义 $\Delta\hat{\boldsymbol{\omega}}_{*k^\star} = \frac{\pi_{l_1}}{\pi_*}\Delta\hat{\boldsymbol{\omega}}_{l_1k^\star} + \frac{\pi_{l_2}}{\pi_*}\Delta\hat{\boldsymbol{\omega}}_{l_2k^\star}$ 可得詹森不等式

$$\pi_{l_1}\|\Delta\hat{\boldsymbol{\omega}}_{l_1k^\star}\|^2 + \pi_{l_2}\|\Delta\hat{\boldsymbol{\omega}}_{l_2k^\star}\|^2 \geq \pi_*\|\Delta\hat{\boldsymbol{\omega}}_{*k^\star}\|^2.$$

因此，

$$\sum_{l\in\mathbb{V}_{k^\star}} \pi_l \|\Delta\hat{\boldsymbol{\omega}}_{lk^\star}\|^2 \geq \sum_{l\in\mathbb{V}_{k^\star}\setminus\{l_1,l_2\}} \pi_l \|\Delta\hat{\boldsymbol{\omega}}_{lk^\star}\|^2 + \pi_*\|\Delta\hat{\boldsymbol{\omega}}_{*k^\star}\|^2,$$

这正是将 $l_1,\ l_2$ 合并到 $*$后对应的贡献。

同样的论证 适用于每个专家块，对于每个 $m \in [M-1]$：由于 $u \mapsto |u|^2$ 和 $u \mapsto \|u\|^2$ 是凸 x,

$$\pi_{l_1}|\Delta\bar{v}_{l_1k^\star m}|^2 + \pi_{l_2}|\Delta\bar{v}_{l_2k^\star m}|^2 \geq \pi_*|\Delta\bar{v}_{*k^\star m}|^2,$$

$$\pi_{l_1}\|\Delta\hat{\boldsymbol{v}}_{l_1k^\star m}\|^2 + \pi_{l_2}\|\Delta\hat{\boldsymbol{v}}_{l_2k^\star m}\|^2 \geq \pi_*\|\Delta\hat{\boldsymbol{v}}_{*k^\star m}\|^2,$$

因此，公式（7）中单元 $\mathbb{V}_{k^\star}$ 的全部单元内二次项在合并后不会增加。

步骤 5：结论。其他所有单元均不受影响，而 $\mathcal{D}_{\mathrm{E}}(G,\ G_0)$ 中的线性（单重覆盖）项保持不变，因为它们仅取决于权重和单例。因此，

$$\mathcal{D}_{\mathrm{V}}\big(G^{(K)}, G_0\big) \ \gtrsim\ \mathcal{D}_{\mathrm{V}}\big(G^{(K-1)}, G_0\big),$$

其中绝对常数（特别是，在单元内比较中，隐含常数可以取1 ）。沿着连续的合并步骤重复同样的论证，即可得到定理 4中所述的完整链条。 □

D.6. 定理 5的证明

在全文中， $\lesssim, \gtrsim$ 中的隐含常数均为通用常数，且可能随行变化。

有限混合测度的 Wasserstein距离。对于两个混合测度

$$G = \sum_{l=1}^{K} \pi_l\, \delta_{\eta_l}, \qquad G' = \sum_{l'=1}^{K'} \pi'_{l'}\, \delta_{\eta'_{l'}},$$

以及任意 $r \geq 1$，定义 Wasserstein-r 距离

$$W_r(G, G') = \left(\inf_{Q \in \Pi(\pi,\pi')} \sum_{l=1}^{K} \sum_{l'=1}^{K'} Q_{ll'}\, \|\eta_l - \eta'_{l'}\|^r \right)^{1/r},$$

其中， $\Pi(\pi,\pi')$ 是 $\pi = (\pi_1, \ldots, \pi_K)$ 与 $\pi' = (\pi'_1, \ldots, \pi'_{K'})$ 之间的耦合集合，即

$$\Pi(\pi,\pi') = \Big\{ Q \in \mathbb{R}_+^{K\times K'} : \sum_{l'=1}^{K'} Q_{ll'} = \pi_l,\ \sum_{l=1}^{K} Q_{ll'} = \pi'_{l'},\ \forall l \in [K],\ l' \in [K'] \Big\}.$$

一个有用的下界（Wasserstein距离控制分量间差异）。固定真实混合测度 $G_0 = \sum_{k=1}^{K_0} \pi_k^0\, \delta_{\eta_k^0} \in \mathcal{O}_{K_0}\ \mathbb{T}$ $G = \sum_{l=1}^{K} \pi_l\, \delta_{\eta_l} \in \mathcal{O}_K\ \mathbb{T}$ $W_r\ G\ G_0\ \to 0$ () 并考虑() 使得(,)。那么（例如，参见（Ho &Nguyen，2019年））存在一个沃罗诺伊分配 $l \mapsto k(l)\ \in [K_0]$ 以及单元 $\mathbb{V}_k = \{l \in [K] :\ k(l)\ = k\}$ 使得

$$W_r^r(G, G_0)\ \gtrsim\ \sum_{k=1}^{K_0} \left(\Big| \sum_{l\in\mathbb{V}_k} \pi_l - \pi_k^0 \Big| + \sum_{l\in\mathbb{V}_k} \pi_l\, \|\eta_l - \eta_k^0\|^r \right). \tag{34}$$

第 1 部分：过拟合层级上的沃罗诺伊划分与高度速率 $\kappa \in \{K_0+1, \ldots, K\}$。令

$$A_N := \Big\{ \mathcal{D}_{\mathrm{TV}}(s_{\widehat{G}_N}, s_{G_0}) \leq C\sqrt{\log N/N} \Big\},$$

其中 $C > 0$是 事实 1中的常数。根据 事实 1， $\mathbb{P}(A_N)\ \geq 1 - c_1 N^{-c_2}$. 在 A_N上，结合 定理 3 与密度界可得

$$\mathcal{D}_{\mathrm{V}}(\widehat{G}_N, G_0)\ \lesssim\ \sqrt{\log N/N}. \tag{35}$$

接下来，根据定理 4，沃罗诺伊损失沿合并链单调变化（差一个常数），因此对于每个 $\kappa \in \{K_0+1,\ \ldots,\ K\}$,

$$\mathcal{D}_{\mathrm{V}}(\widehat{G}_N^{(\kappa)}, G_0)\ \lesssim\ \sqrt{\log N/N}. \tag{36}$$

现在我们来界定树状图高度 $h_N^{(\kappa)}$ 对于 $\kappa \geq K_0+1$。由于 $\kappa > K_0$，在层级 κ 至少存在一个包含 $\widehat{G}_N^{(\kappa)}$ 中两个不同原子的沃罗诺伊单元 $\mathbb{V}_{k^\star}$。固定这样一个单元 $l_1 \neq l_2 \in \mathbb{V}_{k^\star}$并选取两个不同的索引$\widehat{G}_N^{(\kappa)} = \sum_{l=1}^{\kappa} \pi_l^{(\kappa)} \delta_{\eta_l^{(\kappa)}}$。记 π_l 并简写$= \pi_l^{(\kappa)}$， $\eta_l := \eta_l^{(\kappa)}$. 根据 $\mathcal{D}_{\mathrm{V}}$ 在公式（7）和公式（36）中的定义，我们得到

$$\pi_{l_1}\Big(\|\hat{\boldsymbol{\omega}}_{l_1} - \hat{\boldsymbol{\omega}}_{k^\star}^0\|^2 + \sum_{m=1}^{M-1}\big(|\bar{v}_{m,l_1} - \bar{v}_{m,k^\star}^0|^2 + \|\hat{\boldsymbol{v}}_{m,l_1} - \hat{\boldsymbol{v}}_{m,k^\star}^0\|^2\big)\Big)$$
$$+ \pi_{l_2}\Big(\|\hat{\boldsymbol{\omega}}_{l_2} - \hat{\boldsymbol{\omega}}_{k^\star}^0\|^2 + \sum_{m=1}^{M-1}\big(|\bar{v}_{m,l_2} - \bar{v}_{m,k^\star}^0|^2 + \|\hat{\boldsymbol{v}}_{m,l_2} - \hat{\boldsymbol{v}}_{m,k^\star}^0\|^2\big)\Big) \lesssim \sqrt{\log N/N}. \tag{37}$$

利用初等不等式

$$\min\{\pi_{l_1}, \pi_{l_2}\} \geq \frac{\pi_{l_1}\pi_{l_2}}{\pi_{l_1}+\pi_{l_2}},$$

以及三角不等式 $\|\hat{\boldsymbol{\omega}}_{l_1} - \hat{\boldsymbol{\omega}}_{l_2}\| \leq \|\hat{\boldsymbol{\omega}}_{l_1} - \hat{\boldsymbol{\omega}}_{k^\star}^0\| + \|\hat{\boldsymbol{\omega}}_{l_2} - \hat{\boldsymbol{\omega}}_{k^\star}^0\|$(专家参数也同理)，我们可以将 公式（37） 左侧下界为不相似度 $\mathrm{d}(\pi_{l_1}\delta_{\eta_{l_1}}, \pi_{l_2}\delta_{\eta_{l_2}})$的一个常数倍，其中该不相似度定义于公式（9）。具体而言，存在一个通用的 $c>0$ 使得

$$\mathrm{d}\big(\pi_{l_1}\delta_{\eta_{l_1}}, \pi_{l_2}\delta_{\eta_{l_2}}\big) \leq c \cdot \text{LHS of Equation (37)} \lesssim \sqrt{\log N/N}.$$

由于 $h_N^{(\kappa)}$ 是层级 κ中所有原子对之间的最小不相似度，

$$h_N^{(\kappa)} \lesssim \sqrt{\log N/N}, \qquad \forall \kappa \in \{K_0+1, \ldots, K\}. \tag{38}$$

第二部分：欠拟合层级 $\kappa' \in [K_0]$。 在精确层级 K_0，每个沃罗诺伊单元都是单点集。根据公式（7）可知， $\mathcal{D}_{\mathrm{V}}$ $(\widehat{G}^{(K_0)}{}_N, G_0)$ 等价于（相差一个常数因子）两个含有 W_1 个原子的测度之间的 K_0 型差异。特别地，

$$W_1(\widehat{G}_N^{(K_0)}, G_0) \lesssim \sqrt{\log N/N}. \tag{39}$$

记

$$\widehat{G}_N^{(K_0)} = \sum_{k=1}^{K_0} \pi_{k,N}\,\delta_{\eta_{k,N}}, \qquad G_0 = \sum_{k=1}^{K_0} \pi_k^0\,\delta_{\eta_k^0},$$

在由 Voronoi 构造所蕴含的自然匹配下（当 N 在 A_N上足够大时成立）。则 方程 (39)蕴含逐分量控制，例如，

$$|\pi_{k,N} - \pi_k^0| \lesssim \sqrt{\log N/N}, \qquad \|\eta_{k,N} - \eta_k^0\| \lesssim \sqrt{\log N/N}, \qquad k \in [K_0],$$

因此，对任意固定的配对（k_1， k_2），水平 K_0 处的成对不相似度满足扰动界

$$\Big|\mathrm{d}\big(\pi_{k_1,N}\delta_{\eta_{k_1,N}}, \pi_{k_2,N}\delta_{\eta_{k_2,N}}\big) - \mathrm{d}\big(\pi_{k_1}^0\delta_{\eta_{k_1}^0}, \pi_{k_2}^0\delta_{\eta_{k_2}^0}\big)\Big| \lesssim \sqrt{\log N/N}, \tag{40}$$

其中我们用到了所有相关参数均一致有界（且 π_k^0 有界并远离 0）。

现令 （$k_1^\star$， $k_2^\star$）为真实高度的（为简单起见，唯一的）最小化配对 $h^{(K_0)}{}_0 \quad = \min_{k_1 \neq k_2} \mathrm{d}$（$\pi_{k_1}^0\delta_{\eta_{k_1}^0}$， $\pi_{k_2}^0\delta_{\eta_{k_2}^0}$）。由 方程 (40)，当 N 足够大时，水平 K_0 处经验不相似度的最小化配对与 （$k_1^\star$， $k_2^\star$）重合，因此

$$|h_N^{(K_0)} - h_0^{(K_0)}| \lesssim \sqrt{\log N/N}.$$

将该配对合并一次后，合并后的原子参数为如同 方程 (10)中的重心，因此同一利普希茨型扰动论证（利用 方程 (39) 以及参数/权重有界并远离 0）可得

$$W_1\big(\widehat{G}_N^{(K_0-1)}, G_0^{(K_0-1)}\big) \lesssim \sqrt{\log N/N}, \qquad |h_N^{(K_0-1)} - h_0^{(K_0-1)}| \lesssim \sqrt{\log N/N}.$$

沿着链条归纳地重复这一论证，即可证明对每个 $\kappa' \in [K_0]$,

$$|h_N^{(\kappa')} - h_0^{(\kappa')}| \lesssim \sqrt{\log N/N}. \tag{41}$$

结论。 结合 公式（36）、 公式（38）和 公式（41），并回顾 $\mathbb{P}(A_N) \geq 1 - c_1 N^{-c_2}$，完成了 定理 5的证明。 □

D.7. 定理 定理 6 的证明

假设 $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)$ 是从 G_0 中独立同分布抽取的，并将经验测度记为 $P_N := \frac{1}{N} \sum_{n=1}^{N} \delta_{(\boldsymbol{x}_n, y_n)}$（其中 $(\boldsymbol{x}_n, y_n)$ 为观测样本）。对于任意混合测度 $G \in \mathcal{O}_K(\mathbb{T})$，定义经验过程

$$\nu_N(G) := \sqrt{N}\,(P_N - P_{G_0}) \log \frac{s_G}{s_{G_0}}.$$

我们将反复使用以下指数不等式。

事实 2 (（Van de Geer，2000年）中的定理 5.11）。令 R，C，C_1，$a > 0$ 满足

$$a \leq C_1 \sqrt{N}\, R^2 \;\wedge\; 8\sqrt{N}\, R,$$

and

$$a \geq \sqrt{C^2(C_1+1)} \left(\int_{a/(2^6\sqrt{N})}^{R} H_B^{1/2}\left(\frac{u}{\sqrt{2}},\, \{\, s_G :\, h(s_G, s_{G_0}) \leq R \,\},\, \nu \right) du \;\vee\; R \right).$$

Then

$$\mathbb{P}_{G_0}\left(\sup_{h(s_G, s_{G_0}) \leq R} \left| \nu_N(G) \right| \geq a \right) \leq C \exp\left(-\frac{a^2}{C^2(C_1+1)R^2} \right).$$

情形 1： $\kappa \geq K_0$。对于 $\kappa \in \{K_0, \ldots, K\}$，令 $\widehat{G}_N^{(\kappa)}$ 表示沿树状图路径的 κ-原子混合测度，并记 $\bar{\ell}_N^{(\kappa)} = \bar{\ell}_N(\widehat{G}_N^{(\kappa)}) = \frac{1}{N} \sum_{n=1}^{N} \log s_{\widehat{G}_N^{(\kappa)}}(y_n \mid \boldsymbol{x}_n)$。

步骤*1*（凹性界）。利用 log 的凹性，对于任意两个密度 p，q，

$$\frac{1}{2} \log \frac{p}{q} \leq \log \frac{p+q}{2q}.$$

将 $p = s_{\widehat{G}_N^{(\kappa)}}(y \mid \boldsymbol{x})$ 和 $q = s_{G_0}(y \mid \boldsymbol{x})$ 代入可得

$$\frac{1}{2} P_N \log \frac{s_{\widehat{G}_N^{(\kappa)}}}{s_{G_0}} \leq P_N \log \frac{\bar{s}_{\widehat{G}_N^{(\kappa)}}}{s_{G_0}}, \qquad \bar{s}_{\widehat{G}_N^{(\kappa)}} := \frac{s_{\widehat{G}_N^{(\kappa)}} + s_{G_0}}{2}.$$

此外，

$$P_N \log \frac{\bar{s}_{\widehat{G}_N^{(\kappa)}}}{s_{G_0}} = (P_N - P_{G_0}) \log \frac{\bar{s}_{\widehat{G}_N^{(\kappa)}}}{s_{G_0}} - D_{\mathrm{KL}}\left(s_{G_0} \,\|\, \bar{s}_{\widehat{G}_N^{(\kappa)}} \right) \leq (P_N - P_{G_0}) \log \frac{\bar{s}_{\widehat{G}_N^{(\kappa)}}}{s_{G_0}}.$$

因此，

$$P_N \log s_{\widehat{G}_N^{(\kappa)}} - \mathcal{L}(G_0) = P_N \log \frac{s_{\widehat{G}_N^{(\kappa)}}}{s_{G_0}} + (P_N - P_{G_0}) \log s_{G_0} \leq 2(P_N - P_{G_0}) \log \frac{s_{\widehat{G}_N^{(\kappa)}}}{s_{G_0}} + (P_N - P_{G_0}) \log s_{G_0}. \quad (42)$$

Step 2（过拟合层级上的 *Hellinger* 半径）。根据定理 5，以至少 $1 - c_1 N^{-c_2}$的概率，

$$\mathcal{D}_{\mathrm{V}}(\widehat{G}_N^{(\kappa)}, G_0) \;\lesssim\; \sqrt{\log N / N}, \qquad \kappa \in \{K_0, \ldots, K\}.$$

由于 $W_2^2(\cdot,\ \cdot) \lesssim \mathcal{D}_{\mathrm{V}}(\cdot,\ \cdot)$（对于采用沃罗诺伊构造的有界参数集上的测度），我们在同一事件上得到

$$W_2(\widehat{G}_N^{(\kappa)}, G_0) \;\lesssim\; (\log N / N)^{1/4}.$$

使用标准不等式 $h(s_G,\ s_{G_0}) \lesssim W_2(G,\ G_0)$ 针对有界定义域上的这些有限混合/MoE 族（例如参见（Nguyen，2013年）），我们推得存在 $R_N \asymp$(对数 $N/N)^{1/4}$ 以及一个事件 A_N满足 $\mathbb{P}_{G_0}(A_N) \geq 1 - c_1 N^{-c_2}$使得

$$h\left(s_{\widehat{G}_N^{(\kappa)}}, s_{G_0} \right) \leq R_N, \qquad \forall \kappa \in \{K_0, \ldots, K\}. \quad (43)$$

步骤 *3*（经验过程控制）。设定 $R := R_N$ 属于 事实 2 并选择

$$a := \sqrt{N}\, R_N \log(1/R_N) \asymp N^{1/4} \log^{1/4} N \cdot \log N$$

(其阶与 $\sqrt{N R_N \log(1/R_N)}$相同)。利用熵界 $H_B^{1/2}(u, \{s_G : h(s_G, s_{G_0}) \le R\}, \nu) \lesssim \log(1/u)$ (例如，(Nguyen等人, 2024年b)中的引理 4)，当 事实 2 的条件对 N 足够大时成立，从而得到

$$\mathbb{P}_{G_0}\left(\sup_{h(s_G, s_{G_0}) \le R_N} \left|\nu_N(G)\right| \gtrsim a \right) \lesssim N^{-c}$$

对于某个 $c > 0$。等价地，两边除以 $\sqrt{N}$,

$$\mathbb{P}_{G_0}\left(\sup_{h(s_G, s_{G_0}) \le R_N} \left|(P_N - P_{G_0}) \log(s_G / s_{G_0})\right| \gtrsim R_N \log(1/R_N) \right) \lesssim N^{-c}. \tag{44}$$

结合公式（43）与公式（44）可得，对于每个固定的 $\kappa \in \{K_0, \ldots, K\}$,

$$\left|(P_N - P_{G_0}) \log(s_{\widehat{G}_N^{(\kappa)}} / s_{G_0})\right| \lesssim R_N \log(1/R_N) \lesssim (\log N/N)^{1/4} \tag{45}$$

以至少 $1 - c_1' N^{-c_2'}$ 的概率。

步骤 *4*（控制 $(P_N - P_{G_0}) \log s_{G_0}$）。根据切比雪夫不等式,

$$\mathbb{P}_{G_0}\left(\left|(P_N - P_{G_0}) \log s_{G_0}\right| \ge t\right) \le \frac{\mathrm{Var}(\log s_{G_0})}{N t^2}. \tag{46}$$

取 $t \asymp (\log N/N)^{1/4}$ 可得

$$\left|(P_N - P_{G_0}) \log s_{G_0}\right| \lesssim (\log N/N)^{1/4} \tag{47}$$

以至少 $1 - \mathrm{Var}(\log s_{G_0})/\sqrt{\log N}$ 的概率。

步骤 *5*（关于 $\kappa \ge K_0$的结论）。将公式（45）和公式（47）代入公式（42）可得，且其概率趋于1,

$$\bar{\ell}_N^{(\kappa)} - \mathcal{L}(G_0) = P_N \log s_{\widehat{G}_N^{(\kappa)}} - \mathcal{L}(G_0) \lesssim (\log N/N)^{1/4}, \qquad \kappa \in \{K_0, \ldots, K\}.$$

这证明了定理 6的第一个结论。

情形 2： $\kappa = K_0$ **（使用条件 K 的匹配下界）**。在精确拟合层面，定理 5意味着

$$W_1(\widehat{G}_N^{(K_0)}, G_0) \lesssim \sqrt{\log N/N}.$$

令 $\epsilon_N := \sqrt{\log N/N}$。根据 $\mathcal{D}_V$ 在单点沃罗诺伊单元（精确拟合）处的定义，这意味着逐分量界$\|\eta_{k,N} - \eta_k^0\| \lesssim \epsilon_N$ 以及$|\pi_{k,N} - \pi_k^0| \lesssim \epsilon_N$（在规范匹配下），因此条件 K 给出逐点不等式

$$\log s_{\widehat{G}_N^{(K_0)}}(\mathrm{y} \mid \mathrm{x}) \ge (1 + c_\beta \epsilon_N) \log s_{G_0}(\mathrm{y} \mid \mathrm{x}) - c_\alpha \epsilon_N, \qquad (\mathrm{x}, \mathrm{y}) \in \mathbb{X} \times \mathbb{Y},$$

对于某些常数 c_α, $c_\beta > 0$。对 P_N 取平均并减去 $\mathcal{L}(G_0)$，得到

$$\bar{\ell}_N^{(K_0)} - \mathcal{L}(G_0) \ge -c_\alpha \epsilon_N + c_\beta \epsilon_N P_N \log s_{G_0} + (1 + c_\beta \epsilon_N)(P_N - P_{G_0}) \log s_{G_0}.$$

利用 公式（46） 和 $t = \epsilon_N$ 可得

$$(P_N - P_{G_0}) \log s_{G_0} = O_{\mathbb{P}}(\epsilon_N),$$

由于 $P_N \log s_{G_0} = P_{G_0} \log s_{G_0} + O_{\mathbb{P}}(\epsilon_N)$，我们得出

$$\bar{\ell}_N^{(K_0)} - \mathcal{L}(G_0) \ge -C \epsilon_N$$

以 $\mathbb{P}_{G_0}$-概率成立。结合在 $\kappa = K_0$处由情形 1得到的上界，可得

$$\left|\bar{\ell}_N^{(K_0)} - \mathcal{L}(G_0)\right| \lesssim (\log N/N)^{1/4}$$

(至多相差对数因子)，这与 定理 6中的陈述一致。

情形 3： $\kappa < K_0$。固定 $\kappa \in \{2, \ldots, K_0 - 1\}$。由 $\mathbb{X}$ 及参数空间的有界性，以及可积包络函数的存在性 $m(\boldsymbol{x},\ \mathrm{y})$ 使得$\sup_{G \in \mathcal{O}_\kappa}$ |对数因子 $s_G(\mathrm{y} \mid \boldsymbol{x})| \le m(\boldsymbol{x},\ \mathrm{y})$，类别 {对数因子 s_G : $G \in \mathcal{O}_\kappa$} 是 Glivenko– Cantelli 类。因此（例如Keener, 2010年，定理 9.2），

$$\sup_{G \in \mathcal{O}_\kappa} \left| \bar{\ell}_N(G) - \mathcal{L}(G) \right| \to 0 \qquad \text{in } \mathbb{P}_{G_0}\text{-probability.}$$

由于 $\widehat{G}_N^{(\kappa)}$ 在 $\mathcal{O}_\kappa(\mathbb{T})$ 上使 $\bar{\ell}_N(\cdot)$ 达到最大，因此

$$\bar{\ell}_N(\widehat{G}_N^{(\kappa)}) \to \sup_{G \in \mathcal{O}_\kappa(\mathbb{T})} \mathcal{L}(G) \qquad \text{in } \mathbb{P}_{G_0}\text{-probability.}$$

记 $G_0^{(\kappa)}$ 为 $\mathcal{L}(\cdot)$ 在 $\mathcal{O}_\kappa(\mathbb{T})$上的最大化点，由此得到定理 6中的第二个结论：

$$\bar{\ell}_N(\widehat{G}_N^{(\kappa)}) \to \mathcal{L}(G_0^{(\kappa)}) \qquad \text{in } \mathbb{P}_{G_0}\text{-probability.}$$

□

D.8. 第 第 4.6 节中条件 K 的证明

固定 $\boldsymbol{\theta}^0 \in \mathbb{T}$，并假设当 $\epsilon > 0$ 足够小时，$\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\| \le \epsilon$ 成立。根据有限维空间中的范数等价性，只需假设对所有 $m \in [M-1]$,

$$\|\hat{\boldsymbol{v}}_m - \hat{\boldsymbol{v}}_m^0\| \le \epsilon, \qquad |\bar{v}_m - \bar{v}_m^0| \le \epsilon,$$

（并且参考类别 $m = M$ 固定为 $(\hat{\boldsymbol{v}}_M, \bar{v}_M) = (\mathbf{0}, 0)$）。我们证明存在常数 $c_\alpha, c_\beta > 0$，使得对所有 $(\mathbf{x},\mathrm{y}) \in \mathbb{X} \times \mathbb{Y}$,

$$\log s_{\boldsymbol{\theta}}(\mathrm{y} \mid \mathbf{x}) \ge (1 + c_\beta \epsilon) \log s_{\boldsymbol{\theta}^0}(\mathrm{y} \mid \mathbf{x}) - c_\alpha \epsilon.$$

由于 $\mathbb{X}$ 是紧集且参数空间有界，得分映射 $(\hat{\boldsymbol{v}}, \bar{v}, \boldsymbol{x}) \mapsto \log \mathsf{e}(\mathrm{y} \mid \boldsymbol{x}; \hat{\boldsymbol{v}}, \bar{v})$ 关于 $\boldsymbol{x} \in \mathbb{X}$ 一致地局部利普希茨连续。因此，存在 $L > 0$，使得对所有 $m \in [M-1]$ 和所有 $\boldsymbol{x} \in \mathbb{X}$,

$$\left| (\hat{\boldsymbol{v}}_m - \hat{\boldsymbol{v}}_m^0)^\top \boldsymbol{x} \right| \le L\epsilon, \qquad |\bar{v}_m - \bar{v}_m^0| \le \epsilon.$$

此外，映射 $(\{\hat{\boldsymbol{v}}_\ell, \bar{v}_\ell\}_{\ell=1}^{M-1}, \boldsymbol{x}) \mapsto \log\bigl(1 + \sum_{\ell=1}^{M-1} \exp(\bar{v}_\ell + \hat{\boldsymbol{v}}_\ell^\top \boldsymbol{x})\bigr)$ 是光滑的，因此是局部利普希茨连续的，从而其扰动关于 $\boldsymbol{x} \in \mathbb{X}$ 一致为 $O(\epsilon)$。结合这些一致的利普希茨界可得

$$\log \mathsf{e}(\mathrm{y} \mid \boldsymbol{x}; \boldsymbol{\theta}) \ge \log \mathsf{e}(\mathrm{y} \mid \boldsymbol{x}; \boldsymbol{\theta}^0) - C\epsilon$$

对某个 $C > 0$ 以及所有 $(\boldsymbol{x}, \mathrm{y})$。最后，由于 $\log \mathsf{e}(\mathrm{y} \mid \boldsymbol{x}; \boldsymbol{\theta}^0) \le 0$ 且在 $\mathbb{X} \times \mathbb{Y}$ 上有下界，我们可以取充分小的 $c_\beta > 0$ 和充分大的 $c_\alpha > 0$，使得

$$\log \mathsf{e}(\mathrm{y} \mid \boldsymbol{x}; \boldsymbol{\theta}) \ge (1 + c_\beta \epsilon) \log \mathsf{e}(\mathrm{y} \mid \boldsymbol{x}; \boldsymbol{\theta}^0) - c_\alpha \epsilon.$$

对门控部分应用相同的推理，并在 $s_{\boldsymbol{\theta}}$ 内对专家求和，即可得到条件 K。 □

D.9. 定理 7 的证明

召回率 $\bar{\ell}_N^{(\kappa)} := \bar{\ell}_N(\widehat{G}_N^{(\kappa)})$ 和 $\mathrm{DSC}_N^{(\kappa)} := -\bigl(h_N^{(\kappa)} + \omega_N \bar{\ell}_N^{(\kappa)}\bigr)$。由 定理 5,

$$h_N^{(\kappa)} = \begin{cases} O_{\mathbb{P}}\bigl((\log N/N)^{1/2}\bigr), & \kappa > K_0, \\ h_0^{(\kappa)} + O_{\mathbb{P}}\bigl((\log N/N)^{1/2}\bigr), & \kappa \le K_0. \end{cases}$$

由 定理 6,

$$\bar{\ell}_N^{(\kappa)} = \begin{cases} \mathcal{L}(G_0) + O_{\mathbb{P}}\bigl((\log N/N)^{1/4}\bigr), & \kappa \ge K_0, \\ \mathcal{L}(G_0^{(\kappa)}) + o_{\mathbb{P}}(1), & \kappa < K_0, \end{cases}$$

以及 $\mathcal{L}(G_0^{(\kappa)}) < \mathcal{L}(G_0)$，对于 $\kappa < K_0$（等价地，$D_{\mathrm{KL}}(s_{G_0} \| s_{G_0^{(\kappa)}}) > 0$）。

因此,

$$\mathrm{DSC}_N^{(\kappa)} = \begin{cases} -\omega_N \mathcal{L}(G_0) + O_{\mathbb{P}}\big(\omega_N (\log N/N)^{1/4}\big), & \kappa > K_0, \\ -\omega_N \mathcal{L}(G_0) - h_0^{(K_0)} + O_{\mathbb{P}}\big(\omega_N (\log N/N)^{1/4}\big), & \kappa = K_0, \\ -\omega_N \mathcal{L}(G_0^{(\kappa)}) - h_0^{(\kappa)} + o_{\mathbb{P}}(\omega_N), & \kappa < K_0. \end{cases}$$

由于 $\omega_N \to \infty$ 且 ω_N(lo g $N/N)^{1/4} \to 0$，而 $\mathcal{L}(G_0) - \mathcal{L}(G_0^{(\kappa)}) > 0$ 对于 $\kappa < K_0$，我们有，当 N 足够大时 ，

$$\mathrm{DSC}_N^{(K_0)} < \min_{\kappa \neq K_0} \mathrm{DSC}_N^{(\kappa)}$$

以 $\mathbb{P}_{G_0}$-probabilit y趋于1。因此 $\widehat{K}_N = \arg\min^{\kappa \in \{2,\ldots,K\}} \mathrm{DSC}_N^{(\kappa)}$ 满足 $\widehat{K}_N \to K_0$，在 $\mathbb{P}_{G_0}$-probabili ty. □