Dendrograms of Mixing Measures for Softmax-Gated Gaussian Mixture of Experts: Consistency Without Model Sweeps

TienHai Do[⋆,1,2] **Trung Nguyen Mai**[⋆,2,3] **TrungTin Nguyen**[⋆,†,4,5]

Nhat Ho[6] **Binh T. Nguyen**[1,2] **Christopher Drovandi**[4,5]

[1]Faculty of Mathematics and Computer Science, University of Science, Ho Chi Minh City, Vietnam.

[2]Vietnam National University Ho Chi Minh City, Vietnam.

[3]Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam.

[4]ARC Centre of Excellence for the Mathematical Analysis of Cellular Systems.

[5]School of Mathematical Sciences, Queensland University of Technology, Brisbane City, Australia.

[6]Department of Statistics and Data Science, University of Texas at Austin, Austin, USA.

Abstract

We develop a unified statistical framework for softmax-gated Gaussian mixture of experts (SGMoE) that addresses three long-standing obstacles in parameter estimation and model selection: (i) non-identifiability of gating parameters up to common translations, (ii) intrinsic gate-expert interactions that induce coupled differential relations in the likelihood, and (iii) the tight numerator-denominator coupling in the softmax-induced conditional density. Our approach introduces Voronoi-type loss functions aligned with the gate-partition geometry and establishes finite-sample convergence rates for the maximum likelihood estimator (MLE). In over-specified models, we reveal a link between the MLE's convergence rate and the solvability of an associated system of polynomial equations characterizing near-nonidentifiable directions. For model selection, we adapt dendrograms of mixing measures to SGMoE, yielding a consistent, sweep-free selector of the number of experts that attains pointwise-optimal parameter rates under overfitting while avoiding multi-size training. Simulations on synthetic data corroborate the theory, accurately recovering the expert count and achieving the predicted rates for parameter estimation while closely approximating the regression function. Under model misspecification (e.g., ϵ-contamination), the dendrogram selection criterion is robust, recovering the true number of mixture components, while the Akaike information criterion, the Bayesian information criterion, and the integrated completed likelihood tend to overselect as sample size grows. On a maize proteomics dataset of drought-responsive traits, our dendrogram-guided SGMoE selects two experts, exposes a clear mixing-measure hierarchy, stabilizes the likelihood early, and yields interpretable genotype-phenotype maps, outperforming standard criteria without multi-size training.

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Mixture of Experts: Scope and Appeal. Mixture of experts (MoE) were introduced as modular neural architectures in Jacobs et al. (1991); Jordan & Jacobs (1994), where a gating network dispatches inputs to specialized experts. Beyond their practical versatility in speech, language, and vision (Bao et al., 2022; Do et al., 2023; Dosovitskiy et al., 2021; Eigen et al., 2014; Fedus et al., 2022; Liang et al., 2022; Peng et al., 1996;

[⋆]Co-first author, [†]Corresponding author.

Pham et al., 2024; You et al., 2021, 2022), MoE admit strong approximation guarantees and learning theory. Universal approximation results for conditional densities and regressors quantify how MoE improve upon unconditional mixtures by allowing both gates and experts to depend on covariates (Nguyen et al., 2019, 2016, 2021a; Norets, 2010). These developments complement classical approximation and risk bounds for unconditional mixtures (Chong et al., 2024; Genovese & Wasserman, 2000; Ho & Nguyen, 2016a,b; Nguyen et al., 2025b, 2023b, 2020; Nguyen, 2013; Rakhlin et al., 2005; Shen et al., 2013) and are surveyed in Chen et al. (2022); Nguyen & Chamroukhi (2018); Nguyen (2021); Yuksel et al. (2012).

Parameter Estimation: from Unconditional Mixtures to MoE. Over-specified finite mixtures can display slow, nonstandard parameter rates. In unconditional mixtures this is explained by singular Fisher information and merging components. Foundational results start with Chen (1995) for univariate mixtures, and extend via Wasserstein tools to multivariate models and weaker identifiability (Ho & Nguyen, 2016a; Nguyen, 2013), with minimax studies in Heinrich & Kahn (2018); Manole & Ho (2020). Algorithmic guarantees for Expectation-Maximization (EM) and Majorization-Minimization or Minimization-Maximization (MM) algorithms and moments have been analyzed under both exact-fit and over-fit regimes (Anandkumar et al., 2012; Balakrishnan et al., 2017; Doss et al., 2023; Dwivedi et al., 2020a,b; Hardt & Price, 2015; Tran et al., 2026b; Wu & Yang, 2020; Wu & Zhou, 2021). For MoE with covariate-free gates, parameter rates depend on algebraic independence of experts and PDE-type couplings (Do et al., 2025; Ho et al., 2022). In softmax-gated Gaussian mixture of experts (SGMoE), parameter estimation is harder due to translation invariance in softmax gates and intrinsic gate-expert couplings; recent progress includes identifiability, inverse bounds, and finite-sample guarantees for the maximum likelihood estimator (MLE) with unified exact- and over-fit treatments in Nguyen et al. (2024a, 2023a, 2024c).

Model Selection: Information Criteria, Penalties, and Bayes. Choosing the number of experts remains critical despite universal approximation theorems. Classical criteria balance fit and complexity, including AIC (Akaike, 1974; Frühwirth-Schnatter et al., 2018), BIC and its MoE adaptations (Berrettini et al., 2024; Forbes et al., 2022a,b; Ho et al., 2025; Khalili et al., 2024; Nguyen & Nguyen, 2025; Schwarz, 1978), ICL (Biernacki et al., 2000; Frühwirth-Schnatter et al., 2012), eBIC for structured settings (Foygel & Drton, 2010; Nguyen & Li, 2024), and SWIC for dependent data (Nguyen et al., 2025a; Sin & White, 1996; Westerhout et al., 2024). These methods are largely asymptotic and often require multi-size model sweeps. Non-asymptotic penalization brings risk guarantees via weak oracle bounds in high-dimensional MoE (Montuelle & Le Pennec, 2014; Nguyen et al., 2021b, 2022a, 2023c, 2022b, 2023d). Bayesian strategies avoid fixing the order but need careful marginal-likelihood evaluation or post-processing; the merge-truncate-merge approach ensures consistency in related mixture settings yet introduces sensitive tuning (Frühwirth-Schnatter, 2019; Guha et al., 2021; Nguyen et al., 2024d; Zens, 2019). A recent alternative leverages dendrograms of mixing measures for selection without exhaustive sweeps in (Do et al., 2024; Thai et al., 2025; Tran et al., 2026a).

Gaps Specific to SGMoE. Softmax gating creates three intertwined obstacles. First, gate parameters are identifiable only up to common translations, so parameter losses must factor out these symmetries. Second, the softmax numerator-denominator coupling and the expert structure induce exact PDE relations between derivatives, which collapse naive Taylor decompositions and require algebra-aware inverse bounds. Third, when models are over-specified, the first nonvanishing terms in the expansions are ruled by solvability of polynomial systems; the resulting exponents govern slow parameter rates and depend on how many fitted atoms approximate each truth (Ho et al., 2022; Nguyen et al., 2023a). Existing selection criteria do not exploit this rate geometry for the MLE, and sweep-based procedures are computationally heavy for SGMoE.

Contributions. We introduce a fast-rate-aware Voronoi distance for SGMoE that augments the unified exact- and over-fit loss with merged-moment couplings inside multi-covered Voronoi cells (eq. (6)). This exposes slow directions created by redundant atoms, motivates a hierarchical merge operator, and yields an aggregation path (dendrogram) on mixing measures. Along this path we prove a monotone strengthening of the loss (Lemma 1), obtain near-parametric finite-sample rates for the aggregated estimators together with height and likelihood control (Theorems 1 to 3 and Table 1), and derive a sweep-free dendrogram selection criterion (DSC) that is consistent and avoids multi-K training (Theorem 4 and Figures 1 and 3). Empirically, DSC is less prone to overfitting than AIC/BIC/ICL under ϵ-contamination due to its structural penalty on small heights (Figure 4), and it restores fast parameter rates after aggregation in over-specified SGMoE (Figure 2). To our knowledge this is the first method that couples finite-sample, fast-rate-aware merging with consistent model selection for SGMoE, avoiding multi-size training while preserving statistical efficiency.

SGMoE Setting. Let $(\mathbf{x}_n, y_n)_{n=1}^N$ be i.i.d. samples with $\mathbf{x}_n \in \mathbb{R}^D$ and $y_n \in \mathbb{R}$. Assume the data are

Table 1: Summary of density and parameter rates for SGMoE. The Voronoi cells $\mathbb{A}_j$ are defined in eq. (2). The function $\bar{r}(\cdot)$ is determined by solvability of the polynomial systems recalled in eq. (3) (e.g., $\bar{r}(2) = 4$, $\bar{r}(3) = 6$). The merged row and the fast pathwise rates correspond to the aggregation path described in Section 3.

Setting	**Loss**	$p_{G_0}(y \mid \boldsymbol{x})$	$\exp(\omega_{0k}^0)$	$\boldsymbol{\omega}_{1k}^0,\ b_k^0$	$\boldsymbol{a}_k^0,\ \sigma_k^0$
Exact-fit	$\mathrm{D_E}$	$\mathcal{O}\big((\log N/N)^{1/2}\big)$	$\mathcal{O}\big((\log N/N)^{1/2}\big)$	$\mathcal{O}\big((\log N/N)^{1/2}\big)$	$\mathcal{O}\big((\log N/N)^{1/2}\big)$
Over-fit	$\mathrm{D_O}$	$\mathcal{O}\big((\log N/N)^{1/2}\big)$	$\mathcal{O}\big((\log N/N)^{1/2}\big)$	$\mathcal{O}\big((\log N/N)^{1/2\bar{r}(\lvert\mathbb{A}_k\rvert)}\big)$	$\mathcal{O}\big((\log N/N)^{1/\bar{r}(\lvert\mathbb{A}_k\rvert)}\big)$
Merged	$\mathrm{D_{FRA}}$	$\mathcal{O}\big((\log N/N)^{1/2}\big)$	$\mathcal{O}\big((\log N/N)^{1/2}\big)$	$\mathcal{O}\big((\log N/N)^{1/2}\big)$	$\mathcal{O}\big((\log N/N)^{1/2}\big)$

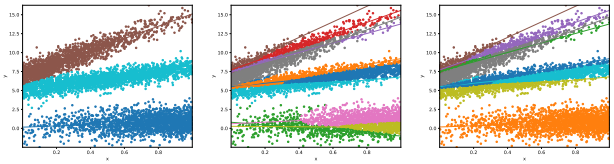
generated by a SGMoE model of order K_0, whose conditional density is

$$p_{G_0}(y \mid \boldsymbol{x}) := \sum_{k=1}^{K_0} \frac{\exp\big((\boldsymbol{\omega}_{1k}^0)^\top \boldsymbol{x} + \omega_{0k}^0\big)}{\sum_{j=1}^{K_0} \exp\big((\boldsymbol{\omega}_{1j}^0)^\top \boldsymbol{x} + \omega_{0j}^0\big)} \times \mathcal{N}\big(y \,\big|\, \boldsymbol{a}_k^{0\top} \boldsymbol{x} + b_k^0,\ \sigma_k^0\big). \tag{1}$$

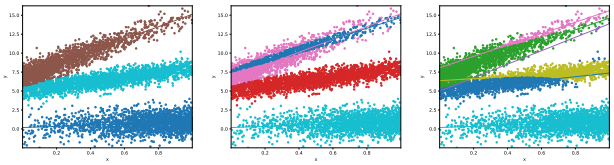
Each expert is Gaussian with mean $\boldsymbol{a}_k^{0\top}\boldsymbol{x} + b_k^0$ and variance $\sigma_k^0 > 0$. We encode parameters via the (not-necessarily normalized) mixing measure

$$G_0 \equiv G_0(K_0) \ := \ \sum_{k=1}^{K_0} \exp(\omega_{0k}^0)\, \delta_{(\boldsymbol{\omega}_{1k}^0,\, \boldsymbol{a}_k^0,\, b_k^0,\, \sigma_k^0)},$$

where $\boldsymbol{\eta}_k^0 := (\omega_{0k}^0, \boldsymbol{\omega}_{1k}^0, \boldsymbol{a}_k^0, b_k^0, \sigma_k^0) \in \boldsymbol{\Theta} \subset \mathbb{R} \times \mathbb{R}^D \times \mathbb{R}^D \times \mathbb{R} \times \mathbb{R}_{>0}$. Assume $\boldsymbol{\Theta}$ is compact and $\mathcal{X} \subset \mathbb{R}^D$, the support of $\mathbf{x}$, is bounded. Assume $\mathbf{x}$ has a continuous distribution so that the model is identifiable under this convention, a standard mild assumption; see Proposition 1 of Nguyen et al. (2023a).



(a) True regression (b) $K = 10$ (c) $K = 8$



(d) $K = 3$ (e) $K = 4$ (f) $K = 6$

Figure 1: Merging procedure from $K = 10$ to $K = 3$ of true mixing measure $G_0(3)$ with $K_0 = 3$ components, defined in eq. (11).

Maximum Likelihood Over At Most K Experts. When the true order K_0 is unknown, we estimate within $\mathcal{O}_K(\boldsymbol{\Theta}) := \Big\{G = \sum_{k=1}^{K'} \exp(\omega_{0k})\, \delta_{(\boldsymbol{\omega}_{1k},\, \boldsymbol{a}_k,\, b_k,\, \sigma_k)} : 1 \le K' \le K,\ (\omega_{0k}, \boldsymbol{\omega}_{1k}, \boldsymbol{a}_k, b_k, \sigma_k) \in \boldsymbol{\Theta}\Big\}$. We analyze the exactly specified case $K = K_0$, the over-specified case $K > K_0$, and the merging scheme using the following maximum likelihood estimator (MLE): $\widehat{G}_N \in \arg\max_{G \in \mathcal{O}_K(\boldsymbol{\Theta})} \frac{1}{N} \sum_{n=1}^N \log\big(p_G(y_n \mid \mathbf{x}_n)\big)$.

Practical Implication. Practitioners can fit a single over-specified SGMoE with moderate $K \ge K_0$, compute its aggregation path, and select $\widehat{K}$ via DSC. This single-fit workflow avoids grid sweeps over K, merges near-duplicate atoms to collapse slow directions within Voronoi cells, accelerates parameter convergence, and often recovers the correct expert count even under mild contamination. Dendrogram heights provide a transparent structural summary.

Paper Organization. Section 2 states the unified parameter-rate result and the algebraic exponents $\bar{r}(\cdot)$. Section 3 introduces the fast-rate-aware distance, merge operator, aggregation path, fast pathwise rates, and DSC. Section 4 illustrates parameter rates, path behaviour, model selection under clean and contaminated regimes, and a real-data application to maize drought-response traits in Section 5. Then, we offer concluding remarks, limitations, and future work in Section 6. Proof sketches appear at the end of Section 3, with full proofs deferred to the appendix. Additional biological background, preprocessing details for the maize dataset, and further geometric and technical discussion are provided in the supplementary material.

Notation. Throughout the paper, for any natural number $N \in \mathbb{N}$ we abbreviate $\{1, 2, \ldots, N\}$ by $[N]$. Given two sequences of positive real numbers $\{a_N\}_{N=1}^\infty$ and $\{b_N\}_{N=1}^\infty$, we write $a_N = \mathcal{O}(b_N)$ (equivalently, $a_N \lesssim b_N$) to mean that there exists a constant $C > 0$ such that $a_N \le C\, b_N$ for all $N \in \mathbb{N}$. For a vector $\boldsymbol{v} \in \mathbb{R}^D$, set $|\boldsymbol{v}| := v_1 + \cdots + v_D$, and let $\|\boldsymbol{v}\|_p$ denote its p-norm; by default, $\|\boldsymbol{v}\|$ refers to the 2-norm unless otherwise stated. We also use $\|\boldsymbol{A}\|$ for the Frobenius norm of a matrix $\boldsymbol{A} \in \mathbb{R}^{D \times D}$. For any set $\mathbb{S}$, $|\mathbb{S}|$ denotes its cardinality. Finally, for two probability density functions p and q with respect to the Lebesgue measure μ, define $\mathrm{D_{TV}}(p, q) := \frac{1}{2} \int |p - q| d\boldsymbol{\mu}$ as their Total

Variation distance, while $\mathrm{D}_{\mathrm{h}}^2(p,q) := \frac{1}{2}\int(\sqrt{p}-\sqrt{q})^2 d\boldsymbol{\mu}$ denotes the squared Hellinger distance between them. Let $\boldsymbol{\Theta}$ be the parameter space. Write $\mathcal{E}_K(\boldsymbol{\Theta})$ for the collection of discrete probability measures on $\boldsymbol{\Theta}$ with exactly K atoms, and $\mathcal{O}_K(\boldsymbol{\Theta}) := \bigcup_{K'\leq K}\mathcal{E}_{K'}(\boldsymbol{\Theta})$ for those with at most K atoms. For a mixing measure $G = \sum_{k=1}^K \pi_k\,\delta_{\boldsymbol{\theta}_k}$, we (slightly abusively) refer to each component $\pi_k\,\delta_{\boldsymbol{\theta}_k}$ as an "atom," comprising both its weight π_k and parameter $\boldsymbol{\theta}_k$. When clear from context, we drop $\boldsymbol{\Theta}$ and simply write $\mathcal{E}_K$ and $\mathcal{O}_K$.

2 PRELIMINARIES

We present a unified result for the parameter estimation rate of the MLE in the SGMoE that simultaneously covers the exact-specified case ($K = K_0$) and the over-specified case ($K > K_0$), building on Nguyen et al. (2023a).

Voronoi Cells. For a candidate mixing measure $G = \sum_{k=1}^K \exp(\omega_{0k})\,\delta_{(\boldsymbol{\omega}_{1k},\boldsymbol{a}_k,b_k,\sigma_k)}$ and the true $G_0 = \sum_{k=1}^{K_0}\exp(\omega_{0k}^0)\,\delta_{(\boldsymbol{\omega}_{1k}^0,\boldsymbol{a}_k^0,b_k^0,\sigma_k^0)}$, define for $k\in[K_0]$:

$$\mathbb{A}_k(G) := \{\ell\in[K] : \|\boldsymbol{\theta}_\ell - \boldsymbol{\theta}_k^0\| \leq \|\boldsymbol{\theta}_\ell - \boldsymbol{\theta}_j^0\|,\ \forall j\neq k\}, \tag{2}$$

where we denote $\boldsymbol{\theta}_\ell := (\boldsymbol{\omega}_{1\ell},\boldsymbol{a}_\ell,b_\ell,\sigma_\ell)$. We use the softmax-translation $(t_0,\boldsymbol{t}_1)$ from identifiability (cf. Proposition 1 of Nguyen et al., 2023a) and the shorthand $\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k} := \boldsymbol{\omega}_{1\ell} - \boldsymbol{\omega}_{1k}^0 - \boldsymbol{t}_1$, $\Delta\boldsymbol{a}_{\ell k} := \boldsymbol{a}_\ell - \boldsymbol{a}_k^0$, $\Delta b_{\ell k} := b_\ell - b_k^0$, $\Delta\sigma_{\ell k} := \sigma_\ell - \sigma_k^0$. For notational simplicity, we write $\mathbb{A}_k$ instead of $\mathbb{A}_k(G)$.

Algebraic Obstruction and Exponents. For $M\geq 2$, let $\bar{r}(M)$ be the smallest integer r determined by the polynomial system as follows: given $0\leq|\boldsymbol{\ell}_1|\leq r$, $0\leq\ell_2\leq r-|\boldsymbol{\ell}_1|$, $|\boldsymbol{\ell}_1|+\ell_2\geq 1$, the polynomial system

$$\sum_{j=1}^M \sum_{(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\alpha_3,\alpha_4)\in\mathbb{I}_{\boldsymbol{\ell}_1,\ell_2}} \frac{p_{5j}^2\,p_{1j}^{\boldsymbol{\alpha}_1}p_{2j}^{\boldsymbol{\alpha}_2}p_{3j}^{\alpha_3}p_{4j}^{\alpha_4}}{\boldsymbol{\alpha}_1!\,\boldsymbol{\alpha}_2!\,\alpha_3!\,\alpha_4!} = 0, \tag{3}$$

admits no non-trivial solution (all $p_{5j}\neq 0$ and at least one $p_{3j}\neq 0$). The ranges of $\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\alpha_3,\alpha_4$ in the above sum satisfy $\mathbb{I}_{\boldsymbol{\ell}_1,\ell_2} = \{\boldsymbol{\alpha}=(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\alpha_3,\alpha_4)\in\mathbb{N}^D\times\mathbb{N}^D\times\mathbb{N}\times\mathbb{N} :\ \boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2=\boldsymbol{\ell}_1,\ |\boldsymbol{\alpha}_2|+\alpha_3+2\alpha_4=\ell_2\}$. For general dimension D and parameter $M\geq 2$, finding the exact value of $\bar{r}(M)$ is a non-trivial central problem in algebraic geometry (Sturmfels, 2002). Known values:

Fact 1 (Nguyen et al., 2023a, Lemma 1)**.** *For any $D\geq 1$: $\bar{r}(2)=4$, $\bar{r}(3)=6$, and $\bar{r}(M)\geq 7$ for $M\geq 4$.*

Classical Overfit-Aware Voronoi Distance. Define a single loss that reduces to the exact-fit metric when each cell has one atom, and adds over-fit penalties otherwise:

$$\begin{aligned}\mathrm{D}_{\mathrm{O}}(G,G_0) &:= \mathrm{D}_{\mathrm{E}}(G,G_0)\\ &+ \inf_{t_0,\boldsymbol{t}_1}\sum_{k:|\mathbb{A}_k|>1}\sum_{\ell\in\mathbb{A}_k}\exp(\omega_{0\ell})\Big(\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k},\Delta b_{\ell k})\|^{\bar{r}(|\mathbb{A}_k|)}\\ &\qquad + \|(\Delta\boldsymbol{a}_{\ell k},\Delta\sigma_{\ell k})\|^{\bar{r}(|\mathbb{A}_k|)/2}\Big), \qquad (4)\\ \mathrm{D}_{\mathrm{E}}(G,G_0) &:= \inf_{t_0,\boldsymbol{t}_1}\sum_{k=1}^{K_0}\Big|\sum_{\ell\in\mathbb{A}_k}\exp(\omega_{0\ell}) - \exp(\omega_{0k}^0+t_0)\Big|\\ &+ \sum_{k:|\mathbb{A}_k|=1}\sum_{\ell\in\mathbb{A}_k}\exp(\omega_{0\ell})\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k},\Delta\boldsymbol{a}_{\ell k},\Delta b_{\ell k},\Delta\sigma_{\ell k})\|.\end{aligned}$$

When $|\mathbb{A}_k|=1$ for all k (i.e., $K=K_0$), eq. (4) equals the exact-fit metric D_{E}; if some $|\mathbb{A}_k|>1$ (i.e., $K>K_0$), eq. (4) adds the higher-order penalties determined by $\bar{r}(\cdot)$.

Fact 2 (Nguyen et al., 2023a, Theorems 1 and 2)**.** *There exist universal constants $C,c>0$ (depending only on G_0 and $\boldsymbol{\Theta}$) s.t. the MLE $\widehat{G}_N$ of order $K\geq K_0$ satisfies*

$$\mathbb{P}\Big(\mathrm{D}_{\mathrm{O}}(\widehat{G}_N,G_0) > C\,(\log N/N)^{1/2}\Big) \lesssim e^{-c\log N}. \tag{5}$$

Remarks. (i) If $K=K_0$ (all $|\mathbb{A}_k|=1$), then $\mathrm{D}_{\mathrm{O}}=\mathrm{D}_{\mathrm{E}}$ and eq. (5) yields the exact-specified rate $\mathbb{P}\big(\mathrm{D}_{\mathrm{E}}(\widehat{G}_N,G_0) > C(\log N/N)^{1/2}\big)\lesssim e^{-c\log N}$, implying parametric ($N^{-1/2}$ up to logs) estimation of $\exp(\omega_{0k}^0)$, $\boldsymbol{\omega}_{1k}^0$ (up to translation), $\boldsymbol{a}_k^0$, b_k^0, σ_k^0 for all $k\in[K_0]$. (ii) If $K>K_0$ (some $|\mathbb{A}_k|>1$), the same bound holds for D_{O}, while the exponents $\bar{r}(|\mathbb{A}_k|)$ inside eq. (4) encode the slower algebraic behavior of over-covered parameters within each Voronoi cell.

3 FAST-RATE-AWARE EXPERT AGGREGATION IN SGMOE

3.1 Why Merge Experts? The Rate Gap

Building on Section 2, identifiability and the unified parameter-rate bound (Fact 2) imply that converting density accuracy into parameter accuracy hinges on a suitable inverse (loss) inequality. When the model is over-specified ($K>K_0$), several fitted atoms may fall into the same Voronoi cell $\mathbb{A}_k$ (defined in eq. (2)), which induces a *rate gap*: single-covered truths achieve (near) parametric rates, whereas multi-covered truths converge more slowly with exponents governed by $\bar{r}(|\mathbb{A}_k|)$ from Section 2. To exploit this, we (i) refine the loss to expose mergeable structure, and (ii) aggregate (merge) near-duplicate atoms to recover fast rates and guide model order selection.

3.2 A Fast-Rate-Aware Voronoi Distance

Our Proposal. Let $\mathrm{D}_{\mathrm{O}}(G, G_0)$ denote the over-fit Voronoi loss from eq. (4) and $\mathbb{A}_k$ be as in eq. (2). We augment it with first-order "merged-moment" couplings inside multi-covered cells to obtain

$$\begin{aligned}\mathrm{D}_{\mathrm{FRA}}(G, G_0) &:= \mathrm{D}_{\mathrm{O}}(G, G_0) \\ &+ \inf_{t_0, \boldsymbol{t}_1} \sum_{k:|\mathbb{A}_k|>1} \Bigg(\Big\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell})(\Delta b_{\ell k}) \Big\| \\ &\quad + \Big\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell})(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}) \Big\| \\ &\quad + \Big\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell})[(\Delta b_{\ell k})^2 + (\Delta \sigma_{\ell k})] \Big\| \\ &\quad + \Big\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell})[(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k})(\Delta b_{\ell k}) + (\Delta \boldsymbol{a}_{\ell k})] \Big\| \\ &\quad + \Big\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell})(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k})(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k})^\top \Big\| \Bigg). \end{aligned} \tag{6}$$

Link to Section 2. The penalties inside eq. (6) are consistent with the exponents $\bar{r}(|\mathbb{A}_k|)$ that appear in the unified loss eq. (4): when $|\mathbb{A}_k| = 1$, $\mathrm{D}_{\mathrm{FRA}}$ reduces to the exact-fit metric D_{E}; when $|\mathbb{A}_k| > 1$, the added block-sums control the slow directions and quantify how well the cell behaves *as if merged*.

Motivation for Merging. Because the slow rates originate from multiple atoms sharing a cell, replacing these atoms by their softmax-weighted aggregate collapses the problematic directions and restores first-order (parametric) behavior for the merged parameters. Thus, $\mathrm{D}_{\mathrm{FRA}}$ both (i) certifies where merging is beneficial (large intra-cell terms) and (ii) predicts the rate improvement obtained by aggregation, which we leverage next for hierarchical merging and model selection.

3.3 A Merge Operator Tailored to SGMoE

Connection to Section 2 and Novelty. The unified rate result in Section 2 shows that parameter convergence hinges on how fitted atoms distribute across Voronoi cells; multi-covered cells induce slower algebraic behavior governed by $\bar{r}(\cdot)$. The merge operator below is the first ingredient of our contribution: it *operationalizes* that insight by collapsing near-duplicate atoms within a cell using softmax-weighted updates. This turns slow, multi-component directions into a single, first-order direction, setting up our fast pathwise rates (Theorem 1) and height/likelihood controls (Theorems 2 and 3).

Rate-Weighted Dissimilarity. For $G^{(K)} = \sum_{k=1}^{K} \exp(\omega_{0k}) \delta_{\boldsymbol{\theta}_k}$ with $\boldsymbol{\theta}_k = (\boldsymbol{\omega}_{1k}, \boldsymbol{a}_k, b_k, \sigma_k)$, define

$$\begin{aligned}&\mathsf{d}\big(\exp(\omega_{0\ell_1})\delta_{\boldsymbol{\theta}_{\ell_1}}, \exp(\omega_{0\ell_2})\delta_{\boldsymbol{\theta}_{\ell_2}}\big) \\ &:= \frac{\exp(\omega_{0\ell_1} + \omega_{0\ell_2})}{\exp(\omega_{0\ell_1}) + \exp(\omega_{0\ell_2})} \|(\boldsymbol{\omega}_{1\ell_1}, b_{\ell_1}) - (\boldsymbol{\omega}_{1\ell_2}, b_{\ell_2})\|^2 \\ &+ \frac{\exp(\omega_{0\ell_1} + \omega_{0\ell_2})}{\exp(\omega_{0\ell_1}) + \exp(\omega_{0\ell_2})} \|(\boldsymbol{a}_{\ell_1}, \sigma_{\ell_1}) - (\boldsymbol{a}_{\ell_2}, \sigma_{\ell_2})\|. \end{aligned} \tag{7}$$

Pick $(i, j) = \arg\min_{\ell_1 \neq \ell_2 \in [K]} \mathsf{d}(\cdot, \cdot)$ and replace the pair by the *softmax-weighted* aggregate

$$\begin{aligned}\omega_{0*} &= \log\left(\exp \omega_{0i} + \exp \omega_{0j}\right), \\ \boldsymbol{\omega}_{1*} &= \exp\left(\omega_{0i} - \omega_{0*}\right)\boldsymbol{\omega}_{1i} + \exp\left(\omega_{0j} - \omega_{0*}\right)\boldsymbol{\omega}_{1j}, \\ b_* &= \exp\left(\omega_{0i} - \omega_{0*}\right) b_i + \exp\left(\omega_{0j} - \omega_{0*}\right) b_j, \\ \boldsymbol{a}_* &= \frac{\exp(\omega_{0i})}{\exp(\omega_{0*})}[(\boldsymbol{\omega}_{1i} - \boldsymbol{\omega}_{1*})(b_i - b_*) + \boldsymbol{a}_i] \\ &\quad + \frac{\exp(\omega_{0j})}{\exp(\omega_{0*})}[(\boldsymbol{\omega}_{1j} - \boldsymbol{\omega}_{1*})(b_j - b_*) + \boldsymbol{a}_j], \\ \sigma_* &= \frac{\exp(\omega_{0i})}{\exp(\omega_{0*})}[(b_i - b_*)^2 + \sigma_i] \\ &\quad + \frac{\exp(\omega_{0j})}{\exp(\omega_{0*})}[(b_j - b_*)^2 + \sigma_j]. \end{aligned} \tag{8}$$

Then we define $G^{(K-1)} = \exp(\omega_{0*})\delta_{(\boldsymbol{\omega}_{1*}, \boldsymbol{a}_*, b_*, \sigma_*)} + \sum_{k \neq i,j} \exp(\omega_{0k})\delta_{(\boldsymbol{\omega}_{1k}, \boldsymbol{a}_k, b_k, \sigma_k)}$. A description of the whole procedure can be seen in Algorithm 1. The choice of merging atoms and deriving the new atom (eqs. (7) and (8)) are in particular faithful to hierarchical clustering and K-means algorithms.

3.4 The Hierarchical View of Aggregation Path

Transition and Main Idea. The merge step converts local redundancy into a single effective atom. Repeating it induces a *global* hierarchy, the aggregation path, along which our new analysis proves a *monotone strengthening* of the loss and, crucially, *fast* convergence at every level. This bridges Section 2 (unified loss but slow rates) with a constructive, data-driven path that achieves the same near-parametric behavior after aggregation.

Having presented the algorithm to choose and merge a mixing measure with K atoms to $K-1$ atoms, we now describe the dendrogram (hierarchical aggregation) of G that emerges by repeatedly applying the merging procedure.

Dendrogram (Hierarchical Aggregation). Iterate the merge in eqs. (7) and (8) from $\kappa = K$ down to 2, generating $\{G^{(\kappa)}\}_{\kappa=2}^{K}$. Define the dendrogram $\mathcal{T}(G) = (\mathcal{V}, \mathcal{E}, \mathcal{H})$ with $\mathcal{V}$ containing K levels, the κ-th level holding the atoms of $G^{(\kappa)}$, $\mathcal{E}$

storing the links between merged pairs across adjacent levels, and $\mathcal{H} = (\mathsf{h}^{(K)}, \ldots, \mathsf{h}^{(2)})$ with $\mathsf{h}^{(\kappa)} := \min\{\mathsf{d}(\cdot,\cdot)$ over pairs in $G^{(\kappa)}\}$. The quantity $\mathsf{h}^{(\kappa)}$ is the height between levels κ and $\kappa-1$.

When we represent $\mathcal{T}(G)$ on a graph, $\mathsf{h}^{(\kappa)}$ is the height between κ-th level and $(\kappa-1)$-th level. The procedure to construct the dendrogram of G is given by Algorithm 2.

Algorithm 1 SGMoE Merge Step (Fast-Rate-Aware)

Require: $G^{(\kappa)} = \sum_{k=1}^{\kappa} \exp(\omega_{0k})\delta_{(\boldsymbol{\omega}_{1k}, \boldsymbol{a}_k, b_k, \sigma_k)}$
1: $(i,j) \leftarrow \arg\min\{\mathsf{d}\big(\exp(\omega_{0\ell_1})\delta_{\boldsymbol{\theta}_{\ell_1}}, \exp(\omega_{0\ell_2})\delta_{\boldsymbol{\theta}_{\ell_2}}\big) : \ell_1 \neq \ell_2 \in [\kappa]\}$
2: Compute $(\omega_{0*}, \boldsymbol{\omega}_{1*}, \boldsymbol{a}_*, b_*, \sigma_*)$ by eq. (8)
3: **return** $G^{(\kappa-1)} = \exp(\omega_{0*})\delta_{(\boldsymbol{\omega}_{1*}, \boldsymbol{a}_*, b_*, \sigma_*)} + \sum_{k\neq i,j} \exp(\omega_{0k})\delta_{\boldsymbol{\theta}_k}$

Algorithm 2 SGMoE Hierarchical Aggregation Path

Require: $G^{(K)} = \sum_{k=1}^{K} \exp(\omega_{0k})\delta_{(\boldsymbol{\omega}_{1k}, \boldsymbol{a}_k, b_k, \sigma_k)}$
1: Initialize $\mathcal{T}(G) = (\mathcal{V}, \mathcal{E}, \mathcal{H})$ with $\mathcal{V}_K = \{\text{atoms of } G^{(K)}\}$, $\mathcal{E} = \varnothing$
2: **for** $\kappa = K, \ldots, 2$ **do**
3: $G^{(\kappa-1)} \leftarrow$ Algorithm 1$(G^{(\kappa)})$
4: Append atoms of $G^{(\kappa-1)}$ to level $\mathcal{V}_{\kappa-1}$, link merged pair in $\mathcal{E}$
5: $\mathsf{h}^{(\kappa)} \leftarrow \min \mathsf{d}(\cdot,\cdot)$ over pairs in $G^{(\kappa)}$; append to $\mathcal{H}$
6: **end for**
7: **return** $\mathcal{T}(G) = (\mathcal{V}, \mathcal{E}, \mathcal{H})$ and $\{G^{(\kappa)}\}_{\kappa=1}^{K}$

Monotone Strengthening of the Loss (Bridge to Fast Rates). The following lemma formalizes that each merge step cannot increase our fast-rate-aware distance to G_0, making the path progressively *easier* to estimate:

Lemma 1. *As* $\mathrm{D}_{\mathrm{FRA}}(G^{(K)}, G_0) \to 0$, $\mathrm{D}_{\mathrm{FRA}}(G^{(K)}, G_0) \gtrsim \mathrm{D}_{\mathrm{FRA}}(G^{(K-1)}, G_0) \gtrsim \cdots \gtrsim \mathrm{D}_{\mathrm{FRA}}(G^{(K_0)}, G_0)$, *with constants depending only on* G_0, $\boldsymbol{\Theta}$, *and* K.

Behavior of the Path for the MLE (Main Fast-Rate Theorem). Leveraging the monotonicity above together with the unified inverse bound from Section 2, we obtain fast rates *at every level* of the path, including the exact-fit and under-fit levels where aggregation recovers optimal parametric rate behavior:

Theorem 1 (Fast convergence rates along the path)**.** *There exist universal constants* $C_1', c_1, C_2', c_2 > 0$ *such that for all* $\kappa \in [K_0+1, K]$ *and* $\kappa' \in [K_0]$, *we have*

$$\mathbb{P}\big(\mathrm{D}_{\mathrm{FRA}}(\widehat{G}_N^{(\kappa)}, G_0) > C_2'(\log N/N)^{1/2}\big) \lesssim e^{-c_2 \log N},$$
$$\mathbb{P}\big(\mathrm{D}_{\mathrm{E}}(\widehat{G}_N^{(\kappa')}, G_0^{(\kappa')}) > C_1'(\log N/N)^{1/2}\big) \lesssim e^{-c_1 \log N}. \tag{9}$$

3.5 Heights and Likelihood Along the Path

Transition from Structure to Statistics. Heights summarize structural redundancy; likelihood captures statistical fit. Our second set of novel guarantees shows (i) heights shrink at a rate dictated by $\bar{r}(\widehat{G}_N) := \max_{k\in[K_0]} \bar{r}(|\mathbb{A}_k(\widehat{G}_N)|)$, and (ii) the empirical likelihood concentrates to its population counterpart along the path.

Height Definitions. For all $\kappa \in [K_0+1, K]$ and $\kappa' \in [K_0]$, let

$$\mathsf{h}_N^{(\kappa)} := \min\Big\{\mathsf{d}\big(\exp(\widehat{\omega}_{0k_1})\delta_{\widehat{\boldsymbol{\theta}}_{k_1}}, \exp(\widehat{\omega}_{0k_2})\delta_{\widehat{\boldsymbol{\theta}}_{k_2}}\big) : k_1 \neq k_2, \text{ atoms of } \widehat{G}_N^{(\kappa)}\Big\}, \tag{10}$$

and let $\mathsf{h}_0^{(\kappa')}$ be the analogous height on the true path. Then:

Theorem 2 (Height control)**.** *For all* $\kappa \in [K_0+1, K]$ *and* $\kappa' \in [K_0]$, $\mathsf{h}_N^{(\kappa)} \lesssim (\log N/N)^{1/\bar{r}(\widehat{G}_N)}$, *and*

$$\big|\mathsf{h}_N^{(\kappa')} - \mathsf{h}_0^{(\kappa')}\big| \lesssim (\log N/N)^{1/2},$$

with constants depending only on G_0, $\boldsymbol{\Theta}$, *and* κ.

Likelihood. We define empirical average log-likelihood and population average log-likelihood as follow: $\bar{\ell}_N(p_G) := N^{-1}\sum_{n=1}^{N} \log p_G(y_n|\mathbf{x}_n)$ and $\mathcal{L}(p_G) := \mathbb{E}_{(\mathbf{x},y)\sim P_{G_0}}[\log p_G(y|\boldsymbol{x})]$.

Condition K. There exist positive constants c_α and c_β such that for all sufficiently small ϵ and $\boldsymbol{\theta}_0, \boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \epsilon$, we have $\log f(\boldsymbol{x}, y|\boldsymbol{\theta}) \geq (1 + c_\beta\epsilon)\log f(\boldsymbol{x}, y|\boldsymbol{\theta}_0) - c_\alpha\epsilon$.

Theorem 3 (Likelihood concentration on the path)**.** *Assume* ***Condition K*** *hold. Then, for any* $\kappa \in [K_0+1, K]$, $\big|\bar{\ell}_N(p_{\widehat{G}_N^{(\kappa)}}) - \mathcal{L}(p_{G_0})\big| \lesssim (\log N/N)^{1/(2\bar{r}(\widehat{G}_N))}$. *Moreover, for* $\kappa' \in [K_0]$, $\bar{\ell}_N(p_{\widehat{G}_N^{(\kappa')}}) \to \mathcal{L}(p_{G_0^{(\kappa')}})$ *in* $\mathbb{P}_{G_0}$*-probability as* $N \to \infty$.

3.6 Choosing the Number of Experts via a Height-Likelihood Rule

Novel Model Selection Principle. By combining structural signal (heights) and statistical fit (likelihood), our DSC favors models that are both well-separated and well-supported by the data, unlike AIC/BIC/ICL, which ignore the geometry of the fitted atoms.

DSC Definition. For each level κ, define

$$\mathrm{DSC}_N^{(\kappa)} := -\Big(\mathsf{h}_N^{(\kappa)} + \epsilon_N\, \bar{\ell}_N\big(p_{\widehat{G}_N^{(\kappa)}}\big)\Big),$$

where the weight ϵ_N satifies $1 \ll \epsilon_N \ll (N/\log N)^{1/(2\bar{r}(\widehat{G}_N))}$. A practical choice is $\epsilon_N := \log N$. Select

$$\widehat{K}_N := \arg\min_{\kappa\in[2,K]} \mathrm{DSC}_N^{(\kappa)}.$$

Theorem 4 (Consistency of model selection)**.** *Assume that data are generated by a softmax-gated Gaussian MoE, the parameter space $\boldsymbol{\Theta}$ is compact, the covariate support $\mathcal{X} \subset \mathbb{R}^D$ is bounded, the DSC uses a penalty ϵ_N satisfying as above, and the true component $K_0 \geq 2$. Then $\widehat{K}_N \to K_0$ in $\mathbb{P}_{G_0}$-probability as $N \to \infty$.*

Interpretation. Unlike pure likelihood criteria (AIC/BIC/ICL), $\mathrm{DSC}_N^{(\kappa)}$ also penalizes *structural closeness* through $\mathsf{h}_N^{(\kappa)}$. Small heights indicate either redundant atoms (near-duplicates) or atoms with tiny softmax weights; both are symptomatic of over-specification. The joint use of heights and likelihood therefore yields a more robust selection rule in SGMoE.

3.7 Proof Sketches

We sketch the proofs of Lemma 1 and Theorems 1 to 4, which together establish monotonicity along the dendrogram path and consistency of the dendrogram-based model selection. We first motivate the fast-rate-aware Voronoi distance in eq. (6). When $\widehat{G}_N \to G_0$, over-specification yields Voronoi cells with $|\mathbb{A}_k^N| > 1$. Repeatedly merging such atoms eventually makes every cell singleton, which motivates our construction. Using the density decomposition

$$Q_N = \Big[\sum_{k=1}^{K_0} \exp\left((\boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1)^\top x + \omega_{0k}^0 + t_0\right)\Big] \times \big[p_{G_N}(y|\boldsymbol{x}) - p_{G_0}(y|\boldsymbol{x})\big],$$

we analyze the sums over indices with $|\mathbb{A}_k^N| > 1$ under $1 \leq |\boldsymbol{\ell}_1| + \ell_2 \leq 2\bar{r}(|\mathbb{A}_k^N|)$. For clarity, we also consider $(\boldsymbol{\ell}_1, \ell_2)$ with $1 \leq |\boldsymbol{\ell}_1| + \ell_2 \leq 2$, which corresponds to $|\mathbb{A}_k^N| = 1$. This reasoning leads to the merging algorithm.

Proof Sketch of Lemma 1. Proceed by induction on $\kappa \in [K_0, K]$ and justify $\mathrm{D}_{\mathrm{FRA}}(G^{(K)}, G_0) \gtrsim \mathrm{D}_{\mathrm{FRA}}(G^{(K-1)}, G_0)$. As $\mathrm{D}_{\mathrm{FRA}}(G^{(K)}, G_0) \to 0$, extract a sequence that satisfies $(\boldsymbol{a}_\ell^N, b_\ell^N, \sigma_\ell^N) \to (\boldsymbol{a}_k^0, b_k^0, \sigma_k^0)$ and there exist $t_0 \in \mathbb{R}$, $\boldsymbol{t}_1 \in \mathbb{R}^D$ with $\sum_{\ell \in \mathbb{A}_k^N} \exp(\omega_{0\ell}^N) \to \exp(\omega_{0k}^0 + t_0)$ and $\boldsymbol{\omega}_{1\ell}^N \to \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1$ for all $\ell \in \mathbb{A}_k^N$. The minimizing pair (ℓ_1, ℓ_2) must belong to a common $\mathbb{A}_k^N$. Using eq. (8) and Jensen's inequality for the convex maps $z \mapsto \|z\|^{\bar{r}_k}$ and $z \mapsto \|z\|^{\bar{r}_k/2}$, it suffices to show

$$(\exp\omega_{0\ell_1}^N + \exp\omega_{0\ell_2}^N)\,\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1*k}^N, \Delta b_{*k}^N)\|^{\bar{r}_k} \lesssim \sum_{j \in \{\ell_1, \ell_2\}} \exp\omega_{0j}^N\,\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1jk}^N, \Delta b_{jk}^N)\|^{\bar{r}_k},$$

$$(\exp\omega_{0\ell_1}^N + \exp\omega_{0\ell_2}^N)\,\|(\Delta\boldsymbol{a}_{*k}^N, \Delta\sigma_{*k}^N)\|^{\bar{r}_k/2} \lesssim \sum_{j \in \{\ell_1, \ell_2\}} \exp\omega_{0j}^N\,\|(\Delta\boldsymbol{a}_{jk}^N, \Delta\sigma_{jk}^N)\|^{\bar{r}_k/2},$$

which yields the desired monotonicity.

Proof Sketch of Theorem 1. Combine Lemma 1 with an inverse bound for $\mathrm{D}_{\mathrm{FRA}}(\widehat{G}_N, G_0)$. Following Nguyen et al., 2023a, establish

$$\mathbb{E}_{\mathbf{x}}\big[\,\mathrm{D}_{\mathrm{TV}}(p_G(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x}))\big] \gtrsim \mathrm{D}_{\mathrm{FRA}}(G, G_0),$$

and use Proposition 2 in Nguyen et al., 2023a,

$$\mathbb{E}_{\mathbf{x}}[\mathrm{D}_{\mathrm{h}}^2(p_{\widehat{G}_N}(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x}))] = \mathcal{O}_{\mathbb{P}}((\log N/N)^{1/2}),$$

to derive the rate for $\widehat{G}_N$. Apply Lemma 1 to obtain the bounds for $\kappa \in [K_0 + 1, K]$. For $\kappa' \in [K_0]$, combine the previous rate with the merging formula to conclude.

Proof Sketch of Theorem 2. Use Theorem 1 and the fact that any merged pair lies in the same Voronoi cell. Inequalities analogous to those in Lemma 1 and Theorem 1 translate parameter rates into height bounds.

Proof Sketch of Theorem 3. Consider three cases. If $\kappa \geq K_0$, invoke empirical process tools (van de Geer, 2000) and comparisons between Hellinger and Wasserstein distances (Chen, 1995; Villani, 2003, 2009). If $\kappa = K_0$, combine Theorem 1 with verification that $u(y|\boldsymbol{x}; \boldsymbol{\omega}_1, \boldsymbol{a}, b, \sigma) := \exp(\boldsymbol{\omega}_1^\top \boldsymbol{x})\mathcal{N}(y|\boldsymbol{a}^\top \boldsymbol{x} + b, \sigma)$ satisfies **Condition K**. If $\kappa < K_0$, conclude via standard convergence arguments.

Finally, Theorem 4 follows from Theorems 2 and 3.

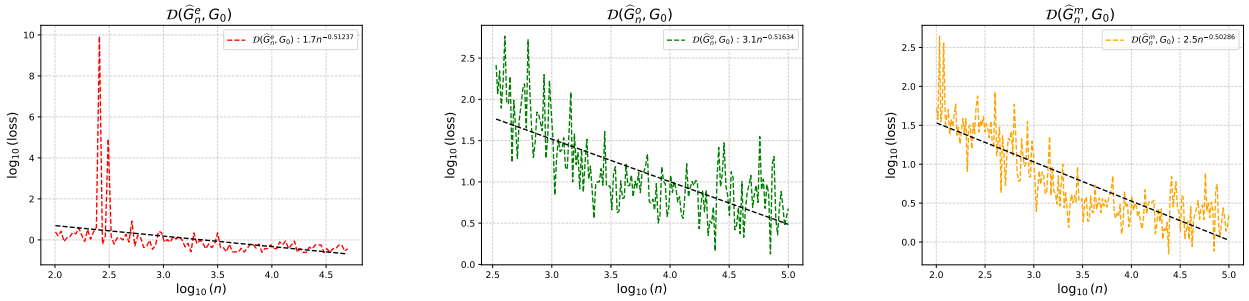
4 SIMULATION STUDIES

We first show that the dendrogram-based merge yields fast convergence of the mixing measure: starting from an over-fitted estimator that converges slowly, the merged estimator approaches the truth quickly. We then assess model selection via DSC against AIC, BIC, and ICL. Unlike these single-shot selectors, the dendrogram offers a hierarchical view of the fitted atoms, clarifying redundancy and structure. All simulations were run in Python 3.12 on a standard Unix-based system.

Numerical Schemes. The ground-truth mixing measure is

$$G_0 \equiv G_0(2) := \sum_{k=1}^{2} \exp(\omega_{0k}^0)\,\delta_{(\boldsymbol{\omega}_{1k}^0, \boldsymbol{a}_k, \boldsymbol{b}_k, \sigma_k)} = \exp(-8)\,\delta_{(25,\,-20,\,15,\,0.3)} + \exp(0)\,\delta_{(0,\,20,\,-5,\,0.4)}.$$

For each experiment, N varies on a logarithmic grid from $\log_{10}(N_{\min})$ to $\log_{10}(N_{\max})$, yielding N_{num} sizes in $[N_{\min}, N_{\max}]$. At each N, we generate N_{rep} datasets from G_0 and compute the exact-fitted MLE $\widehat{G}_N^e \in \mathcal{E}_2$ and the over-fitted MLE $\widehat{G}_N^o \in \mathcal{O}_4$ $(K = 4)$ using an EM variant of Chamroukhi et al. (2009). EM stops at tolerance $\epsilon = 10^{-6}$ or 2000 iterations. Because the



(a) Exact-fitted $\hat{G}^e_n$. (b) Over-fitted $\hat{G}^o_n$. (c) Merged $\hat{G}^m_n$.

Figure 2: Convergence under three settings: (a) exact-fitted, (b) over-fitted, and (c) merged mixing measures.

softmax gate in eq. (1) is translation-invariant, we fix a baseline by setting $\omega^0_{0K_0} = 0$ and $\boldsymbol{\omega}^0_{1K_0} = 0$.

To stabilize estimation and highlight asymptotics, EM is favorably initialized. For each replication and (K, K_0), split $[K]$ into K_0 disjoint sets $\mathbb{S}_1, \ldots, \mathbb{S}_{K_0}$, each nonempty. For $k \in \mathbb{S}_t$, draw $\boldsymbol{\eta}^0_k = (\omega^0_{0k}, \boldsymbol{\omega}^0_{1k}, \boldsymbol{a}^0_k, b^0_k, \sigma^0_k)$ from a Gaussian centered at $\boldsymbol{\eta}^0_t = (\omega^0_{0t}, \boldsymbol{\omega}^0_{1t}, \boldsymbol{a}^0_t, b^0_t, \sigma^0_t)$ with small covariance. After estimating $\widehat{G}^o_N$, apply the merging procedure in Algorithm 2 to obtain $\widehat{G}^m_N \in \mathcal{E}_2$.

Fast Parameter Estimation via the Dendrogram. We measure accuracy with the Voronoi distance in eq. (6). For the exact-fitted setting, we use 30 replicates over 100 sample sizes with $N \in [10^2, 5\times10^4]$; for the over-fitted setting, 40 replicates over 165 sizes with $N \in [338, 10^5]$; for the merged estimator, 40 replicates over 200 sizes with $N \in [10^2, 10^5]$. The average loss and a reference slope $N^{-1/2}$ are shown in Figure 2. Results match Theorem 1: the exact-fitted and merged estimators attain the optimal $N^{-1/2}$ rate toward G_0, while merging drives the over-fitted estimator to the exact-fit level. For illustration of Algorithm 2, Figure 1 considers $G_0(3)$ as follows:

$$e^{-2}\,\delta_{(3,\,1,\,0,\,1)} + e^{1}\,\delta_{(-3.5,\,8,\,7,\,0.8)} + e^{0}\,\delta_{(0,\,3,\,5,\,0.6)}. \qquad (11)$$

Model Selection with DSC. We compare DSC to AIC, BIC, and ICL over 32 sample sizes with $N \in [10^3, 5 \times 10^4]$ and $N_{\text{rep}} = 25$. For each method, we report the selection frequency of K_0 and the average selected size (see Figure 3). AIC/BIC/ICL fit a model for each $\kappa \in [K]$ via EM and pick the best by the corresponding criterion. DSC fits a single SGMoE with $K = 4$, builds its dendrogram, and evaluates the criterion with $\omega_N = \log N$ (Section 3.6). AIC tends to overestimate at small N, while all methods recover K_0 for large N.

Misspecified Regime. We study ϵ-contamination with $p_0 = (1-\epsilon)p_{G_0} + \epsilon q$, where q is Laplace$(0, 1)$. Figure 4a shows the contaminated sample ($n = 5000$). Figures 4b and 4c report the proportion of correct selections and the average selected size. AIC/BIC/ICL behave similarly: they may find $K_0 = 2$ at small N, but tend to overselect as N grows, indicating sensitivity to contamination. DSC, leveraging dendrogram structure, is more robust and continues to select K_0 with non-negligible frequency even at large N.

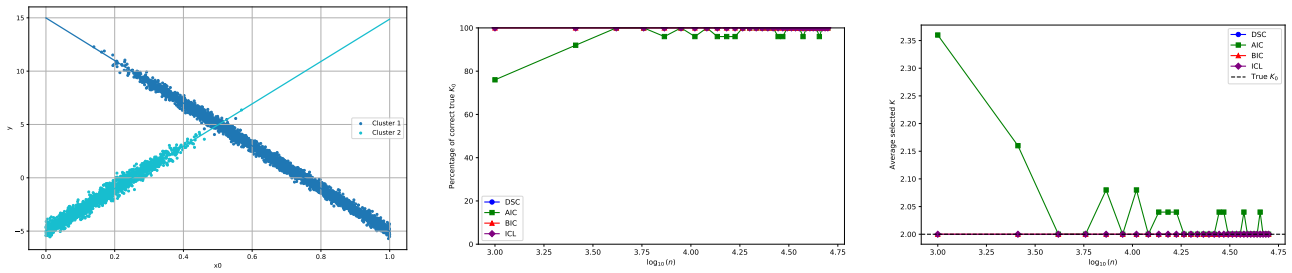
5 REAL DATA APPLICATION

We illustrate the dendrogram of mixing measures obtained from our SGMoE model using a real dataset from the study in Blein-Nicolas et al. (2024). The data originate from a large-scale experiment on maize aimed at understanding the genetic and molecular bases of drought-responsive traits from proteins expressed in the leaf (Blein-Nicolas et al., 2020; Prado et al., 2018), where 254 genotypes representing the genetic diversity of dent maize were grown under two watering conditions and phenotyped for seven ecophysiological traits.

After preprocessing and removing missing data as described in Blein-Nicolas et al. (2024), the final dataset consists of 233 maize genotypes ($N = 233$), two ecophysiological traits (outputs), which are *water use* (WU) and the proteins quantified under the *water deficit* (WD) condition, and 973 protein variables (inputs, $D = 973$). To reduce dimensionality and remove irrelevant features, we apply a Lasso procedure to select $D = 10$ protein variables most associated with the target trait and primarily focus on the ecophysiological trait WU.

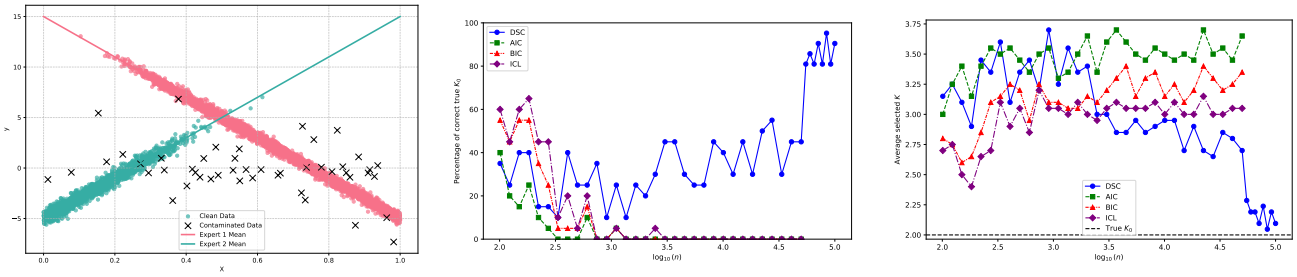
We then fit the SGMoE model with $K = 20$ clusters. To ensure a more robust initialization, we first cluster the data into 20 groups using the K-Means algorithm. The resulting cluster assignments are then used to initialize the gating and expert parameters of the SGMoE model, providing a stable starting point for the subsequent steps of the EM algorithm.

Figure 5a displays the dendrogram of the fitted mixing measure obtained by Algorithm 2, which reveals the hierarchical structure underlying the data. In this ex-



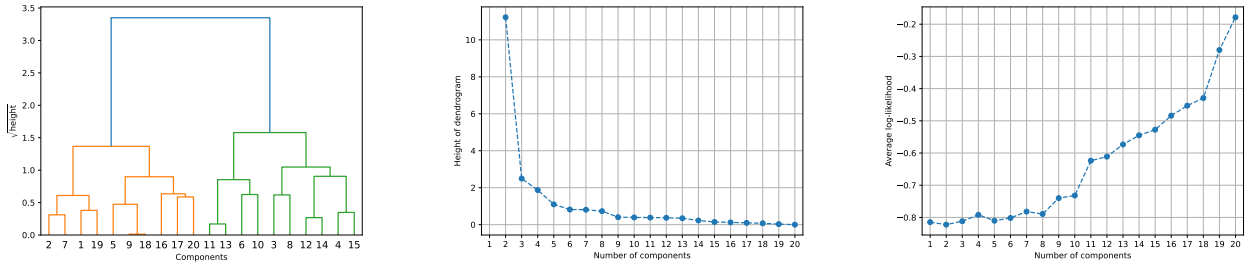
(a) Data clusters for G_0 ($n = 5000$). (b) Proportion of correct selections. (c) Average selected components.

Figure 3: DSC vs. AIC, BIC, and ICL for selecting $K_0 = 2$ of G_0.



(a) Contaminated sample ($n = 5000$). (b) Proportion of correct selections. (c) Average selected components.

Figure 4: Model selection under ϵ-contamination with $K_0 = 2$. After AIC, BIC, and ICL fail to recover K_0, we further test DSC on 8 sample sizes between 5.5×10^4 and 10^5, where it still recovers K_0 with high frequency.



(a) Dendrogram of mixing measure. (b) Heights between levels. (c) Average log-likelihood across levels.

Figure 5: Dendrogram of mixing measure inferred from maize drought-responsive traits dataset.

periment, both BIC and ICL select a single component, while DSC selects 2 components, and AIC overestimates with 18 components. The corresponding heights and average log-likelihoods across levels are shown in Figure 5b and Figure 5c, respectively. We observe that the merging heights generally decrease and approach zero, while the average log-likelihood stabilizes in a few initial levels. Notably, the height at level 2 is much larger than those at subsequent levels, suggesting that there should be two clusters in the data.

The dendrogram not only facilitates effective model selection but also unveils the hierarchical relationships among mixture components, thereby enhancing the interpretability of the estimated parameters in complex biological data settings.

6 CONCLUSION

This work shows that rate-aware geometry, realized through a Voronoi distance together with merging and dendrograms of mixing measures, delivers both fast parameter estimation and consistent, sweep-free model selection in SGMoE. We hope these ideas spur further advances in structured mixture models and expert architectures. Our analysis assumes linear softmax gates, Gaussian experts, compact $\mathbf{\Theta}$, and bounded covariate support. Extending the theory beyond these settings will require additional regularity and tail controls. Exact values of $\bar{r}(M)$ are known for $M \leq 3$; for $M \geq 4$ only lower bounds are available. While our guarantees use these bounds, sharper algebraic results would further tighten rates.

Acknowledgments

This project was funded primarily by the Australian Research Council Centre of Excellence for the Mathematical Analysis of Cellular Systems (CE230100001), which supported TrungTin Nguyen and Christopher Drovandi. Christopher Drovandi was also supported by an Australian Research Council Future Fellowship (FT210100260). Additional support was provided by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number A2025-18-02. The authors also acknowledge Dr. Dat Do (University of Chicago) for helpful discussions about the dendrogram of mixing measures for mixture models (Do et al., 2024).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. (Cited on page 2.)

Anandkumar, A., Hsu, D., & Kakade, S. M. (2012). A method of moments for mixture models and hidden markov models. In *COLT*. (Cited on page 2.)

Balakrishnan, S., Wainwright, M. J., & Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, 45, 77–120. (Cited on page 2.)

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

Berrettini, M., Galimberti, G., Ranciati, S., & Murphy, T. B. (2024). Identifying Brexit voting patterns in the British house of commons: an analysis based on Bayesian mixture models with flexible concomitant covariate effects. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(3), 621–638. (Cited on page 2.)

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7), 719–725. (Cited on page 2.)

Blein-Nicolas, M., Devijver, E., Gallopin, M., & Perthame, E. (2024). Nonlinear network-based quantitative trait prediction from biological data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(3), 796–815. (Cited on pages 8, 15, and 16.)

Blein-Nicolas, M., Negro, S. S., Balliau, T., Welcker, C., Cabrera-Bosquet, L., Nicolas, S. D., Charcosset, A., & Zivy, M. (2020). A systems genetics approach reveals environment-dependent associations between snps, protein coexpression, and drought-related traits in maize. *Genome Research*, 30(11), 1593–1604. (Cited on pages 8 and 15.)

Chamroukhi, F., Samé, A., Govaert, G., & Aknin, P. (2009). Time series modeling by a regression approach based on a latent process. *Neural Networks*, 22(5–6), 593–602. (Cited on page 7.)

Chen, J. (1995). Optimal Rate of Convergence for Finite Mixture Models. *The Annals of Statistics*, 23(1), 221 – 233. Publisher: Institute of Mathematical Statistics. (Cited on pages 2 and 7.)

Chen, Z., Deng, Y., Wu, Y., Gu, Q., & Li, Y. (2022). Towards Understanding the Mixture-of-Experts Layer in Deep Learning. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in Neural Information Processing Systems*. (Cited on page 2.)

Chong, M. C., Nguyen, H. D., & TrungTin Nguyen (2024). Risk Bounds for Mixture Density Estimation on Compact Domains via the h-Lifted Kullback–Leibler Divergence. *Transactions on Machine Learning Research*. (Cited on page 2.)

Dai, D., Deng, C., Zhao, C., Xu, R., Gao, H., Chen, D., Li, J., Zeng, W., Yu, X., Wu, Y., Xie, Z., Li, Y., Huang, P., Luo, F., Ruan, C., Sui, Z., & Liang, W. (2024). DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1280–1297). Bangkok, Thailand: Association for Computational Linguistics. (Cited on page 17.)

Do, D., Do, L., McKinley, S. A., Terhorst, J., & Nguyen, X. (2024). Dendrogram of mixing measures: Learning latent hierarchy and model selection for finite mixture models. *arXiv preprint arXiv:2403.01684*. (Cited on pages 2, 10, and 17.)

Do, D., Do, L., & Nguyen, X. (2025). Strong identifiability and parameter learning in regression with heterogeneous response. *Electronic Journal of Statistics*, 19(1), 131 – 203. Publisher: Institute of Mathematical Statistics and Bernoulli Society. (Cited on page 2.)

Do, T. G., Le, H. K., Nguyen, T., Pham, Q., Nguyen, B. T., Doan, T.-N., Liu, C., Ramasamy, S., Li, X., & HOI, S. (2023). HyperRouter: Towards Efficient Training and Inference of Sparse Mixture of Experts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* Singapore: Association for Computational Linguistics. (Cited on pages 1 and 17.)

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

Doss, N., Wu, Y., Yang, P., & Zhou, H. H. (2023). Optimal estimation of high-dimensional Gaussian location mixtures. *The Annals of Statistics*, 51(1), 62 – 95. Publisher: Institute of Mathematical Statistics. (Cited on page 2.)

Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M. J., Jordan, M. I., & Yu, B. (2020a). Sharp analysis of expectation-maximization for weakly identifiable models. *AISTATS*. (Cited on page 2.)

Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M. J., Jordan, M. I., & Yu, B. (2020b). Singularity, misspecification, and the convergence rate of EM. *Annals of Statistics*, 44, 2726–2755. (Cited on page 2.)

Eigen, D., Ranzato, M., & Sutskever, I. (2014). Learning factored representations in a deep mixture of experts. In *ICLR Workshops*. (Cited on page 1.)

Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23, 1–39. (Cited on pages 1 and 17.)

Forbes, F., Nguyen, H. D., Nguyen, T., & Arbel, J. (2022a). Mixture of expert posterior surrogates for approximate Bayesian computation. In *JDS 2022 - 53èmes Journées de Statistique de la Société Française de Statistique (SFdS)* Lyon, France. (Cited on page 2.)

Forbes, F., Nguyen, H. D., Nguyen, T., & Arbel, J. (2022b). Summary statistics and discrepancy measures for approximate Bayesian computation via surrogate posteriors. *Statistics and Computing*, 32(5), 85. (Cited on page 2.)

Foygel, R. & Drton, M. (2010). Extended Bayesian Information Criteria for Gaussian Graphical Models. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems*, volume 23: Curran Associates, Inc. (Cited on page 2.)

Frühwirth-Schnatter, S. (2019). Keeping the balance—Bridge sampling for marginal likelihood estimation in finite mixture, mixture of experts and Markov mixture models. *Brazilian Journal of Probability and Statistics*, 33(4), 706 – 733. (Cited on page 2.)

Frühwirth-Schnatter, S., Pamminger, C., Weber, A., & Winter-Ebmer, R. (2012). Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering. *Journal of Applied Econometrics*, 27(7), 1116–1137. (Cited on page 2.)

Frühwirth-Schnatter, S., Pittner, S., Weber, A., & Winter-Ebmer, R. (2018). Analysing plant closure effects using time-varying mixture-of-experts Markov chain clustering. *The Annals of Applied Statistics*, 12(3), 1796 – 1830. Publisher: Institute of Mathematical Statistics. (Cited on page 2.)

Genovese, C. R. & Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *The Annals of Statistics*, 28(4), 1105 – 1127. (Cited on page 2.)

Guha, A., Ho, N., & Nguyen, X. (2021). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli*, 27(4), 2159 – 2188. (Cited on page 2.)

Hardt, M. & Price, E. (2015). Tight bounds for learning a mixture of two gaussians. In *STOC*. (Cited on page 2.)

Heinrich, P. & Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6), 2844–2870. (Cited on page 2.)

Ho, B. H., Chi, L. N., Nguyen, T., Hoang, V. H., Nguyen, B. T., & Drovandi, C. (2025). A Unified Framework for Variable Selection in Model-Based Clustering with Missing Not at Random. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. (Cited on page 2.)

Ho, N. & Nguyen, X. (2016a). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6), 2726 – 2755. (Cited on pages 2 and 34.)

Ho, N. & Nguyen, X. (2016b). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10(1), 271–307. (Cited on page 2.)

Ho, N., Yang, C.-Y., & Jordan, M. I. (2022). Convergence Rates for Gaussian Mixtures of Experts. *Journal of Machine Learning Research*, 23(323), 1–81. (Cited on page 2.)

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1), 79–87. (Cited on page 1.)

Jiang, W. & Tanner, M. A. (1999). On the identifiability of mixtures-of-experts. *Neural Networks*, 12(9), 1253–1258. (Cited on page 25.)

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

Khalili, A., Yang, A. Y., & Da, X. (2024). Estimation and group-feature selection in sparse mixture-of-experts with diverging number of parameters. *Journal of Statistical Planning and Inference*, (pp. 106250). (Cited on page 2.)

Liang, H., Fan, Z., Sarkar, R., Jiang, Z., Chen, T., Zou, K., Cheng, Y., Hao, C., & Wang, Z. (2022). M^{3}ViT: Mixture-of-Experts Vision Transformer for Efficient Multi-task Learning with Model-Accelerator Co-design. In *NeurIPS*. (Cited on page 1.)

Manole, T. & Ho, N. (2020). Uniform convergence rates for maximum likelihood estimation under two-component gaussian mixture models. *arXiv preprint arXiv:2006.00704*. (Cited on page 2.)

Montuelle, L. & Le Pennec, E. (2014). Mixture of Gaussian regressions model with logistic weights, a penalized maximum likelihood approach. *Electronic Journal of Statistics*, 8(1), 1661–1695. (Cited on page 2.)

Nguyen, D. N. & Li, Z. (2024). Joint learning of Gaussian graphical models in heterogeneous dependencies of high-dimensional transcriptomic data. In *The 16th Asian Conference on Machine Learning (Conference Track)*. (Cited on page 2.)

Nguyen, H., Akbarian, P., Nguyen, T., & Ho, N. (2024a). A General Theory for Softmax Gating Multinomial Logistic Mixture of Experts. In *Proceedings of The 41st International Conference on Machine Learning*. (Cited on page 2.)

Nguyen, H., Akbarian, P., Yan, F., & Ho, N. (2024b). Statistical perspective of top-k sparse softmax gating mixture of experts. (Cited on page 17.)

Nguyen, H., Nguyen, T., & Ho, N. (2023a). Demystifying Softmax Gating Function in Gaussian Mixture of Experts. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems*, volume 36 (pp. 4624–4652).: Curran Associates, Inc. (Cited on pages 2, 3, 4, 7, 17, 19, 21, 25, 30, and 35.)

Nguyen, H., Nguyen, T., Nguyen, K., & Ho, N. (2024c). Towards Convergence Rates for Parameter Estimation in Gaussian-gated Mixture of Experts. In S. Dasgupta, S. Mandt, & Y. Li (Eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research* (pp. 2683–2691).: PMLR. (Cited on page 2.)

Nguyen, H. D. & Chamroukhi, F. (2018). Practical and theoretical aspects of mixture-of-experts modeling: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1246. (Cited on page 2.)

Nguyen, H. D., Chamroukhi, F., & Forbes, F. (2019). Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model. *Neurocomputing*, 366, 208–214. (Cited on page 2.)

Nguyen, H. D., Gupta, M., Westerhout, J., & Nguyen, T. (2025a). On the large-sample limits of some Bayesian model evaluation statistics. *arXiv preprint arXiv:2502.03846*. (Cited on page 2.)

Nguyen, H. D., Lloyd-Jones, L. R., & McLachlan, G. J. (2016). A universal approximation theorem for mixture-of-experts models. *Neural computation*, 28(12), 2585–2593. (Cited on page 2.)

Nguyen, H. D. & Nguyen, T. (2025). Modifications of the BIC for order selection in finite mixture models. *2506.20124*. (Cited on page 2.)

Nguyen, H. D., Nguyen, T., Chamroukhi, F., & McLachlan, G. J. (2021a). Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models. *Journal of Statistical Distributions and Applications*, 8(1), 13. (Cited on page 2.)

Nguyen, H. D., Nguyen, T., Westerhout, J., & Guo, X. (2025b). Approximation rates for finite mixtures of location-scale models. *arXiv preprint arXiv:2508.10612*. (Cited on page 2.)

Nguyen, T. (2021). *Model Selection and Approximation in High-dimensional Mixtures of Experts Models: from Theory to Practice*. PhD Thesis, Normandie Université. (Cited on page 2.)

Nguyen, T., Chamroukhi, F., Nguyen, H. D., & Forbes, F. (2021b). Non-asymptotic model selection in block-diagonal mixture of polynomial experts models. *Preprint. arXiv:2104.08959*. (Cited on page 2.)

Nguyen, T., Chamroukhi, F., Nguyen, H. D., & Forbes, F. (2022a). Model selection by penalization in mixture of experts models with a non-asymptotic approach. In *JDS 2022 - 53èmes Journées de Statistique de la Société Française de Statistique (SFdS)* Lyon, France. (Cited on page 2.)

Nguyen, T., Chamroukhi, F., Nguyen, H. D., & McLachlan, G. J. (2023b). Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces. *Communications in Statistics - Theory and Methods*, 52(14), 5048–5059. (Cited on page 2.)

Nguyen, T., Forbes, F., Arbel, J., & Duy Nguyen, H. (2024d). Bayesian nonparametric mixture of experts for inverse problems. *Journal of Nonparametric Statistics*, (pp. 1–60). (Cited on page 2.)

Nguyen, T., Nguyen, D. N., Nguyen, H. D., & Chamroukhi, F. (2023c). A non-asymptotic theory for

model selection in high-dimensional mixture of experts via joint rank and variable selection. In *AJCAI Australasian Joint Conference on Artificial Intelligence 2023*. (Cited on page 2.)

Nguyen, T., Nguyen, H. D., Chamroukhi, F., & Forbes, F. (2022b). A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models. *Electronic Journal of Statistics*, 16(2), 4742 – 4822. (Cited on page 2.)

Nguyen, T., Nguyen, H. D., Chamroukhi, F., & McLachlan, G. J. (2020). Approximation by finite mixtures of continuous density functions that vanish at infinity. *Cogent Mathematics & Statistics*, 7(1), 1750861. (Cited on page 2.)

Nguyen, T., Nguyen, H. D., Chamroukhi, F., & McLachlan, G. J. (2023d). Non-asymptotic oracle inequalities for the Lasso in high-dimensional mixture of experts. *arXiv:2009.10622*. (Cited on page 2.)

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1), 370–400. (Cited on page 2.)

Norets, A. (2010). Approximation of conditional densities by smooth mixtures of regressions. *The Annals of Statistics*, 38(3), 1733 – 1766. (Cited on page 2.)

Peng, F., Jacobs, R. A., & Tanner, M. A. (1996). Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixtures-of-Experts Models With an Application to Speech Recognition. *Journal of the American Statistical Association*, 91(435), 953–960. (Cited on page 1.)

Pham, Q., Do, G., Nguyen, H., Nguyen, T., Liu, C., Sartipi, M., Nguyen, B. T., Ramasamy, S., Li, X., Hoi, S., & others (2024). CompeteSMoE–Effective Training of Sparse Mixture of Experts via Competition. *arXiv preprint arXiv:2402.02526*. (Cited on pages 2 and 17.)

Prado, S. A., Cabrera-Bosquet, L., Grau, A., Coupel-Ledru, A., Millet, E. J., Welcker, C., & Tardieu, F. (2018). Phenomics allows identification of genomic regions affecting maize stomatal conductance with conditional effects of water deficit and evaporative demand. *Plant, Cell & Environment*, 41(2), 314–326. (Cited on pages 8 and 15.)

Rakhlin, A., Panchenko, D., & Mukherjee, S. (2005). Risk bounds for mixture density estimation. *ESAIM: PS*, 9, 220–229. (Cited on page 2.)

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464. (Cited on page 2.)

Shen, W., Tokdar, S. T., & Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3), 623–640. (Cited on page 2.)

Sin, C.-Y. & White, H. (1996). Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics*, 71(1), 207–225. (Cited on page 2.)

Sturmfels, B. (2002). *Solving systems of polynomial equations*. Number 97 in CBMS Regional Conference Series in Mathematics. American Mathematical Soc. (Cited on page 4.)

Thai, T., Nguyen, T., Do, D., Ho, N., & Drovandi, C. (2025). Model Selection for Gaussian-gated Gaussian Mixture of Experts Using Dendrograms of Mixing Measures. *arXiv preprint arXiv:2505.13052*. (Cited on pages 2 and 17.)

Tran, T., Nguyen, T., Bashar, M. A., Ho, N., Nayak, R., & Drovandi, C. (2026a). Fast Model Selection and Stable Optimization for Softmax-Gated Multinomial-Logistic Mixture of Experts Models. _eprint: 2602.07997. (Cited on page 2.)

Tran, T., Nguyen, T., Fort, G., Doan, T., Nguyen, H. D., Nguyen, B. T., Forbes, F., & Drovandi, C. (2026b). Revisiting Incremental Stochastic Majorization-Minimization Algorithms with Applications to Mixture of Experts. *arXiv preprint arXiv:2601.19811*. (Cited on page 2.)

van de Geer, S. (2000). *Empirical Processes in M-estimation*, volume 6. Cambridge university press. (Cited on pages 7, 22, and 34.)

Villani, C. (2003). *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society. (Cited on page 7.)

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer. (Cited on page 7.)

Westerhout, J., Nguyen, T., Guo, X., & Nguyen, H. D. (2024). On the Asymptotic Distribution of the Minimum Empirical Risk. In *Forty-first International Conference on Machine Learning*. (Cited on page 2.)

Wu, Y. & Yang, P. (2020). Optimal estimation of Gaussian mixtures via denoised method of moments. *The Annals of Statistics*, 48, 1987–2007. (Cited on page 2.)

Wu, Y. & Zhou, H. H. (2021). Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $o(\sqrt{n})$ iterations. *Mathematical Statistics and Learning*, 4, 143–220. (Cited on page 2.)

You, Z., Feng, S., Su, D., & Yu, D. (2021). Speechmoe: Scaling to large acoustic models with dynamic routing mixture of experts. In *Interspeech*. (Cited on page 2.)

You, Z., Feng, S., Su, D., & Yu, D. (2022). Speechmoe2: Mixture-of-experts model with improved routing. In *ICASSP 2022 - 2022 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7217–7221). (Cited on page 2.)

Yuksel, S. E., Wilson, J. N., & Gader, P. D. (2012). Twenty Years of Mixture of Experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8), 1177–1193. (Cited on page 2.)

Zens, G. (2019). Bayesian shrinkage in mixture-of-experts models: identifying robust determinants of class membership. *Advances in Data Analysis and Classification*, 13(4), 1019–1051. (Cited on page 2.)

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials for "Dendrograms of Mixing Measures for Softmax-Gated Gaussian Mixture of Experts: Consistency without Model Sweeps"

Supplementary Organization. This supplement has six parts. *First*, Appendix A provides additional biological background and preprocessing details for the maize drought-response dataset used in the main-paper illustration. *Second*, Appendix B gives a unified overview of unconditional mixtures, MoE, and SGMoE, highlighting their geometric and statistical differences and clarifying the motivation for the dendrogram framework. *Third*, Appendix C illustrates the Voronoi cells and the merge step underlying the SGMoE aggregation path. *Fourth*, Appendix D details the main technical challenges, namely softmax translation invariance, gate–expert PDE couplings, and algebraic cancellations; it also discusses in more detail the connection between the polynomial equations in eq. (3) and the over-specified SGMoE setting. *Fifth*, Appendix E expands the proof sketches for Lemma 1 and Theorems 1 to 3, highlighting how the Voronoi geometry drives the analysis. *Finally*, Appendix F presents the full proofs, together with the main proof ingredients and notational conventions used throughout the appendix.

A ADDITIONAL DETAILS ON THE MAIZE DROUGHT-RESPONSE DATA

This appendix provides additional biological context and a more explicit description of the data-processing pipeline used for the real-data illustration in the main paper. The goal is to clarify the provenance of the dataset, the meaning of the response and predictor variables, and the rationale for the preprocessing choices, while avoiding repetition of the main-text discussion.

Biological motivation and data provenance. The dataset comes from a broader systems-genetics effort on dent maize aimed at linking molecular variation in the leaf proteome to drought-related ecophysiological traits. In that line of work, a genetically diverse maize panel was grown under contrasting watering conditions and characterised using both high-throughput phenotyping and proteomics, with the broader objective of understanding how genotype-dependent molecular responses are related to drought adaptation and plant water-use behaviour (Blein-Nicolas et al., 2020; Prado et al., 2018). The statistical prediction study of Blein-Nicolas et al. (2024) used these biological measurements as a benchmark for multivariate trait prediction from high-dimensional proteomic covariates, specifically considering drought-related traits measured on a panel of 233 maize genotypes with 973 protein predictors (Blein-Nicolas et al., 2024). Systems-genetics reports associated with the same experimental programme also emphasise that the maize data were designed to study drought-related traits by integrating proteomic and genomic information (Blein-Nicolas et al., 2020).

Why this dataset is relevant here. This dataset is well suited to our SGMoE framework for three reasons. First, the sample is biologically heterogeneous: the maize panel spans substantial genetic diversity, so one should not expect all genotypes to follow a single homogeneous regression relationship. Second, drought response is known to be multi-mechanistic, with different molecular programmes potentially associated with different water-use strategies or stress-response profiles. Third, the predictor space is high-dimensional relative to the sample size, which makes model structure and interpretability especially important. These features make the dataset a natural test bed for a method that combines flexible conditional modelling with hierarchical aggregation and model selection. The resulting fitted components can then be interpreted as latent subgroups of genotypes sharing similar proteomic-to-phenotypic relationships rather than as merely algorithmic clusters.

Raw variables used in the illustration. Following Blein-Nicolas et al. (2024), we work with a cleaned subset of the original experiment after removing observations with missing values. The final analysis set contains $N = 233$ maize genotypes. The biological study recorded two drought-related ecophysiological outputs together with quantitative protein abundances measured under water-deficit conditions, yielding a predictor matrix with 973

protein variables before dimension reduction. In the present illustration, we focus primarily on the ecophysiological trait *water use* (WU), while the predictors are the leaf protein abundances measured under the water-deficit regime. This choice is scientifically meaningful because WU is directly linked to drought adaptation and integrates the cumulative effect of genotype-specific physiological regulation under stress.

Preprocessing strategy. The preprocessing follows the protocol used for the statistical benchmark in Blein-Nicolas et al. (2024), with the same starting point of a cleaned matrix after exclusion of incomplete observations. Since the original proteomic representation is very high-dimensional compared with the number of genotypes, we apply a supervised screening step before fitting the SGMoE. Concretely, we use a Lasso-based variable-selection procedure to extract a smaller subset of proteins that are most strongly associated with the target trait, and we retain $D = 10$ proteins for the analysis shown in the main paper. This reduction serves two purposes. Statistically, it improves stability in the small-N, large-D regime and reduces the risk that the fitted experts are driven by noise dimensions. Biologically, it yields a more interpretable model by restricting attention to a compact set of drought-informative protein signals. We stress that this Lasso step is used only as a preprocessing device; the clustering, aggregation path, and model selection are all performed by the SGMoE methodology thereafter.

Model fitting for the SGMoE path. After preprocessing, we fit an over-specified SGMoE with $K = 20$ initial components. Because mixture models can be sensitive to starting values, we initialise the fit using a preliminary K-means partition of the genotypes. These initial groups are then used to seed the gating and expert parameters before running the estimation procedure. The purpose of this intentionally over-specified fit is not to interpret all 20 initial components literally, but rather to create a rich starting representation from which the dendrogram path can merge redundant atoms and reveal a more stable low-dimensional structure. In this sense, the over-specified fit plays the same exploratory role as in the synthetic studies: it allows the subsequent aggregation path to separate persistent large-scale structure from small within-cell duplications.

Interpretation of the fitted path. In the main-text illustration, the fitted dendrogram suggests a pronounced split at level 2, while the average log-likelihood stabilises quickly along the path. From a biological viewpoint, this pattern is consistent with the idea that the maize panel contains a small number of broad genotype groups with distinct proteomic-response profiles under drought, rather than many sharply separated subpopulations. Thus, the selected two-expert solution should be read as a parsimonious summary of two dominant genotype–phenotype response regimes. The value of the dendrogram is therefore twofold: it provides a data-driven model-selection tool, and it offers a hierarchical view of how more complex over-specified representations collapse into a small number of biologically interpretable regimes.

Why the real-data example is informative for our methodology. Unlike the synthetic experiments, this dataset does not come with a known ground-truth number of experts. Its role is instead to illustrate the practical behaviour of the pathwise procedure on a genuinely heterogeneous biological problem. In particular, it shows that the dendrogram can remain informative even when standard information criteria disagree strongly, and that the selected solution can still be interpreted in domain terms through genotype–phenotype structure and early likelihood stabilisation. This complements the theory by demonstrating that the SGMoE aggregation path is not only a technical device for proving rates, but also a practically useful summary of heterogeneity in complex omics-assisted prediction problems.

B OVERVIEW OF MIXTURE AND MOE GEOMETRY

This section provides a unified overview of unconditional mixtures, MoE, and SGMoE, clarifying the geometric and statistical differences that motivate our dendrogram framework.

First, we recall the definitions of unconditional mixtures, MoE, and covariate-free gates.

- *Unconditional mixture:*

$$p(y) = \sum_{k=1}^{K} \pi_k \times f(y; \boldsymbol{\eta}_k).$$

- *MoE:*
$$p(y \mid \boldsymbol{x}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{x}) \times f(y; \boldsymbol{\eta}_k(\boldsymbol{x})),$$
where both the weights and the experts depend on $\boldsymbol{x}$.
- *Covariate-free gates:*
$$\pi_k(\boldsymbol{x}) \equiv \pi_k,$$
which reduces to a **mixture of regressions** when $\boldsymbol{\eta}_k(x)$ is a regression map.

Next, we analyze the difference from existing dendrogram approaches and compare them to Gaussian-gated Gaussian MoE (GGMoE) (Thai et al., 2025).

Difference from existing dendrogram approaches. Our framework differs from classical dendrogram methods in three key aspects. First, we introduce Voronoi-type losses in the gate space that respect softmax symmetry (common translations). Second, our method is tailored to conditional SGMoE geometry and provides finite-sample predictive parameter rates, building on Nguyen et al. (2023a). Third, earlier dendrogram approaches were developed for **unconditional** finite mixtures (Do et al., 2024) and rely on standard Wasserstein-type losses; they are not tailored to the **conditional** geometry and softmax-induced couplings of SGMoE. We introduce a **fast-rate-aware Voronoi loss** $\mathrm{D}_{\mathrm{FRA}}$ that (i) reduces to the exact-fit loss when cells are singletons and (ii) adds merged-moment block sums precisely in the slow directions created by Voronoi multi-coverage. This is motivated by the insufficiency of Wasserstein for SGMoE parameter geometry (and even the limitations of early Voronoi losses in Nguyen et al. (2023a)) and is spelled out in our appendix overview and the formal $\mathrm{D}_{\mathrm{FRA}}$/merge analysis.

Compare to GGMoE. GGMoE is a **generative** MoE that models covariates and gates via Bayes' rule, enabling closed-form EM M-steps but **not matching modern deep MoE practice**. Our framework aligns with contemporary **discriminative softmax/top-k** gating **learned end-to-end** over features: we **directly** optimize the conditional density $p(y \mid \boldsymbol{x})$ **without** a generative model for $\boldsymbol{x}$ (Dai et al., 2024; Do et al., 2023; Fedus et al., 2022; Nguyen et al., 2024b; Pham et al., 2024). This conditional focus aligns with predictive use but is **analytically harder**: softmax gating introduces a tight numerator–denominator coupling and **nontrivial gate–expert interactions** that do not arise in GGMoE's EM updates. Beyond this objective mismatch, we contribute **Voronoi-type losses** aligned with the **gate-induced partition** and establish **finite-sample MLE convergence rates** for SGMoE in both **exact-fit** and **over-specified** regimes, addressing the conditional SGMoE geometry directly rather than relying on a generative model for x. Empirically, beyond synthetic studies, we analyze a **maize proteomics** dataset of drought-responsive traits: the **dendrogram-guided SGMoE path** selects **two experts**, stabilizes the likelihood early, reveals a clear **hierarchical structure** in the mixing measure, and yields interpretable **genotype–phenotype** mappings, **complementing** GGMoE-centric work whose experiments are primarily synthetic (Thai et al., 2025).

C ILLUSTRATION OF VORONOI CELLS AND MERGE STEPS FOR SGMoE

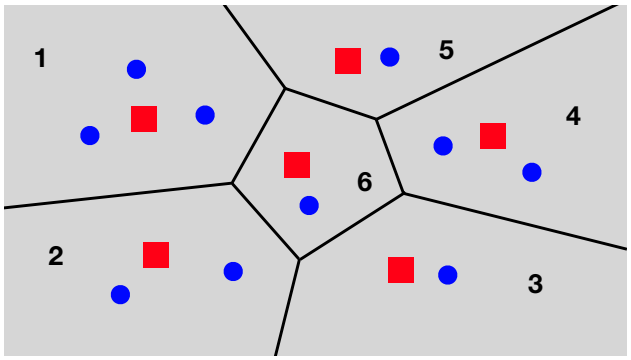
For a candidate mixing measure $G = \sum_{k=1}^{K} \exp(\omega_{0k})\, \delta_{(\boldsymbol{\omega}_{1k}, \boldsymbol{a}_k, b_k, \sigma_k)}$ and the true $G_0 = \sum_{k=1}^{K_0} \exp(\omega_{0k}^0)\, \delta_{(\boldsymbol{\omega}_{1k}^0, \boldsymbol{a}_k^0, b_k^0, \sigma_k^0)}$, define, for $k \in [K_0]$, the (parameter-space) Voronoi cell

$$\mathbb{A}_k(G) := \big\{\ell \in [K] : \ \|\boldsymbol{\theta}_\ell - \boldsymbol{\theta}_k^0\| \leq \|\boldsymbol{\theta}_\ell - \boldsymbol{\theta}_j^0\|, \ \forall j \neq k\big\}, \tag{12}$$

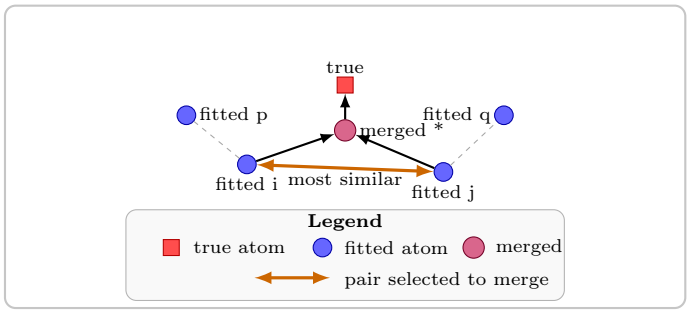
where $\boldsymbol{\theta}_\ell := (\boldsymbol{\omega}_{1\ell}, \boldsymbol{a}_\ell, b_\ell, \sigma_\ell)$. We use the softmax translation $(t_0, \boldsymbol{t}_1)$ from identifiability (cf. Proposition 1 of Nguyen et al., 2023a) and the shorthand $\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k} := \boldsymbol{\omega}_{1\ell} - \boldsymbol{\omega}_{1k}^0 - \boldsymbol{t}_1$, $\Delta\boldsymbol{a}_{\ell k} := \boldsymbol{a}_\ell - \boldsymbol{a}_k^0$, $\Delta b_{\ell k} := b_\ell - b_k^0$, $\Delta\sigma_{\ell k} := \sigma_\ell - \sigma_k^0$. For brevity we write $\mathbb{A}_k$ for $\mathbb{A}_k(G)$. (We restate eq. (12) only for completeness; throughout we reference the main-paper definition eq. (2).)

Explanation. Figure 6 summarizes the geometry and the merge step used by our method for an example with $K_0 = 6$ and $K = 10$: red squares denote true atoms of G_0, blue circles denote fitted atoms of G. Each Voronoi cell is generated by one true atom, and its cardinality $|\mathbb{A}_k|$ equals the number of fitted atoms assigned to that true atom (e.g., two circles in a cell imply $|\mathbb{A}_k| = 2$). Panel Figure 6a shows the Voronoi partition $\{\mathbb{A}_k\}_{k\in[K_0]}$ induced by G_0 as in eq. (2). Cells with $|\mathbb{A}_k| > 1$ reveal redundancy: multiple fitted atoms approximate the same

truth and create slow directions. Panel Figure 6b zooms into one such multi-covered cell and depicts the merge step at a visual level: the closest pair (w.r.t. our rate-weighted dissimilarity) is merged into a single aggregate; iterating this operation produces the aggregation path. Panel Figure 6c links the visuals to the mathematics: labels "fitted i," "fitted j," and "merged *" correspond to $\exp(\omega_{0i})\delta_{\boldsymbol{\theta}_i}$, $\exp(\omega_{0j})\delta_{\boldsymbol{\theta}_j}$, and $\exp(\omega_{0*})\delta_{\boldsymbol{\theta}_*}$. Pair selection uses d from eq. (7), and the softmax-weighted update rules are given in eq. (8). Together, these steps collapse slow directions within a cell, strengthen the loss along the path $\mathrm{D}_{\mathrm{FRA}}$ (eq. (6)), and enable our fast pathwise guarantees and sweep-free model selection via DSC.



(a) Voronoi cells $\{\mathbb{A}_k\}_{k\in[K_0]}, K_0 = 6, K = 10$ induced by G_0 as in eq. (2). Red squares are true atoms $\{\boldsymbol{\theta}_k^0\}$. Blue circles are fitted atoms $\{\boldsymbol{\theta}_\ell\}$. The cardinality $|\mathbb{A}_k|$ equals the number of fitted atoms approximating the true atom in that cell.



(b) Visual merge in a **multi-covered cell** $|\mathbb{A}_k| > 1$. Among four fitted atoms, the closest pair (i, j) by a dissimilarity is merged first; repeating yields the aggregation path.

Math key and merge equations. Visual labels i, j, and * correspond to

$$\text{fitted i: } \exp(\omega_{0i})\delta_{\boldsymbol{\theta}_i}, \qquad \text{fitted j: } \exp(\omega_{0j})\delta_{\boldsymbol{\theta}_j}, \qquad \text{merged *: } \exp(\omega_{0*})\delta_{\boldsymbol{\theta}_*}.$$

Pair selection uses the rate-weighted dissimilarity d in eq. (7). The softmax-weighted merge (eq. (8)) is

$$\omega_{0*} = \log\big(e^{\omega_{0i}} + e^{\omega_{0j}}\big), \quad \alpha_i = \frac{e^{\omega_{0i}}}{e^{\omega_{0i}} + e^{\omega_{0j}}}, \; \alpha_j = \frac{e^{\omega_{0j}}}{e^{\omega_{0i}} + e^{\omega_{0j}}},$$

$$\boldsymbol{\omega}_{1*} = \alpha_i\,\boldsymbol{\omega}_{1i} + \alpha_j\,\boldsymbol{\omega}_{1j}, \qquad b_* = \alpha_i\,b_i + \alpha_j\,b_j,$$

$$\boldsymbol{a}_* = \alpha_i\big[(\boldsymbol{\omega}_{1i} - \boldsymbol{\omega}_{1*})(b_i - b_*) + \boldsymbol{a}_i\big] + \alpha_j\big[(\boldsymbol{\omega}_{1j} - \boldsymbol{\omega}_{1*})(b_j - b_*) + \boldsymbol{a}_j\big],$$

$$\sigma_* = \alpha_i\big[(b_i - b_*)^2 + \sigma_i\big] + \alpha_j\big[(b_j - b_*)^2 + \sigma_j\big].$$

(c) Mathematical notation and closed-form merge in Section 3.3.

Figure 6: Voronoi geometry and merge step for SGMoE. Multi-covered cells $|\mathbb{A}_k| > 1$ signal redundant fitted atoms. The merge operator collapses them to a single aggregate that aligns with the true atom and improves the rate as formalized by our pathwise guarantees.

D THEORETICAL CHALLENGES: MORE DETAILS

The geometric picture above motivates the analytic tools below. We now detail three fundamental challenges in the statistical analysis of SGMoE that create substantial obstacles for parameter estimation and model selection:

(i) *Softmax translation invariance.* Gating parameters are identifiable only up to common translations. Unlike covariate-independent gating functions, the softmax gate is invariant under simultaneous shifts of intercepts and slopes, which makes the parameterization non-unique. As a result, standard identifiability arguments break down, and it becomes necessary to design translation-invariant loss functions. We address this by introducing the Voronoi partition and loss (see eq. (6)), which takes an infimum over translations and thereby aligns the loss with the geometry of gating partitions.

(ii) *Gate-expert PDE couplings.* The likelihood function exhibits intrinsic gate-expert interactions that induce coupled differential relations among parameters. These relations lead to numerous linear dependencies among derivative terms in Taylor expansions, which prevents a direct decomposition of density discrepancies $p_{\widehat{G}_N}(y|\boldsymbol{x}) - p_{G_0}(y|\boldsymbol{x})$ into independent components. Moreover, the parameters of the softmax gating numerators and the Gaussian experts are intrinsically linked through explicit PDEs,

$$\frac{\partial^2 u}{\partial \boldsymbol{\omega}_1 \partial b} = \frac{\partial u}{\partial \boldsymbol{a}}, \qquad \frac{\partial^2 u}{\partial b^2} = 2\frac{\partial u}{\partial \sigma}, \tag{13}$$

where $u(y|\boldsymbol{x};\boldsymbol{\omega},\boldsymbol{a},b,\sigma) := \exp(\boldsymbol{\omega}_1^\top \boldsymbol{x})\mathcal{N}(y|\boldsymbol{a}^\top \boldsymbol{x}+b,\sigma)$. Our analysis requires a systematic reorganization of these dependent terms to recover a meaningful set of independent directions.

(iii) *Algebraic cancellations.* Due to the tight coupling between numerators and denominators in the softmax-induced conditional density, higher-order cancellations in the expansions give rise to systems of polynomial equations introduced in eq. (3). The solvability of these systems determines the order of the first non-vanishing terms and directly controls the convergence rates of the MLE in over-specified models. This algebraic obstruction is a key source of non-standard, slower rates unique to SGMoE.

These challenges indicate that previously used loss functions, such as the Wasserstein distance, are insufficient for analyzing parameter quantities in either standard mixture models or mixtures with covariate-free gating functions. Moreover, the convergence rates of parameter estimates, as reported in Nguyen et al. (2023a), remain relatively slow due to the influence of the associated polynomial systems. Therefore, developing a dedicated method or algorithm, such as our DSC approach in Section 3, for models of this type is well motivated.

In addition, we also clarify the relationship between polynomial equations in eq. (3) and SGMoE in the over - specified case. Following Theorem 1, at $\kappa = K$, we can see that the convergence rate of Fast-Rate-Aware Voronoi distance $\mathrm{D}_{\mathrm{FRA}}$ is

$$\mathrm{D}_{\mathrm{FRA}}(\widehat{G}_N, G_0) \lesssim \left(\frac{\log N}{N}\right)^{1/2},$$

this is an "optimal" rate for a mixing measure. However, the convergence rate of some parameters such as $\boldsymbol{\omega}_1, \boldsymbol{a}, b, \sigma$ are not $\left(\frac{\log N}{N}\right)^{1/2}$ (following the definition of $\mathrm{D}_{\mathrm{FRA}}$). In particular, in the over-specified case, respectively $|\mathbb{A}_k| \geq 2$, then the associated parameters suffer slower rates of the order $N^{-1/(2\bar{r}(|\mathbb{A}_k|))}$ or $N^{-1/\bar{r}(|\mathbb{A}_k|)}$ (see Table 1).

To explain this connection, we revisit our proof for over-specified case. Firstly, we want to show that $\mathbb{E}_{\mathbf{x}}\left[V\left(p_G(\cdot \mid \boldsymbol{x}), p_{G_0}(\cdot \mid \boldsymbol{x})\right)\right] \gtrsim \mathrm{D}_{\mathrm{FRA}}\left(G, G_0\right)$ because we can see that if we obtain this argument, we will get the "optimal" convergence rate of $\mathrm{D}_{\mathrm{FRA}}$. We can rewrite the quantity Q_N as follows:

$$\begin{aligned} Q_N = \sum_{k=1}^{K_0}\sum_{\ell\in\mathbb{A}_k} \exp\left(\omega_{0\ell}^N\right)\Big[u\left(y|\boldsymbol{x};\boldsymbol{\omega}_{1\ell}^N,\boldsymbol{a}_\ell^N,b_\ell^N,\sigma_\ell^N\right) - u\left(y|\boldsymbol{x};\boldsymbol{\omega}_{1k}^0,\boldsymbol{a}_k^0,b_k^0,\sigma_k^0\right) - v\left(y|\boldsymbol{x};\boldsymbol{\omega}_{1\ell}^N\right) \\ + v\left(y|\boldsymbol{x};\boldsymbol{\omega}_{1k}^0\right)\Big] + \sum_{k=1}^{K_0}\left(\sum_{\ell\in\mathbb{A}_k}\exp\left(\omega_{0\ell}^N\right) - \exp\left(\omega_{0k}^0\right)\right)\left[u\left(y|\boldsymbol{x};\omega_{0k}^0,\boldsymbol{a}_k^0,b_k^0,\sigma_k^0\right) - v\left(y|\boldsymbol{x};\boldsymbol{\omega}_{1k}^0\right)\right], \end{aligned}$$

where we define $u\left(y|\boldsymbol{x};\boldsymbol{\omega}_1,\boldsymbol{a},b,\sigma\right) := \exp\left(\boldsymbol{\omega}_1^\top\boldsymbol{x}\right)\mathcal{N}\left(y|\boldsymbol{a}^\top\boldsymbol{x}+b,\sigma\right)$ and $v\left(y|\boldsymbol{x};\boldsymbol{\omega}_1\right) := \exp\left(\boldsymbol{\omega}_1^\top\boldsymbol{x}\right)p_{G_N}(y|\boldsymbol{x})$. Next, for each $k\in[K_0]$ and $\ell\in\mathbb{A}_k$, we denote $h_1\left(\boldsymbol{x},\boldsymbol{a}_k^0,b_k^0\right) := \left(\boldsymbol{a}_k^0\right)^\top\boldsymbol{x}+b_k^0$ and then apply the Taylor expansions to the functions $u\left(y|\boldsymbol{x};\boldsymbol{\omega}_{1\ell}^N,\boldsymbol{a}_\ell^N,b_\ell^N,\sigma_\ell^N\right)$ and $v\left(y|\boldsymbol{x};\boldsymbol{\omega}_{1\ell}^N\right)$ up to orders r_{1k} and r_{2k} (which we will choose later), respectively, as follows:

$$\begin{aligned} &u\left(y|\boldsymbol{x};\boldsymbol{\omega}_{1\ell}^N,\boldsymbol{a}_\ell^N,b_\ell^N,\sigma_\ell^N\right) - u\left(y|\boldsymbol{x};\boldsymbol{\omega}_{1k}^0,\boldsymbol{a}_k^0,b_k^0,\sigma_k^0\right) \\ &= \sum_{|\ell_1|+\ell_2=1}^{2r_{1k}} T_{\ell_1,\ell_2}^N(k)\boldsymbol{x}^{\ell_1}\exp\left(\left(\boldsymbol{\omega}_{1k}^0\right)^\top\boldsymbol{x}\right)\frac{\partial^{\ell_2}\mathcal{N}}{\partial h_1^{\ell_2}}\left(y|\left(\boldsymbol{a}_k^0\right)^\top\boldsymbol{x}+b_k^0,\sigma_k^0\right) + R_{1\ell k}(\boldsymbol{x},y), \\ &v\left(y|\boldsymbol{x};\boldsymbol{\omega}_{1\ell}^N\right) - v\left(y|\boldsymbol{x};\boldsymbol{\omega}_{1k}^0\right) = \sum_{|\gamma|=1}^{r_{2k}} S_\gamma^N(k)\boldsymbol{x}^\gamma\exp\left(\left(\boldsymbol{\omega}_{1k}^0\right)^\top\boldsymbol{x}\right)p_{G_N}(y|\boldsymbol{x}) + R_{2\ell k}(\boldsymbol{x},y), \end{aligned}$$

where $R_{1\ell k}(\boldsymbol{x}, y)$ and $R_{2\ell k}(\boldsymbol{x}, y)$ are Taylor remainders such that $R_{\rho\ell k}(\boldsymbol{x}, y)/\operatorname{D_{FRA}}(G_N, G_0)$ vanishes as $N \to \infty$ for $\rho \in \{1, 2\}$. As a result, the limit of $Q_N/\operatorname{D_{FRA}}(G_N, G_0)$ when n goes to infinity can be seen as a linear combination of elements of the following set:

$$\mathcal{W} := \left\{ \boldsymbol{x}^{\ell_1} \exp\left(\left(\boldsymbol{\omega}_{1k}^0\right)^\top \boldsymbol{x} \right) \frac{\partial^{\ell_2} \mathcal{N}}{\partial h_1^{\ell_2}} \left(y | \left(\boldsymbol{a}_k^0\right)^\top \boldsymbol{x} + b_k^0, \sigma_k^0 \right) : k \in [K_0], 0 \leq 2\left|\ell_1\right| + \ell_2 \leq 2r_{1k} \right\}$$
$$\cup \left\{ \boldsymbol{x}^{\gamma} \exp\left(\left(\boldsymbol{\omega}_{1k}^0\right)^\top \boldsymbol{x} \right) p_{G_0}(y|\boldsymbol{x}) : k \in [K_0], 0 \leq |\gamma| \leq r_{2k} \right\}$$

which is shown to be linearly independent. By the Fatou's lemma, we demonstrate that $Q_N/\operatorname{D_{FRA}}(G_N, G_0)$ goes to zero as $N \to \infty$, implying that all the coefficients in the representation of $Q_N/\operatorname{D_{FRA}}(G_N, G_0)$, denoted by $T_{\ell_1,\ell_2}^N(k)/\operatorname{D_{FRA}}(G_N, G_0)$ and $S_\gamma^N(k)/\operatorname{D_{FRA}}(G_N, G_0)$, vanish when $N \to \infty$. Given that result, we aim to select the Taylor orders r_{1k} and r_{2k} such that at least one among the limits of $T_{\ell_1,\ell_2}^N(k)/\operatorname{D_{FRA}}(G_N, G_0)$ and $S_\gamma^N(k)/\operatorname{D_{FRA}}(G_N, G_0)$ is different from zero, which leads to a contradiction. In the over-specified case, we assume that all the limits of $T_{\ell_1,\ell_2}^N(k)/\operatorname{D_{FRA}}(G_N, G_0)$ and $S_\gamma^N(k)/\operatorname{D_{FRA}}(G_N, G_0)$ equal zero. After some steps of considering typical limits as in the previous setting which requires $r_{2k} = 2$ for all $k \in [K_0]$, we encounter the following system of polynomial equations:

$$\sum_{\ell \in \mathbb{A}_k} \sum_{(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \alpha_3, \alpha_4) \in \mathbb{I}_{\ell_1,\ell_2}} \frac{p_{5\ell}^2 p_{1\ell}^{\boldsymbol{\alpha}_1} p_{2\ell}^{\boldsymbol{\alpha}_2} p_{3\ell}^{\alpha_3} p_{4\ell}^{\alpha_4}}{\boldsymbol{\alpha}_1! \boldsymbol{\alpha}_2! \alpha_3! \alpha_4!} = 0$$

for all $(\ell_1, \ell_2) \in \mathbb{N}^D \times \mathbb{N}$ such that $0 \leq |\ell_1| \leq r_{1k}, 0 \leq \ell_2 \leq r_{1k} - |\ell_1|$ and $|\ell_1| + \ell_2 \geq 1$ for some $k \in [K_0]$. Due to the construction of this system, it must have at least one non-trivial solution. Therefore, we choose $r_{1k} = \bar{r}(|\mathbb{A}_k|)$ for all $k \in [K_0]$.

To discuss about the value of $\bar{r}(M)$ with $M \geq 2$ in general, by Fact 1, we obtain $\bar{r}(M) = 2M$ for $M = 2, 3$ and as M increases, so does $\bar{r}(M)$. Hence, we predict that $\bar{r}(M) = 2M$. With this conjecture, we can see that the slow convergence rate of parameter estimation of SGMoE before we apply the merging atoms process.

E PROOF SKETCHES

In this section we expand the sketches for Lemma 1, Theorems 1 to 3.

Why the $\operatorname{D_{FRA}}$ loss in eq. (6)? When $\widehat{G}_N \to G_0$ with $K > K_0$, some Voronoi cells $\mathbb{A}_k$ are multi-covered. The slow directions in $\operatorname{D_O}$ (with exponents $\bar{r}(|\mathbb{A}_k|)$) arise from these cells. $\operatorname{D_{FRA}}$ augments $\operatorname{D_O}$ with first-order *merged-moment* block-sums that vanish when a cell behaves as a single aggregate. Thus $\operatorname{D_{FRA}}$ is simultaneously (i) exact-fit consistent, it reduces to $\operatorname{D_E}$ when $|\mathbb{A}_k| = 1$, and (ii) overfit-aware, penalizing precisely the slow directions that merging removes. In the over-specified case, cells with $|\mathbb{A}_k| > 1$ may persist; repeatedly merging atoms within such cells yields singletons and restores first-order behavior. Formally, using the density decomposition

$$Q_N = \Big[\sum_{k=1}^{K_0} \exp\left((\boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1)^\top \boldsymbol{x} + \omega_{0k}^0 + t_0 \right) \Big] \cdot \left[p_{G_N}(y|\boldsymbol{x}) - p_{G_0}(y|\boldsymbol{x}) \right],$$

we analyze the sums over indices with $|\mathbb{A}_k| > 1$ under $1 \leq |\boldsymbol{\ell}_1| + \ell_2 \leq 2\bar{r}(|\mathbb{A}_k|)$; for clarity, we also isolate the case $1 \leq |\boldsymbol{\ell}_1| + \ell_2 \leq 2$, corresponding to $|\mathbb{A}_k| = 1$. This leads directly to the merge operator and the aggregation path.

E.1 Proof Sketch of Lemma 1

We argue for the first merge $G^{(K)} \to G^{(K-1)}$; the rest follows by induction. Assume $\operatorname{D_{FRA}}(G^{(K)}, G_0) \to 0$. Then, for the Voronoi partition $\{\mathbb{A}_k\}$, there exist $(t_0, \boldsymbol{t}_1)$ such that, for every k,

$$\sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}) \to \exp(\omega_{0k}^0 + t_0), \qquad (\boldsymbol{\omega}_{1\ell}, \boldsymbol{a}_\ell, b_\ell, \sigma_\ell) \to (\boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1, \boldsymbol{a}_k^0, b_k^0, \sigma_k^0).$$

The minimizing pair (i, j) of d must lie in the same cell $\mathbb{A}_k$. Let the merged atom be $\exp(\omega_{0*})\delta_{(\boldsymbol{\omega}_{1*}, \boldsymbol{a}_*, b_*, \sigma_*)}$ as in eq. (8). Using the convexity of $z \mapsto \|z\|^m$ for $m \in \{\bar{r}(|\mathbb{A}_k|), \bar{r}(|\mathbb{A}_k|)/2\}$ and the identities implicit in eq. (8), we obtain the two key comparisons

$$(\exp\omega_{0i} + \exp\omega_{0j})\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1*k}, \Delta b_{*k})\|^{\bar{r}(|\mathbb{A}_k|)} \lesssim \sum_{t\in\{i,j\}} \exp\omega_{0t}\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1tk}, \Delta b_{tk})\|^{\bar{r}(|\mathbb{A}_k|)},$$
$$(\exp\omega_{0i} + \exp\omega_{0j})\|(\Delta\boldsymbol{a}_{*k}, \Delta\sigma_{*k})\|^{\bar{r}(|\mathbb{A}_k|)/2} \lesssim \sum_{t\in\{i,j\}} \exp\omega_{0t}\|(\Delta\boldsymbol{a}_{tk}, \Delta\sigma_{tk})\|^{\bar{r}(|\mathbb{A}_k|)/2}.$$

The block-sum terms in $\mathrm{D}_{\mathrm{FRA}}$ also decrease since the merged parameters are softmax-weighted averages. Collecting terms yields $\mathrm{D}_{\mathrm{FRA}}(G^{(K)}, G_0) \gtrsim \mathrm{D}_{\mathrm{FRA}}(G^{(K-1)}, G_0)$, proving monotonicity.

E.2 Proof Sketch of Theorem 1

(A) Inverse bound. We first prove an inverse inequality: there exists $C > 0$ depending only on G_0 and $\boldsymbol{\Theta}$ such that, for any $G \in \mathcal{O}_K(\boldsymbol{\Theta})$,

$$\mathbb{E}_{\mathbf{x}}\big[\,\mathrm{D}_{\mathrm{TV}}(p_G(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x}))\big] \;\geq\; C\;\mathrm{D}_{\mathrm{FRA}}(G, G_0). \tag{14}$$

The proof follows the *density decomposition* strategy in Nguyen et al. (2023a) but keeps all merged-moment block-sums that define $\mathrm{D}_{\mathrm{FRA}}$. Let

$$Q_N(\boldsymbol{x}, y) = \Big[\textstyle\sum_{k=1}^{K_0} \exp\big((\boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1)^\top\boldsymbol{x} + \omega_{0k}^0 + t_0\big)\Big] \cdot \big[p_G(y|\boldsymbol{x}) - p_{G_0}(y|\boldsymbol{x})\big].$$

A multi-index Taylor expansion (around $(\boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1, \boldsymbol{a}_k^0, b_k^0, \sigma_k^0)$ within each cell $\mathbb{A}_k$) up to order $\bar{r}(|\mathbb{A}_k|)$, together with the PDE identities $\partial^2 u/\partial\boldsymbol{\omega}_1\partial b = \partial u/\partial\boldsymbol{a}$ and $\partial^2 u/\partial b^2 = 2\,\partial u/\partial\sigma$, rewrites Q_N as a linear combination of basis functions

$$\boldsymbol{x}^{\boldsymbol{\ell}_1}\exp\big((\boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1)^\top\boldsymbol{x}\big)\,\frac{\partial^{\ell_2}}{\partial h_1^{\ell_2}}\mathcal{N}\big(y|\boldsymbol{a}_k^{0\top}\boldsymbol{x} + b_k^0, \sigma_k^0\big), \quad 1 \leq |\boldsymbol{\ell}_1| + \ell_2 \leq 2\bar{r}(|\mathbb{A}_k|),$$

with coefficients that are precisely the atomwise sums appearing in $\mathrm{D}_{\mathrm{FRA}}$ (up to constants). If eq. (14) failed, all these coefficients would have to vanish at a rate faster than $\mathrm{D}_{\mathrm{FRA}}(G, G_0)$, forcing a non-trivial solution to the polynomial system of eq. (3), in contradiction with the definition of $\bar{r}(\cdot)$ (Fact 1). This yields eq. (14).

(B) Applying density rates. By Proposition 2 of Nguyen et al. (2023a), $\mathbb{E}_{\mathbf{x}}[\mathrm{D}_{\mathrm{h}}^2(p_{\widehat{G}_N}(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x}))] = \mathcal{O}_{\mathbb{P}}((\log N/N)^{1/2})$. Using the inequality $\mathrm{D}_{\mathrm{TV}} \leq \sqrt{2}\;\mathrm{D}_{\mathrm{h}}^{2\,1/2}$ and eq. (14) with $G = \widehat{G}_N$, we obtain

$$\mathrm{D}_{\mathrm{FRA}}(\widehat{G}_N, G_0) = \mathcal{O}_{\mathbb{P}}\big((\log N/N)^{1/2}\big).$$

Now apply Lemma 1 along the aggregation path: for every $\kappa \in [K_0 + 1, K]$,

$$\mathrm{D}_{\mathrm{FRA}}(\widehat{G}_N^{(\kappa)}, G_0) \;\lesssim\; \mathrm{D}_{\mathrm{FRA}}(\widehat{G}_N, G_0) = \mathcal{O}_{\mathbb{P}}\big((\log N/N)^{1/2}\big).$$

For the exact-fit and under-fit levels $\kappa' \leq K_0$, $\mathrm{D}_{\mathrm{FRA}} = \mathrm{D}_{\mathrm{E}}$ by definition, which gives the second claim.

E.3 Proof Sketch of Theorem 2

For $\kappa \in [K_0 + 1, K]$, the height $\mathsf{h}_N^{(\kappa)}$ is the minimum d-distance between any two atoms of $\widehat{G}_N^{(\kappa)}$. Inside a multi-covered cell $\mathbb{A}_k(\widehat{G}_N)$, the Taylor/merged-moment analysis from the proof of eq. (14) implies that

$$\mathsf{d}\big(\exp(\widehat{\omega}_{0i})\delta_{\widehat{\boldsymbol{\theta}}_i}, \exp(\widehat{\omega}_{0j})\delta_{\widehat{\boldsymbol{\theta}}_j}\big) \;\lesssim\; \big\|(\Delta_{\boldsymbol{t}_1}\widehat{\boldsymbol{\omega}}_{1ik}, \Delta\widehat{b}_{ik})\big\|^2 + \big\|(\Delta\widehat{\boldsymbol{a}}_{ik}, \Delta\widehat{\sigma}_{ik})\big\|.$$

The right-hand side is controlled by $\mathrm{D}_{\mathrm{FRA}}(\widehat{G}_N^{(\kappa)}, G_0)$ with the exponents $\bar{r}(|\mathbb{A}_k|)$, hence $\mathsf{h}_N^{(\kappa)} \lesssim (\log N/N)^{1/\bar{r}(\widehat{G}_N)}$. For $\kappa' \leq K_0$, heights converge at parametric rate because atoms are separated and $\mathrm{D}_{\mathrm{E}}(\widehat{G}_N^{(\kappa')}, G_0^{(\kappa')}) = \mathcal{O}_{\mathbb{P}}((\log N/N)^{1/2})$.

E.4 Proof Sketch of Theorem 3

Let $\bar{\ell}_N(p_G) = N^{-1}\sum_{n=1}^N \log p_G(y_n|\mathbf{x}_n)$ and $\mathcal{L}(p_G) = \mathbb{E}_{(\mathbf{x},y)\sim P_{G_0}}[\log p_G(y|\boldsymbol{x})]$. Under Condition K, a local Lipschitz/curvature argument yields

$$\left|\bar{\ell}_N(p_G) - \mathcal{L}(p_G)\right| \lesssim \mathbb{E}_{(\mathbf{x},y)\sim P_{G_0}}\left[\mathrm{D_{TV}}(p_G(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x}))\right] + \text{empirical fluctuation}.$$

For $\kappa \geq K_0$, combine the inverse bound $\mathbb{E}[\mathrm{D_{TV}}] \gtrsim \mathrm{D_{FRA}}$ with $\mathrm{D_{FRA}}(\widehat{G}_N^{(\kappa)}, G_0) = \mathcal{O}_{\mathbb{P}}((\log N/N)^{1/2})$ and standard empirical-process bounds (e.g., van de Geer, 2000) to obtain $|\bar{\ell}_N(p_{\widehat{G}_N^{(\kappa)}}) - \mathcal{L}(p_{G_0})| \lesssim (\log N/N)^{1/(2\bar{r}(\widehat{G}_N))}$. For $\kappa' \leq K_0$, $\widehat{G}_N^{(\kappa')}$ is exact/under-fit and converges at parametric rate, hence $\bar{\ell}_N(p_{\widehat{G}_N^{(\kappa')}}) \to \mathcal{L}(p_{G_0^{(\kappa')}})$ in probability.

F PROOF OF MAIN RESULTS

Before proving the main results, we fix notation used throughout this appendix. For any natural number $N \in \mathbb{N}$, write $[N] := \{1, 2, \ldots, N\}$. Given two sequences of positive real numbers $\{a_N\}_{N=1}^\infty$ and $\{b_N\}_{N=1}^\infty$, we write $a_N = \mathcal{O}(b_N)$ (equivalently, $a_N \lesssim b_N$) to mean that there exists a constant $C > 0$ such that $a_N \leq C\, b_N$ for all $N \in \mathbb{N}$. For a vector $\boldsymbol{v} \in \mathbb{R}^D$ and any multi-index $\boldsymbol{p} \in \mathbb{N}^D$, set $|\boldsymbol{p}| := p_1 + \cdots + p_D$, $\boldsymbol{v}^{\boldsymbol{p}} := v_1^{p_1} v_2^{p_2} \cdots v_D^{p_D}$, $\boldsymbol{p}! := p_1! p_2! \cdots p_D!$, and let $\|\boldsymbol{v}\|_p$ denote its p-norm; by default, $\|\boldsymbol{v}\|$ refers to the 2-norm unless otherwise stated. We also use $\|\boldsymbol{A}\|$ for the Frobenius norm of a matrix $\boldsymbol{A} \in \mathbb{R}^{D\times D}$. For any set $\mathbb{S}$, $|\mathbb{S}|$ denotes its cardinality. For two probability density functions p and q with respect to the Lebesgue measure μ, define $\mathrm{D_{TV}}(p,q) := \frac{1}{2}\int |p-q|\,\mathrm{d}\mu$ as their total variation distance, while $\mathrm{D_h^2}(p,q) := \frac{1}{2}\int(\sqrt{p}-\sqrt{q})^2\,\mathrm{d}\mu$ denotes the squared Hellinger distance. Moreover, for $\boldsymbol{\mu} \in \mathbb{R}^D$, $\boldsymbol{\alpha} \in \mathbb{N}^D$, and a differentiable function f of $\boldsymbol{\mu}$, we write the partial derivative of order $|\boldsymbol{\alpha}|$ as

$$\frac{\partial^{|\boldsymbol{\alpha}|}}{\partial^{\boldsymbol{\alpha}}\boldsymbol{\mu}} f(\boldsymbol{\mu}) := \frac{\partial^{|\boldsymbol{\alpha}|}}{\partial\mu_1^{\boldsymbol{\alpha}_1}\partial\mu_2^{\boldsymbol{\alpha}_2}\cdots\partial\mu_D^{\boldsymbol{\alpha}_D}} f(\boldsymbol{\mu}).$$

Let $\boldsymbol{\Theta}$ be the parameter space. Write $\mathcal{E}_K(\boldsymbol{\Theta})$ for the collection of discrete probability measures on $\boldsymbol{\Theta}$ with exactly K atoms, and $\mathcal{O}_K(\boldsymbol{\Theta}) := \bigcup_{K'\leq K}\mathcal{E}_{K'}(\boldsymbol{\Theta})$ for those with at most K atoms. For a mixing measure $G = \sum_{k=1}^K \pi_k\,\delta_{\boldsymbol{\theta}_k}$, we (slightly abusively) refer to each component $\pi_k\,\delta_{\boldsymbol{\theta}_k}$ as an "atom," comprising both its weight π_k and parameter $\boldsymbol{\theta}_k$. Finally, the domain of parameters in the SGMoE is $\boldsymbol{\Theta}$, where $\boldsymbol{\eta}_k^0 := (\omega_{0k}^0, \boldsymbol{\omega}_{1k}^0, \boldsymbol{a}_k^0, b_k^0, \sigma_k^0) \in \boldsymbol{\Theta} \subset \mathbb{R}\times\mathbb{R}^D\times\mathbb{R}^D\times\mathbb{R}\times\mathbb{R}_{>0}$. Furthermore, assume $\boldsymbol{\Theta}$ is compact and $\mathcal{X} \subset \mathbb{R}^D$, the support of $\mathbf{x}$, is bounded. When clear from context, we drop $\boldsymbol{\Theta}$ and simply write $\mathcal{E}_K$ and $\mathcal{O}_K$.

F.1 Proof of Lemma 1

We prove the inequality $\mathrm{D_{FRA}}(G^{(K)}, G_0) \gtrsim \mathrm{D_{FRA}}(G^{(K-1)}, G_0)$, and the rest are similar.

Assume that $G_N := G_N^{(K)} = \sum_{k=1}^K \exp(\omega_{0k}^N)\delta_{(\boldsymbol{\omega}_{1k}^N, \boldsymbol{a}_k^N, b_k^N, \sigma_k^N)} \in \mathcal{E}_K$ varies so that $\mathrm{D_{FRA}}(G_N, G_0) \to 0$. We consider the Voronoi cells $\mathbb{A}_k^N := \mathbb{A}_k(G_N)$, for $k \in [K_0]$, of the mixing measure G_N generated by the true components of G_0. Since the argument in this proof is asymptotic, we assume without loss of generality that those Voronoi cells are independent of N for all $N \in \mathbb{N}$, i.e, $\mathbb{A}_k = \mathbb{A}_k^N$.

Then, we have $(\boldsymbol{a}_\ell^N, b_\ell^N, \sigma_\ell^N) \to (\boldsymbol{a}_k^0, b_k^0, \sigma_k^0)$, and there exist $t_0 \in \mathbb{R}$ and $\boldsymbol{t}_1 \in \mathbb{R}^D$ such that $\sum_{\ell\in\mathbb{A}_k^N} \exp(\omega_{0\ell}^N) \to \exp(\omega_{0k}^0 + t_0)$ and $\boldsymbol{\omega}_{1\ell}^N \to \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1$ for any $\ell \in \mathbb{A}_k$ and $k \in [K_0]$ as N approaches infinity.

We are going to show that the merging pair of indices (ℓ_1, ℓ_2) must belong to a common $\mathbb{A}_k$. Indeed, for every pair (ℓ_1, ℓ_2) in a common $\mathbb{A}_k$, since $(\boldsymbol{a}_{\ell_1}^N, b_{\ell_1}^N, \sigma_{\ell_1}^N) \to (\boldsymbol{a}_k^0, b_k^0, \sigma_k^0)$ and $\boldsymbol{\omega}_{1\ell_1}^N \to \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1$, and $(\boldsymbol{a}_{\ell_2}^N, b_{\ell_2}^N, \sigma_{\ell_2}^N) \to (\boldsymbol{a}_k^0, b_k^0, \sigma_k^0)$ and $\boldsymbol{\omega}_{1\ell_2}^N \to \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1$, we have

$$\mathsf{d}\left(\exp(\omega_{0\ell_1}^N)\delta_{(\boldsymbol{\omega}_{1\ell_1}^N, \boldsymbol{a}_{\ell_1}^N, b_{\ell_1}^N, \sigma_{\ell_1}^N)}, \exp(\omega_{0\ell_2}^N)\delta_{(\boldsymbol{\omega}_{1\ell_2}^N, \boldsymbol{a}_{\ell_2}^N, b_{\ell_2}^N, \sigma_{\ell_2}^N)}\right) \to 0, \quad \text{as } N \to \infty.$$

On the other hand, for every pair $(\ell, \ell') \in \mathbb{A}_k \times \mathbb{A}_{k'}$, where $k \neq k'$, because $(\boldsymbol{a}_\ell^N, b_\ell^N, \sigma_\ell^N) \to (\boldsymbol{a}_k^0, b_k^0, \sigma_k^0)$ and $\boldsymbol{\omega}_{1\ell}^N \to \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1$, and $(\boldsymbol{a}_{\ell'}^N, b_{\ell'}^N, \sigma_{\ell'}^N) \to (\boldsymbol{a}_{k'}^0, b_{k'}^0, \sigma_{k'}^0)$ and $\boldsymbol{\omega}_{1\ell'}^N \to \boldsymbol{\omega}_{1k'}^0 + \boldsymbol{t}_1$, we have

$$\mathsf{d}\left(\exp(\omega_{0\ell}^N)\delta_{(\boldsymbol{\omega}_{1\ell}^N, \boldsymbol{a}_\ell^N, b_\ell^N, \sigma_\ell^N)}, \exp(\omega_{0\ell'}^N)\delta_{(\boldsymbol{\omega}_{1\ell'}^N, \boldsymbol{a}_{\ell'}^N, b_{\ell'}^N, \sigma_{\ell'}^N)}\right) \gtrsim \|(\boldsymbol{\omega}_{1k}^0, b_k^0) - (\boldsymbol{\omega}_{1k'}^0, b_{k'}^0)\|^2 + \|(\boldsymbol{a}_k^0, \sigma_k^0) - (\boldsymbol{a}_{k'}^0, \sigma_{k'}^0)\|,$$

where the multiplicative constant is not dependent on N. Hence, the merging pair must belong to a common $\mathbb{A}_k$.

Next, for any $(t_0, \boldsymbol{t}_1) \in \mathbb{R} \times \mathbb{R}^D$ such that $\omega_{0k}^0 + t_0$ and $\boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1$ still lie inside the domain of the parameter space $\boldsymbol{\Theta}$, we define $\mathcal{D}(G_N, G_0, t_0, \boldsymbol{t}_1)$ as

$$
\begin{aligned}
\mathcal{D}(G_N, G_0, t_0, \boldsymbol{t}_1) := & \sum_{k:|\mathbb{A}_k|>1} \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) \Big(\|(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N, \Delta b_{\ell k}^N)\|^{\bar{r}_k} + \|(\Delta \boldsymbol{a}_{\ell k}^N, \Delta \sigma_{\ell k}^N)\|^{\bar{r}_k/2} \Big) \\
& + \sum_{k:|\mathbb{A}_k|=1} \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) \|(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N, \Delta \boldsymbol{a}_{\ell k}^N, \Delta b_{\ell k}^N, \Delta \sigma_{\ell k}^N)\| + \sum_{k=1}^{K_0} \Big| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) - \exp(\omega_{0k}^0 + t_0) \Big| \\
& + \sum_{k:|\mathbb{A}_k|>1} \Bigg(\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N)(\Delta b_{\ell k}^N) \| + \| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N)(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N) \| + \| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N)[(\Delta b_{\ell k}^N)^2 + (\Delta \sigma_{\ell k}^N)] \| \\
& \quad + \| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N)[(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N)(\Delta b_{\ell k}^N) + (\Delta \boldsymbol{a}_{\ell k}^N)] \| + \| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N)(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N)(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N)^\top \| \Bigg),
\end{aligned}
$$

in which $\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N := \boldsymbol{\omega}_{1\ell}^N - \boldsymbol{\omega}_{1k}^0 - \boldsymbol{t}_1$, $\Delta \boldsymbol{a}_{\ell k}^N := \boldsymbol{a}_\ell^N - \boldsymbol{a}_k^0$, $\Delta b_{\ell k}^N := b_\ell^N - b_k^0$, $\Delta \sigma_{\ell k}^N := \sigma_\ell^N - \sigma_k^0$, and $\bar{r}_k := \bar{r}\left(\mathbb{A}_k(\widehat{G}_N)\right)$.

We prove that $\mathcal{D}(G_N^{(K-1)}, G_0, t_0, \boldsymbol{t}_1) \lesssim \mathcal{D}(G_N^{(K)}, G_0, t_0, \boldsymbol{t}_1)$. Let the merging pair of indices (ℓ_1, ℓ_2) in the Voronoi cell $\mathbb{A}_k$, then $|\mathbb{A}_k| > 1$ and the merged atom is $\exp(\omega_{0*}^N) \delta_{(\boldsymbol{\omega}_{1*}^N, \boldsymbol{a}_*^N, b_*^N, \sigma_*^N)}$, i.e,

$$
\begin{aligned}
\omega_{0*}^N &= \log\left(\exp \omega_{0\ell_1}^N + \exp \omega_{0\ell_2}^N\right), \\
\boldsymbol{\omega}_{1*}^N &= \exp\left(\omega_{0\ell_1}^N - \omega_{0*}^N\right) \boldsymbol{\omega}_{1\ell_1}^N + \exp\left(\omega_{0\ell_2}^N - \omega_{0*}^N\right) \boldsymbol{\omega}_{1\ell_2}^N, \\
b_*^N &= \exp\left(\omega_{0\ell_1}^N - \omega_{0*}^N\right) b_{\ell_1}^N + \exp\left(\omega_{0\ell_2}^N - \omega_{0*}^N\right) b_{\ell_2}^N, \\
\boldsymbol{a}_*^N &= \frac{\exp(\omega_{0\ell_1}^N)}{\exp(\omega_{0*}^N)} \Big[(\boldsymbol{\omega}_{1\ell_1}^N - \boldsymbol{\omega}_{1*}^N)(b_{\ell_1}^N - b_*^N) + \boldsymbol{a}_{\ell_1}^N \Big] + \frac{\exp(\omega_{0\ell_2}^N)}{\exp(\omega_{0*}^N)} \Big[(\boldsymbol{\omega}_{1\ell_2}^N - \boldsymbol{\omega}_{1*}^N)(b_{\ell_2}^N - b_*^N) + \boldsymbol{a}_{\ell_2}^N \Big], \\
\sigma_*^N &= \frac{\exp(\omega_{0\ell_1}^N)}{\exp(\omega_{0*}^N)} \Big[(b_{\ell_1}^N - b_*^N)^2 + \sigma_{\ell_1}^N \Big] + \frac{\exp(\omega_{0\ell_2}^N)}{\exp(\omega_{0*}^N)} \Big[(b_{\ell_2}^N - b_*^N)^2 + \sigma_{\ell_2}^N \Big].
\end{aligned}
$$

Hence, we have that

$$
\begin{aligned}
\Big| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) - \exp(\omega_{0k}^0 + t_0) \Big| &= \Big| \sum_{\ell \in \mathbb{A}_k, \ell \notin \{\ell_1, \ell_2\}} \exp(\omega_{0\ell}^N) + \exp(\omega_{0*}^N) - \exp(\omega_{0k}^0 + t_0) \Big|, \\
\exp(\omega_{0*}^N)\, \Delta b_{*k}^N &= \exp(\omega_{0*}^N) \left(b_*^N - b_k^0\right) \\
&= \exp(\omega_{0*}^N) \left(\exp(\omega_{0\ell_1}^N - \omega_{0*}^N)\, b_{\ell_1}^N + \exp(\omega_{0\ell_2}^N - \omega_{0*}^N)\, b_{\ell_2}^N - b_k^0\right) \\
&= \exp(\omega_{0\ell_1}^N)\, b_{\ell_1}^N + \exp(\omega_{0\ell_2}^N)\, b_{\ell_2}^N - \exp(\omega_{0*}^N)\, b_k^0 \\
&= \exp(\omega_{0\ell_1}^N)(b_{\ell_1}^N - b_k^0) + \exp(\omega_{0\ell_2}^N)(b_{\ell_2}^N - b_k^0) \\
&= \exp(\omega_{0\ell_1}^N)\, \Delta b_{\ell_1 k}^N + \exp(\omega_{0\ell_2}^N)\, \Delta b_{\ell_2 k}^N.
\end{aligned}
$$

It follows that the term

$$
\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N)(\Delta b_{\ell k}^N) \| = \| \sum_{\ell \in \mathbb{A}_k \setminus \{\ell_1, \ell_2\}} \exp(\omega_{0\ell}^N)(\Delta b_{\ell k}^N) + \exp(\omega_{0*}^N)(\Delta b_{*k}^N) \|.
$$

Similarly, we can show that

$$\begin{aligned}
\|\sum_{\ell\in\mathbb{A}_k}\exp(\omega_{0\ell}^N)(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N)\| &= \|\sum_{\ell\in\mathbb{A}_k\setminus\{\ell_1,\ell_2\}}\exp(\omega_{0\ell}^N)(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N)+\exp(\omega_{0*}^N)(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1*k}^N)\|,\\
\|\sum_{\ell\in\mathbb{A}_k}\exp(\omega_{0\ell}^N)[(\Delta b_{\ell k}^N)^2+(\Delta\sigma_{\ell k}^N)]\| &= \Big\|\sum_{\ell\in\mathbb{A}_k\setminus\{\ell_1,\ell_2\}}\exp(\omega_{0\ell}^N)[(\Delta b_{\ell k}^N)^2+(\Delta\sigma_{\ell k}^N)]\\
&\quad+\exp(\omega_{0*}^N)[(\Delta b_{*k}^N)^2+(\Delta\sigma_{*k}^N)]\Big\|,\\
\|\sum_{\ell\in\mathbb{A}_k}\exp(\omega_{0\ell}^N)[(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N)(\Delta b_{\ell k}^N)+(\Delta\boldsymbol{a}_{\ell k}^N)]\| &= \Big\|\sum_{\ell\in\mathbb{A}_k\setminus\{\ell_1,\ell_2\}}\exp(\omega_{0\ell}^N)[(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N)(\Delta b_{\ell k}^N)+(\Delta\boldsymbol{a}_{\ell k}^N)]\\
&\quad+\exp(\omega_{0*}^N)[(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1*k}^N)(\Delta b_{*k}^N)+(\Delta\boldsymbol{a}_{*k}^N)]\Big\|,\\
\|\sum_{\ell\in\mathbb{A}_k}\exp(\omega_{0\ell}^N)(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N)(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N)^\top\| &= \Big\|\sum_{\ell\in\mathbb{A}_k\setminus\{\ell_1,\ell_2\}}\exp(\omega_{0\ell}^N)(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N)(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N)^\top\\
&\quad+\exp(\omega_{0*}^N)(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1*k}^N)(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1*k}^N)^\top\Big\|.
\end{aligned}$$

To this end, we show the key convexity step in detail. Firstly, we define α_i and $\boldsymbol{x}_i$ $(i=1,2)$ as follow:

$$\begin{aligned}
&\alpha_i := \exp(\omega_{0\ell_i}^N-\omega_{0*}^N),\ i=1,2,\\
&\boldsymbol{x}_i := \big(\boldsymbol{\omega}_{1\ell_i}^N-\boldsymbol{\omega}_{1k}^0-\boldsymbol{t}_1,\ b_{\ell_i}^N-b_k^0\big)=\big(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell_i k}^N,\ \Delta b_{\ell_i k}^N\big),\qquad i=1,2.
\end{aligned}$$

Note that $\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2\in(0,1)$ and $\boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2=1$ and by the definition of the merged atom eq. (8), we have the convex combination identity $\big(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1*k}^N,\ \Delta b_{*k}^N\big)=\boldsymbol{\alpha}_1\boldsymbol{x}_1+\boldsymbol{\alpha}_2\boldsymbol{x}_2$.

By the fact that $\bar r\geq 4$ (since $|\mathbb{A}_k|>1$), so we can use Jensen's inequality (convexity of the map $z\mapsto\|z\|^m$ with $m=\bar r_k$):

$$\big\|\boldsymbol{\alpha}_1\boldsymbol{x}_1+\boldsymbol{\alpha}_2\boldsymbol{x}_2\big\|^{\bar r_k}\leq\boldsymbol{\alpha}_1\|\boldsymbol{x}_1\|^{\bar r_k}+\boldsymbol{\alpha}_2\|\boldsymbol{x}_2\|^{\bar r_k}.$$

Multiply both sides by $\exp(\omega_{0*}^N)$ and substitute $\alpha_i=\exp(\omega_{0\ell_i}^N-\omega_{0*}^N)$:

$$\begin{aligned}
\big(\exp(\omega_{0\ell_1}^N)+\exp(\omega_{0\ell_2}^N)\big)\big\|\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1*k}^N,\Delta b_{*k}^N\big\|^{\bar r_k} &= \exp(\omega_{0*}^N)\big\|\boldsymbol{\alpha}_1\boldsymbol{x}_1+\boldsymbol{\alpha}_2\boldsymbol{x}_2\big\|^{\bar r_k}\\
&\leq \exp(\omega_{0*}^N)\big(\boldsymbol{\alpha}_1\|\boldsymbol{x}_1\|^{\bar r_k}+\boldsymbol{\alpha}_2\|\boldsymbol{x}_2\|^{\bar r_k}\big)\\
&= \exp(\omega_{0\ell_1}^N)\|\boldsymbol{x}_1\|^{\bar r_k}+\exp(\omega_{0\ell_2}^N)\|\boldsymbol{x}_2\|^{\bar r_k}\\
&= \exp(\omega_{0\ell_1}^N)\big\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell_1 k}^N,\Delta b_{\ell_1 k}^N)\big\|^{\bar r_k}+\exp(\omega_{0\ell_2}^N)\big\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell_2 k}^N,\Delta b_{\ell_2 k}^N)\big\|^{\bar r_k}.
\end{aligned}$$

Analogously, we can show that

$$\exp(\omega_{0\ell_1}^N)\,\big\|(\Delta\boldsymbol{a}_{\ell_1 k}^N,\ \Delta\sigma_{\ell_1 k}^N)\big\|^{\bar r_k/2}+\exp(\omega_{0\ell_2}^N)\,\big\|(\Delta\boldsymbol{a}_{\ell_2 k}^N,\ \Delta\sigma_{\ell_2 k}^N)\big\|^{\bar r_k/2}\gtrsim\big(\exp(\omega_{0\ell_1}^N)+\exp(\omega_{0\ell_2}^N)\big)\big\|(\Delta\boldsymbol{a}_{*k}^N,\ \Delta\sigma_{*k}^N)\big\|^{\bar r_k/2}.$$

Combining the two inequalities above gives the claimed comparison between the contribution of the merged atom and the contributions of the two original atoms:

$$\begin{aligned}
&\exp(\omega_{0\ell_1}^N)\Big(\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell_1 k}^N,\Delta b_{\ell_1 k}^N)\|^{\bar r_k}+\|(\Delta\boldsymbol{a}_{\ell_1 k}^N,\Delta\sigma_{\ell_1 k}^N)\|^{\bar r_k/2}\Big)+\exp(\omega_{0\ell_2}^N)\Big(\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell_2 k}^N,\Delta b_{\ell_2 k}^N)\|^{\bar r_k}+\|(\Delta\boldsymbol{a}_{\ell_2 k}^N,\Delta\sigma_{\ell_2 k}^N)\|^{\bar r_k/2}\Big)\\
&\qquad\gtrsim\big(\exp(\omega_{0\ell_1}^N)+\exp(\omega_{0\ell_2}^N)\big)\Big(\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1*k}^N,\Delta b_{*k}^N)\|^{\bar r_k}+\|(\Delta\boldsymbol{a}_{*k}^N,\Delta\sigma_{*k}^N)\|^{\bar r_k/2}\Big).
\end{aligned}$$

Hence

$$\mathcal{D}(G_N^{(K)},G_0,t_0,\boldsymbol{t}_1)\ \gtrsim\ \mathcal{D}(G_N^{(K-1)},G_0,t_0,\boldsymbol{t}_1),$$

and therefore

$$\mathrm{D}_{\mathrm{FRA}}(G_N^{(K-1)},G_0)\ \lesssim\ \mathrm{D}_{\mathrm{FRA}}(G_N^{(K)},G_0).$$

F.2 Proof of Theorem 1

First of all, we study the convergence rate of the MLE $\widehat{G}_N \in \mathcal{E}_K$ of the SGMoE; that is, we will show the inverse bound for SGMoE. We revisit the following result on the identifiability of the SGMoE models, which was previously studied in Jiang & Tanner (1999); Nguyen et al. (2023a).

Fact 3 (Nguyen et al., 2023a, Proposition 1). *For any mixing measures $G = \sum_{k=1}^{K} \exp(\omega_{0k})\delta_{(\boldsymbol{\omega}_{1k}, \boldsymbol{a}_k, b_k, \sigma_k)}$ and $G' = \sum_{k=1}^{K'} \exp(\omega'_{0k})\delta_{(\boldsymbol{\omega}'_{1k}, \boldsymbol{a}'_k, b'_k, \sigma'_k)}$, if we have $p_G(y|\boldsymbol{x}) = p_{G'}(y|\boldsymbol{x})$ for almost surely $(\boldsymbol{x}, y)$, then it follows that $K = K'$ and $G \equiv G'_{t_0, \boldsymbol{t}_1}$ where $G'_{t_0,\boldsymbol{t}_1} := \sum_{k=1}^{K'} \exp(\omega'_{0k} + t_0)\delta_{(\boldsymbol{\omega}'_{1k}+\boldsymbol{t}_1, \boldsymbol{a}'_k, b'_k, \sigma'_k)}$ for some $t_0 \in \mathbb{R}$ and $\boldsymbol{t}_1 \in \mathbb{R}^D$.*

The identifiability of the softmax gating Gaussian mixture of experts guarantees that the MLE $\widehat{G}_N$ converges to the true mixing measure G_0 (up to the translation of the parameters in the softmax gating).

Given the consistency of the MLE, it is natural to ask about its convergence rate to the true parameters. Our next result establishes the convergence rate of conditional density estimation $p_{\widehat{G}_N}(y|\boldsymbol{x})$ to the true conditional density $p_{G_0}(y|\boldsymbol{x})$, which lays an important foundation for the study of MLE's convergence rate.

Fact 4 (Nguyen et al., 2023a, Proposition 2). *The density estimation $p_{\widehat{G}_N}(y|\boldsymbol{x})$ converges to the true density $p_{G_0}(y|\boldsymbol{x})$ under the Hellinger distance $\mathrm{D}^2_{\mathrm{h}}(\cdot,\cdot)$ at the following rate:*

$$\mathbb{E}_{\mathbf{x}}[\mathrm{D}^2_{\mathrm{h}}(p_{\widehat{G}_N}(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x}))] = \mathcal{O}_P(\sqrt{\log(N)/N}).$$

That is,

$$\mathbb{P}(\mathbb{E}_{\mathbf{x}}[\mathrm{D}^2_{\mathrm{h}}(p_{\widehat{G}_N}(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x}))] > C(\log(N)/N)^{1/2}) \lesssim \exp(-c\log N),$$

where c and C are universal constants.

The result of Fact 4 indicates that under either the exact-specified or over-specified cases of the SGMoE, the rate of the conditional density function $p_{\widehat{G}_N}(y|\boldsymbol{x})$ to the true one $p_{G_0}(y|\boldsymbol{x})$ under Hellinger distance is of order $\mathcal{O}(N^{-1/2})$ (up to some logarithmic factors), which is parametric on the sample size.

Now, we establish the convergence rate of the MLE under the over-specified case of the SGMoE via the Fast-Rate-Aware Voronoi Distance $\mathrm{D}_{\mathrm{FRA}}$.

Theorem 5. *Under the over-specified case of the SGMoE, namely, when $K > K_0$, we obtain that*

$$\mathbb{E}_{\mathbf{x}}[\mathrm{D}^2_{\mathrm{h}}(p_G(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x}))] \geq C \cdot \mathrm{D}_{\mathrm{FRA}}(G, G_0),$$

for any $G \in \mathcal{O}_K$ where C is some universal constant depending only on G_0 and $\boldsymbol{\Theta}$. Therefore, that lower bound leads to the following convergence rate of the MLE:

$$\mathbb{P}(\mathrm{D}_{\mathrm{FRA}}(\widehat{G}_N, G_0) > C'(\log(N)/N)^{1/2}) \lesssim \exp(-c\log N), \tag{15}$$

where C' and c are some universal constants.

Proof of Theorem 5. We are going to prove that there exists a constant $C > 0$ depending only on G_0 and $\boldsymbol{\Theta}$ such that, for any $G \in \mathcal{O}_K$,

$$\mathbb{E}_{\mathbf{x}}\big[\mathrm{D}_{\mathrm{TV}}(p_G(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x}))\big] \gtrsim \mathrm{D}_{\mathrm{FRA}}(G, G_0). \tag{16}$$

Then, by the Fact 4, we get the convergence rate of the MLE of SGMoE.

Local version: Firstly, we prove the local version of the eq. (16):

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{O}_K : \mathrm{D}_{\mathrm{FRA}}(G, G_0) \leq \varepsilon} \mathbb{E}_{\mathbf{x}}\big[\mathrm{D}_{\mathrm{TV}}(p_G(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x}))\big] / \mathrm{D}_{\mathrm{FRA}}(G, G_0) > 0. \tag{17}$$

Assume that the inequality in eq. (17) does not hold true, there exists a sequence of mixing measures $G_N := \sum_{k=1}^{K_N} \exp(\omega^N_{0k})\delta_{(\boldsymbol{\omega}^N_{1k}, \boldsymbol{a}^N_k, b^N_k, \sigma^N_k)} \in \mathcal{O}_K$ such that

$$\mathbb{E}_{\mathbf{x}}[\mathrm{D}_{\mathrm{TV}}(p_{G_N}(\cdot|x), p_{G_0}(\cdot|x)]/\,\mathrm{D}_{\mathrm{FRA}}(G_N, G_0) \to 0,$$
$$\mathrm{D}_{\mathrm{FRA}}(G_N, G_0) \to 0,$$

when N to infinity. Since the proof argument is asymptotic, we also assume that $K_N = K' \leq K$ for all $N \geq 1$. Next, we consider the Voronoi cells $\mathbb{A}_k^N := \mathbb{A}_k(G_N)$, for $k \in [K_0]$, of the mixing measure G_N generated by the true components of G_0. And we can assume without loss of generality (WLOG) that those Voronoi cells are independent of N for all $N \in \mathbb{N}$, i.e. $\mathbb{A}_k = \mathbb{A}_k^N$. Additionally, since $\mathrm{D}_{\mathrm{FRA}}(G_N, G_0) \to 0$, we have $(\boldsymbol{a}_\ell^N, b_\ell^N, \sigma_\ell^N) \to (\boldsymbol{a}_\ell^0, b_\ell^0, \sigma_\ell^0)$ for any $\ell \in \mathbb{A}_k$ as $N \to \infty$. Furthermore, there exist $t_0 \in \mathbb{R}$ and $\boldsymbol{t}_1 \in \mathbb{R}^D$ such that $\sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) \to \exp(\omega_{0k}^0 + t_0)$ and $\boldsymbol{\omega}_{1\ell}^N \to \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1$ as N approaches infinity for any $\ell \in \mathbb{A}_k$ and $k \in [K_0]$. It suggests that we can upper bound $\mathrm{D}_{\mathrm{FRA}}$ as $\mathrm{D}_{\mathrm{FRA}}(G_N, G_0) \leq \mathrm{D}_{\mathrm{V}}(G_N, G_0)$, where

$$
\begin{aligned}
\mathrm{D}_{\mathrm{V}}(G_N, G_0) &:= \sum_{k:|\mathbb{A}_k|>1} \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) \Big(\|(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N, \Delta b_{\ell k}^N)\|^{\bar{r}(|\mathbb{A}_k|)} + \|(\Delta \boldsymbol{a}_{\ell k}^N, \Delta \sigma_{\ell k}^N)\|^{\bar{r}(|\mathbb{A}_k|)/2} \Big) \\
&+ \sum_{k:|\mathbb{A}_k|=1} \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) \|(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N, \Delta \boldsymbol{a}_{\ell k}^N, \Delta b_{\ell k}^N, \Delta \sigma_{\ell k}^N)\| + \sum_{k=1}^{K_0} \Big| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) - \exp(\omega_{0k}^0 + t_0) \Big| \\
&+ \sum_{k:|\mathbb{A}_k|>1} \Bigg(\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N)(\Delta b_{\ell k}^N) \| + \| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N)(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N) \| + \| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N)[(\Delta b_{\ell k}^N)^2 + (\Delta \sigma_{\ell k}^N)] \| \\
&\quad + \| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N)[(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N)(\Delta b_{\ell k}^N) + (\Delta \boldsymbol{a}_{\ell k}^N)] \| + \| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N)(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N)(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N)^\top \| \Bigg),
\end{aligned}
$$

in which $\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N := \boldsymbol{\omega}_{1\ell}^N - \boldsymbol{\omega}_{1k}^0 - \boldsymbol{t}_1$, $\Delta \boldsymbol{a}_{\ell k}^N := \boldsymbol{a}_\ell^N - \boldsymbol{a}_k^0$, $\Delta b_{\ell k}^N := b_\ell^N - b_k^0$, $\Delta \sigma_{\ell k}^N := \sigma_\ell^N - \sigma_k^0$.

Step 1: Density Decomposition

In this step, we try to find a density decomposition for the quatity $Q_N = \Big[\sum_{k=1}^{K_0} \exp\Big(\big(\boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1\big)^\top \boldsymbol{x} + \omega_{0k}^0 + t_0\Big)\Big] \cdot [p_{G_N}(y|\boldsymbol{x}) - p_{G_0}(y|\boldsymbol{x})]$:

$$
\begin{aligned}
Q_N &= \sum_{k=1}^{K_0} \sum_{\ell \in \mathbb{A}_k} \exp\big(\omega_{0\ell}^N\big) \left[u\left(y|\boldsymbol{x}; \boldsymbol{\omega}_{1\ell}^N, \boldsymbol{a}_\ell^N, b_\ell^N, \sigma_\ell^N\right) - u\left(y|\boldsymbol{x}; \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1, \boldsymbol{a}_k^0, b_k^0, \sigma_k^0\right) \right] \\
&\quad - \sum_{k=1}^{K_0} \sum_{\ell \in \mathbb{A}_k} \exp\big(\omega_{0\ell}^N\big) \left[v\left(y|\boldsymbol{x}; \boldsymbol{\omega}_{1\ell}^N\right) - v\left(y|\boldsymbol{x}; \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1\right) \right] \\
&\quad + \sum_{k=1}^{K_0} \left(\sum_{\ell \in \mathbb{A}_k} \exp\big(\omega_{0\ell}^N\big) - \exp\big(\omega_{0k}^0 + t_0\big) \right) \left[u\left(y|\boldsymbol{x}; \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1, \boldsymbol{a}_k^0, b_k^0, \sigma_k^0\right) - v\left(y|\boldsymbol{x}; \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1\right) \right], \\
&:= A_N + B_N + E_N,
\end{aligned}
$$

where we denote $u\left(y|\boldsymbol{x}; \boldsymbol{\omega}_1, \boldsymbol{a}, b, \sigma\right) := \exp\left(\boldsymbol{\omega}_1^\top \boldsymbol{x}\right) \mathcal{N}\left(y|\boldsymbol{a}^\top \boldsymbol{x} + b, \sigma\right)$ and $v\left(y|\boldsymbol{x}; \boldsymbol{\omega}_1\right) := \exp\left(\boldsymbol{\omega}_1^\top \boldsymbol{x}\right) p_{G_N}(y|\boldsymbol{x})$.

Since each Voronoi cell $\mathbb{A}_k$ possibly has more than one element, we continue to decompose A_N as follows:

$$
\begin{aligned}
A_N &= \sum_{k:|\mathbb{A}_k|>1} \sum_{\ell \in \mathbb{A}_k} \exp\big(\omega_{0\ell}^N\big) \left[u\left(y|\boldsymbol{x}; \boldsymbol{\omega}_{1\ell}^N, \boldsymbol{a}_\ell^N, b_\ell^N, \sigma_\ell^N\right) - u\left(y|\boldsymbol{x}; \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1, \boldsymbol{a}_k^0, b_k^0, \sigma_k^0\right) \right] \\
&\quad + \sum_{k:|\mathbb{A}_k|=1} \sum_{\ell \in \mathbb{A}_k} \exp\big(\omega_{0\ell}^N\big) \left[u\left(y|\boldsymbol{x}; \boldsymbol{\omega}_{1\ell}^N, \boldsymbol{a}_\ell^N, b_\ell^N, \sigma_\ell^N\right) - u\left(y|\boldsymbol{x}; \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1, \boldsymbol{a}_k^0, b_k^0, \sigma_k^0\right) \right] \\
&:= A_{N,1} + A_{N,2}.
\end{aligned}
$$

Now, we perform Taylor expansion up to the $\bar{r}(|\mathbb{A}_k|)-$th order, and then rewrite $A_{N,1}$ with a note that $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \alpha_3, \alpha_4) \in \mathbb{N}^D \times \mathbb{N}^D \times \mathbb{N} \times \mathbb{N}$ as follows:

$$
\begin{aligned}
A_{N,1} &= \sum_{k:|\mathbb{A}_k|>1} \sum_{\ell \in \mathbb{A}_k} \sum_{|\boldsymbol{\alpha}|=1}^{\bar{r}(|\mathbb{A}_k|)} \frac{\exp\left(\omega_{0\ell}^N\right)}{\boldsymbol{\alpha}!} (\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N)^{\boldsymbol{\alpha}_1} (\Delta \boldsymbol{a}_{\ell k}^N)^{\boldsymbol{\alpha}_2} (\Delta b_{\ell k}^N)^{\alpha_3} (\Delta \sigma_{\ell k}^N)^{\alpha_4} \\
&\quad \times \frac{\partial^{|\boldsymbol{\alpha}_1|+|\boldsymbol{\alpha}_2|+\alpha_3+\alpha_4}}{\partial \boldsymbol{\omega}_1^{\boldsymbol{\alpha}_1} \partial \boldsymbol{a}^{\boldsymbol{\alpha}_2} \partial b^{\alpha_3} \partial \sigma^{\alpha_4}} u(y|\boldsymbol{x}; \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1, \boldsymbol{a}_k^0, b_k^0, \sigma_k^0) + R_1^N(\boldsymbol{x}, y),
\end{aligned}
$$

where $R_1^N(\boldsymbol{x}, y)$ is the remainder term such that

$$R_1^N(\boldsymbol{x}, y) = o\left(\sum_{k:|\mathbb{A}_k|>1}\sum_{\ell\in\mathbb{A}_k}\exp(\omega_{0\ell}^N)(\|\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N\|^{\bar{r}(|\mathbb{A}_k|)} + \|\Delta\boldsymbol{a}_{\ell k}^N\|^{\bar{r}(|\mathbb{A}_k|)} + \|\Delta b_{\ell k}^N\|^{\bar{r}(|\mathbb{A}_k|)} + \|\Delta\sigma_{\ell k}^N\|^{\bar{r}(|\mathbb{A}_k|)})\right).$$

Next, for each $k \in [K_0]$ and $\ell \in \mathbb{A}_k$, we denote $h_1(\boldsymbol{x}, \boldsymbol{a}, b) := (\boldsymbol{a})^\top \boldsymbol{x} + b$. By the partial differential equations

$$\frac{\partial^2 u}{\partial \boldsymbol{\omega}_1 \partial b} = \frac{\partial u}{\partial \boldsymbol{a}}; \quad \frac{\partial^2 u}{\partial b^2} = 2\frac{\partial u}{\partial \sigma},$$

we have

$$\frac{\partial^{|\boldsymbol{\alpha}_2|} u}{\partial \boldsymbol{a}^{\boldsymbol{\alpha}_2}} = \frac{\partial^{2|\boldsymbol{\alpha}_2|} u}{\partial \boldsymbol{\omega}_1^{\boldsymbol{\alpha}_2} \partial b^{|\boldsymbol{\alpha}_2|}}; \quad \frac{\partial^{\alpha_4} u}{\partial \sigma^{\alpha_4}} = \frac{1}{2^{\alpha_4}} \cdot \frac{\partial^{2\alpha_4} u}{\partial b^{2\alpha_4}}.$$

Hence

$$\frac{\partial^{|\boldsymbol{\alpha}_1|+|\boldsymbol{\alpha}_2|+\alpha_3+\alpha_4} u}{\partial \boldsymbol{\omega}_1^{\boldsymbol{\alpha}_1} \partial \boldsymbol{a}^{\boldsymbol{\alpha}_2} \partial b^{\alpha_3} \partial \sigma^{\alpha_4}} = \frac{1}{2^{\alpha_4}} \cdot \frac{\partial^{(|\boldsymbol{\alpha}_1|+|\boldsymbol{\alpha}_2|)+(|\boldsymbol{\alpha}_2|+\alpha_3+2\alpha_4)} u}{\partial \boldsymbol{\omega}_1^{\boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2} \partial b^{|\boldsymbol{\alpha}_2|+\alpha_3+2\alpha_4}}.$$

It follows that

$$\begin{aligned}
A_{N,1} = &\sum_{k:|\mathbb{A}_k|>1}\sum_{\ell\in\mathbb{A}_k}\sum_{|\boldsymbol{\alpha}|=1}^{\bar{r}(|\mathbb{A}_k|)} \frac{\exp(\omega_{0\ell}^N)}{\boldsymbol{\alpha}!}(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N)^{\boldsymbol{\alpha}_1}(\Delta\boldsymbol{a}_{\ell k}^N)^{\boldsymbol{\alpha}_2}(\Delta b_{\ell k}^N)^{\alpha_3}(\Delta\sigma_{\ell k}^N)^{\alpha_4} \\
&\times \frac{1}{2^{\alpha_4}} \cdot \frac{\partial^{(|\boldsymbol{\alpha}_1|+|\boldsymbol{\alpha}_2|)+(|\boldsymbol{\alpha}_2|+\alpha_3+2\alpha_4)}}{\partial \boldsymbol{\omega}_1^{\boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2} \partial b^{|\boldsymbol{\alpha}_2|+\alpha_3+2\alpha_4}} u(y|\boldsymbol{x}; \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1, \boldsymbol{a}_k^0, b_k^0, \sigma_k^0) + R_1^N(\boldsymbol{x}, y) \\
= &\sum_{k:|\mathbb{A}_k|>1}\sum_{\ell\in\mathbb{A}_k}\sum_{|\boldsymbol{\ell}_1|+\ell_2=1}^{2\bar{r}(|\mathbb{A}_k|)}\sum_{\boldsymbol{\alpha}\in\mathbb{I}_{\boldsymbol{\ell}_1,\ell_2}} \frac{\exp(\omega_{0\ell}^N)}{2^{\alpha_4}\boldsymbol{\alpha}!}(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N)^{\boldsymbol{\alpha}_1}(\Delta\boldsymbol{a}_{\ell k}^N)^{\boldsymbol{\alpha}_2}(\Delta b_{\ell k}^N)^{\alpha_3}(\Delta\sigma_{\ell k}^N)^{\alpha_4} \\
&\times \boldsymbol{x}^{\boldsymbol{\ell}_1} \exp\left((\boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1)^\top \boldsymbol{x}\right) \cdot \frac{\partial^{\ell_2}\mathcal{N}}{\partial h_1^{\ell_2}}(y|(\boldsymbol{a}_k^0)^\top \boldsymbol{x} + b_k^0, \sigma_k^0) + R_1^N(\boldsymbol{x}, y),
\end{aligned}$$

where $\mathbb{I}_{\boldsymbol{\ell}_1,\ell_2} = \{\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \alpha_3, \alpha_4) \in \mathbb{N}^D \times \mathbb{N}^D \times \mathbb{N} \times \mathbb{N} : \ \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 = \boldsymbol{\ell}_1, \ |\boldsymbol{\alpha}_2| + \alpha_3 + 2\alpha_4 = \ell_2\}$.
Similarly, we can decompose $A_{N,2}$ by the first-order Taylor expansion as

$$\begin{aligned}
A_{N,2} = &\sum_{k:|\mathbb{A}_k|=1}\sum_{\ell\in\mathbb{A}_k}\sum_{|\boldsymbol{\ell}_1|+\ell_2=1}^{2}\sum_{\boldsymbol{\alpha}\in\mathbb{I}_{\ell_1,\ell_2}} \frac{\exp(\omega_{0\ell}^N)}{2^{\alpha_4}\boldsymbol{\alpha}!}(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N)^{\boldsymbol{\alpha}_1}(\Delta\boldsymbol{a}_{\ell k}^N)^{\boldsymbol{\alpha}_2}(\Delta b_{\ell k}^N)^{\alpha_3}(\Delta\sigma_{\ell k}^N)^{\alpha_4} \\
&\times \boldsymbol{x}^{\boldsymbol{\ell}_1} \exp\left((\boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1)^\top \boldsymbol{x}\right) \cdot \frac{\partial^{|\ell_2|}\mathcal{N}}{\partial h_1^{|\ell_2|}}(y|(\boldsymbol{a}_k^0)^\top \boldsymbol{x} + b_k^0, \sigma_k^0) + R_2^N(\boldsymbol{x}, y),
\end{aligned}$$

where

$$R_2^N(\boldsymbol{x}, y) = o\left(\sum_{k:|\mathbb{A}_k|=1}\sum_{\ell\in\mathbb{A}_k}\exp(\omega_{0\ell}^N)(\|\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N\| + \|\Delta\boldsymbol{a}_{\ell k}^N\| + \|\Delta b_{\ell k}^N\| + \|\Delta\sigma_{\ell k}^N\|)\right).$$

Analogously, B_N can be rewritten as

$$\begin{aligned}
B_N &= B_{N,1} + B_{N,2} \\
&= -\sum_{k:|\mathbb{A}_k|>1}\sum_{\ell\in\mathbb{A}_k}\sum_{|\gamma|=1}^{2} \frac{\exp(\omega_{0\ell}^N)}{\gamma!}\left(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1ik}^N\right)^\gamma \cdot \boldsymbol{x}^\gamma \exp\left(\left(\boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1\right)^\top \boldsymbol{x}\right) p_{G_N}(y|\boldsymbol{x}) + R_3^N(\boldsymbol{x}, y) \\
&\quad -\sum_{k:|\mathbb{A}_k|=1}\sum_{\ell\in\mathbb{A}_k}\sum_{|\gamma|=1} \frac{\exp(\omega_{0\ell}^N)}{\gamma!}\left(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1ik}^N\right)^\gamma \cdot \boldsymbol{x}^\gamma \exp\left(\left(\boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1\right)^\top \boldsymbol{x}\right) p_{G_N}(y|\boldsymbol{x}) + R_4^N(\boldsymbol{x}, y)
\end{aligned}$$

where

$$R_3^N(\boldsymbol{x},y) = o\left(\sum_{k:|\mathbb{A}_k|>1}\sum_{\ell\in\mathbb{A}_k}\exp(\omega_{0\ell}^N)(\|\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N\|^2)\right),$$
$$R_4^N(\boldsymbol{x},y) = o\left(\sum_{k:|\mathbb{A}_k|=1}\sum_{\ell\in\mathbb{A}_k}\exp(\omega_{0\ell}^N)(\|\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N\|)\right).$$

Therefore, Q_N can be represented as

$$\begin{aligned}
Q_N &= \sum_{k=1}^{K_0}\sum_{|\boldsymbol{\ell}_1|+\ell_2=1}^{2\bar{r}(|\mathbb{A}_k|)} T_{\boldsymbol{\ell}_1,\ell_2}^N(k)\cdot\boldsymbol{x}^{\boldsymbol{\ell}_1}\exp\left(\left(\boldsymbol{\omega}_{1k}^0+\boldsymbol{t}_1\right)^\top\boldsymbol{x}\right)\frac{\partial^{\ell_2}\mathcal{N}}{\partial h_1^{\ell_2}}\left(y|\boldsymbol{a}_k^{0\top}\boldsymbol{x}+b_k^0,\sigma_k^0\right)\\
&\quad+\sum_{k=1}^{K_0}\sum_{|\gamma|=1}^{1+\mathbf{1}_{\{|\mathbb{A}_k|>1\}}} S_\gamma^N(k)\cdot\boldsymbol{x}^\gamma\exp\left(\left(\boldsymbol{\omega}_{1k}^0+\boldsymbol{t}_1\right)^\top\boldsymbol{x}\right)p_{G_N}(y|\boldsymbol{x})+\sum_{\rho=1}^4 R_\rho^N(\boldsymbol{x},y)\\
&\quad+\sum_{k=1}^{K_0}\left(\sum_{\ell\in\mathbb{A}_k}\exp\left(\omega_{0\ell}^N\right)-\exp\left(\omega_{0k}^0+t_0\right)\right)\left[u\left(y|\boldsymbol{x};\boldsymbol{\omega}_{1k}^0+\boldsymbol{t}_1,\boldsymbol{a}_k^0,b_k^0,\sigma_k^0\right)-v\left(y|\boldsymbol{x};\boldsymbol{\omega}_{1k}^0+\boldsymbol{t}_1\right)\right]\\
&= \sum_{k=1}^{K_0}\sum_{|\boldsymbol{\ell}_1|+\ell_2=0}^{2\bar{r}(|\mathbb{A}_k|)} T_{\boldsymbol{\ell}_1,\ell_2}^N(k)\cdot\boldsymbol{x}^{\boldsymbol{\ell}_1}\exp\left(\left(\boldsymbol{\omega}_{1k}^0+\boldsymbol{t}_1\right)^\top\boldsymbol{x}\right)\frac{\partial^{\ell_2}\mathcal{N}}{\partial h_1^{\ell_2}}\left(y|\boldsymbol{a}_k^{0\top}\boldsymbol{x}+b_k^0,\sigma_k^0\right)\\
&\quad+\sum_{k=1}^{K_0}\sum_{|\gamma|=0}^{1+\mathbf{1}_{\{|\mathbb{A}_k|>1\}}} S_\gamma^N(k)\cdot\boldsymbol{x}^\gamma\exp\left(\left(\boldsymbol{\omega}_{1k}^0+\boldsymbol{t}_1\right)^\top\boldsymbol{x}\right)p_{G_N}(y|\boldsymbol{x})+\sum_{\rho=1}^4 R_\rho^N(\boldsymbol{x},y),
\end{aligned}\tag{18}$$

with coefficients $T_{\boldsymbol{\ell}_1,\ell_2}^N(k)$ and $S_\gamma^N(k)$ are defined for any $k\in[K_0]$, $0\le|\boldsymbol{\ell}_1|+\ell_2\le 2\bar{r}\left(|\mathbb{A}_k|\right)$ and $0\le|\gamma|\le 2$ as

$$T_{\boldsymbol{\ell}_1,\ell_2}^N(k) = \begin{cases}\sum_{\ell\in\mathbb{A}_k}\sum_{\boldsymbol{\alpha}\in\mathbb{I}_{\ell_1,\ell_2}}\frac{\exp\left(\omega_{0\ell}^N\right)}{2^{\alpha_4}\boldsymbol{\alpha}!}\left(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N\right)^{\boldsymbol{\alpha}_1}\left(\Delta\boldsymbol{a}_{\ell k}^N\right)^{\boldsymbol{\alpha}_2}\left(\Delta b_{\ell k}^N\right)^{\alpha_3}\left(\Delta\sigma_{\ell k}^N\right)^{\alpha_4}, & (\boldsymbol{\ell}_1,\ell_2)\neq(0_D,0),\\ \sum_{\ell\in\mathbb{A}_k}\exp(\omega_{0\ell}^N)-\exp(\omega_{0k}^0+t_0), & (\boldsymbol{\ell}_1,\ell_2)=(0_D,0),\end{cases}$$
$$S_\gamma^N(k) = \begin{cases}-\sum_{\ell\in\mathbb{A}_k}\frac{\exp\left(\omega_{0\ell}^N\right)}{\gamma!}\left(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N\right)^\gamma, & |\gamma|\neq 0,\\ -\sum_{\ell\in\mathbb{A}_k}\exp(\omega_{0\ell}^N)+\exp(\omega_{0k}^0+t_0), & |\gamma|=0.\end{cases}$$

Step 2: Non-vanishing coefficients

Next, we will show that not all the quatities $T_{\boldsymbol{\ell}_1,\ell_2}^N(k)/\operatorname{D_V}(G_N,G_0)$ and $S_\gamma^N(k)/\operatorname{D_V}(G_N,G_0)$ go to 0 as $N\to\infty$. We assume that all of them go to 0 as $N\to\infty$. Then, by assumption $T_{0_D,0}^N(k)/\operatorname{D_V}(G_N,G_0)\to 0$, we have

$$\frac{1}{\operatorname{D_V}(G_N,G_0)}\sum_{k=1}^{K_0}\left|\sum_{\ell\in\mathbb{A}_k}\exp(\omega_{0\ell}^N)-\exp(\omega_{0k}^0+t_0)\right|\to 0.\tag{19}$$

For any k such that $|\mathbb{A}_k|=1$, consider all $(|\boldsymbol{\ell}_1|,\ell_2)$ implying $1\le|\boldsymbol{\ell}_1|+\ell_2\le 2$, we have $T_{\boldsymbol{\ell}_1,\ell_2}^N(k)/\operatorname{D_V}(G_N,G_0)\to 0$ for all k such that $|\mathbb{A}_k|=1$. Hence

$$\frac{1}{\operatorname{D_V}(G_N,G_0)}\left(\sum_{k:|\mathbb{A}_k|=1}\sum_{\ell\in\mathbb{A}_k}\exp(\omega_{0\ell}^N)\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1\ell k}^N,\Delta\boldsymbol{a}_{\ell k}^N,\Delta b_{\ell k}^N,\Delta\sigma_{\ell k}^N)\|\right)\to 0.\tag{20}$$

Next, we consider k such that $|\mathbb{A}_k|>1$ and $(|\boldsymbol{\ell}_1|,\ell_2)$ such that $1\le|\boldsymbol{\ell}_1|+\ell_2\le 2$:

- For $(|\boldsymbol{\ell}_1|,\ell_2)=(0,1)$, then
$$\frac{1}{\operatorname{D_V}(G_N,G_0)}\|\sum_{\ell\in\mathbb{A}_k}\exp\left(\omega_{0\ell}^N\right)(\Delta b_{\ell k}^N)\|\to 0.$$

- For $(|\boldsymbol{\ell}_1|, \ell_2) = (1, 0)$, then
$$\frac{1}{\mathrm{D_V}(G_N, G_0)} \| \sum_{\ell \in \mathbb{A}_k} \exp\left(\omega_{0\ell}^N\right) (\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N) \| \to 0.$$

- For $(|\boldsymbol{\ell}_1|, \ell_2) = (1, 1)$, then
$$\frac{1}{\mathrm{D_V}(G_N, G_0)} \| \sum_{\ell \in \mathbb{A}_k} \exp\left(\omega_{0\ell}^N\right) [(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N)(\Delta b_{\ell k}^N) + (\Delta \boldsymbol{a}_{\ell k}^N)] \| \to 0.$$

- For $(|\boldsymbol{\ell}_1|, \ell_2) = (0, 2)$, then
$$\frac{1}{\mathrm{D_V}(G_N, G_0)} \| \sum_{\ell \in \mathbb{A}_k} \exp\left(\omega_{0\ell}^N\right) [(\Delta b_{\ell k}^N)^2 + (\Delta \sigma_{\ell k}^N)] \| \to 0.$$

- For $(|\boldsymbol{\ell}_1|, \ell_2) = (2, 0)$, then
$$\frac{1}{\mathrm{D_V}(G_N, G_0)} \| \sum_{\ell \in \mathbb{A}_k} \exp\left(\omega_{0\ell}^N\right) (\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N)(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N)^\top \| \to 0.$$

Combining the above limit and the formulation of $\mathrm{D_{FRA}}(G_N, G_0)$ together, it follows that
$$\frac{1}{\mathrm{D_V}(G_N, G_0)} \cdot \sum_{k:|\mathbb{A}_k|>1} \sum_{\ell \in \mathbb{A}_k} \exp\left(\omega_{0\ell}\right) \left(\left\| \left(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell k}^N, \Delta b_{\ell k}^N\right) \right\|^{\bar{r}(|\mathbb{A}_k|)} + \left\| \left(\Delta \boldsymbol{a}_{\ell k}^N, \Delta \sigma_{\ell k}^N\right) \right\|^{\bar{r}(|\mathbb{A}_k|)/2} \right) \nrightarrow 0$$

which implies that there exists some index $k^* \in [K_0]$ such that $|\mathbb{A}_{k^*}| > 1$ and
$$\frac{1}{\mathrm{D_V}(G_N, G_0)} \cdot \sum_{\ell \in \mathbb{A}_{k^*}} \exp\left(\omega_{0\ell}\right) \left(\left\| \left(\Delta_{vt_1} \boldsymbol{\omega}_{1ellk^*}^N, \Delta b_{\ell k^*}^N\right) \right\|^{\bar{r}(|\mathbb{A}_k|)} + \left\| \left(\Delta \boldsymbol{a}_{\ell k^*}^N, \Delta \sigma_{\ell k^*}^N\right) \right\|^{\bar{r}(|\mathbb{A}_k|)/2} \right) \nrightarrow 0$$

for all $\boldsymbol{t}_1 \in \mathbb{R}^D$. WLOG, we assume that $k^* = 1$. For $(\boldsymbol{\ell}_1, \ell_2) \in \mathbb{N}^D \times \mathbb{N}$ such that $1 \leq |\boldsymbol{\ell}_1| + \ell_2 \leq \bar{r}\left(|\mathbb{A}_1|\right)$, we have $T_{\boldsymbol{\ell}_1, \ell_2}^N(1) / \mathrm{D_V}(G_N, G_0) \to 0$ as $N \to \infty$. Thus, by dividing this ratio and the left hand side of the above equation and let $\boldsymbol{t}_1 = 0$, we have
$$\frac{\sum_{\ell \in \mathbb{A}_1} \sum_{\boldsymbol{\alpha} \in \mathbb{I}_{\boldsymbol{\ell}_1, \ell_2}} \frac{\exp\left(\omega_{0\ell}^N\right)}{2^{\alpha_4} \alpha!} \left(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell 1}^N\right)^{\boldsymbol{\alpha}_1} \left(\Delta \boldsymbol{a}_{\ell 1}^N\right)^{\boldsymbol{\alpha}_2} \left(\Delta b_{\ell 1}^N\right)^{\alpha_3} \left(\Delta \sigma_{\ell 1}^N\right)^{\alpha_4}}{\sum_{\ell \in \mathbb{A}_1} \exp\left(\omega_{0\ell}^N\right) \left(\left\| \left(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell 1}^N, \Delta b_{\ell 1}^N\right) \right\|^{\bar{r}(|\mathbb{A}_1|)} + \left\| \left(\Delta \boldsymbol{a}_{\ell 1}^N, \Delta \sigma_{\ell 1}^N\right) \right\|^{\bar{r}(|\mathbb{A}_1|)/2} \right)} \to 0 \tag{21}$$

for all $(\boldsymbol{\ell}_1, \ell_2)$ such that $1 \leq |\boldsymbol{\ell}_1| + \ell_2 \leq \bar{r}\left(|\mathbb{A}_1|\right)$.
Let us define $\bar{M}_N := \max\left\{ \left\|\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell 1}^N\right\|, \left\|\Delta \boldsymbol{a}_{\ell 1}^N\right\|^{1/2}, \left|\Delta b_{\ell 1}^N\right|, \left|\Delta \sigma_{\ell 1}^N\right|^{1/2} : \ell \in \mathbb{A}_1 \right\}$ and $\bar{\omega}_N := \max_{\ell \in \mathbb{A}_1} \exp\left(\omega_{0\ell}^N\right)$. Since the sequence $\exp\left(\omega_{0\ell}^N\right) / \bar{\omega}_N$ is bounded, we can replace it by its subsequence that has a positive limit $p_{5\ell}^2 := \lim_{N \to \infty} \exp\left(\omega_{0\ell}^N\right) / \bar{\omega}_N$. Hence, at least one among $p_{5\ell}^2$, for $\ell \in \mathbb{A}_1$, equals 1.

Similarly, we also define
$$\begin{aligned} \left(\Delta_{\boldsymbol{t}_1} \boldsymbol{\omega}_{1\ell 1}^N\right) / \bar{M}_N \to p_{1\ell}, \left(\Delta \boldsymbol{a}_{\ell 1}^N\right) / \bar{M}_N \to p_{2\ell}, \\ \left(\Delta b_{\ell 1}^N\right) / \bar{M}_N \to p_{3\ell}, \left(\Delta \sigma_{\ell 1}^N\right) / \left[2 \bar{M}_N\right] \to p_{4\ell}. \end{aligned}$$

Here, at least one of $p_{1\ell}, p_{2\ell}, p_{3\ell}$ and $p_{4\ell}$ for $\ell \in \mathbb{A}_1$ equals either 1 or -1 . Next, we divide both the numerator and the denominator of the ratio in eq. (21) by $\bar{\omega}_N \bar{M}_N^{\boldsymbol{\ell}_1 + \ell_2}$, and then achieve the following system of polynomial equations:
$$\sum_{\ell \in \mathbb{A}_1} \sum_{\boldsymbol{\alpha} \in \mathbb{I}_{\boldsymbol{\ell}_1, \ell_2}} \frac{1}{\alpha!} \cdot p_{5\ell}^2 p_{1\ell}^{\boldsymbol{\alpha}_1} p_{2\ell}^{\boldsymbol{\alpha}_2} p_{3\ell}^{\alpha_3} p_{4\ell}^{\alpha_4} = 0$$

for all $(\boldsymbol{\ell}_1, \ell_2) \in \mathbb{N}^D \times \mathbb{N}$ such that $1 \leq |\boldsymbol{\ell}_1| + \ell_2 \leq \bar{r}\left(|\mathbb{A}_1|\right)$. However, based on the definition of $\bar{r}\left(|\mathbb{A}_1|\right)$, the above system has no non-trivial solutions, which is a contradiction. Thus, not all the quantities $T^N_{\boldsymbol{\ell}_1,\ell_2}(k)/\,\mathrm{D_V}(G_N, G_0)$ and $S^N_\gamma(k)/\,\mathrm{D_V}(G_N, G_0)$ go to 0 as $N \to \infty$.

Step 3: Fatou's lemma involvement

Following this, we define by m_N be the maximum of the absolute values of those quantities. Based on the result in Step 2, we know that $1/m_N \nrightarrow \infty$. Then, by applying the Fatou's lemma, we obtain that

$$\lim_{N\to\infty} \frac{\mathbb{E}_{\mathbf{x}}\left[\,\mathrm{D_{TV}}(p_G(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x}))\right]}{m_N \cdot \mathrm{D_V}(G_N, G_0)} \geq \int \liminf_{N\to\infty} \frac{|p_G(y|\boldsymbol{x}), p_{G_0}(y|\boldsymbol{x})|}{2m_N \cdot \mathrm{D_V}(G_N, G_0)}\,\mathrm{d}(\boldsymbol{x}, y). \tag{22}$$

By assumption, the left-hand side of eq. (22) equals to 0, so the integrand in the right-hand side also equals to 0 for almost surely $(\boldsymbol{x}, y)$. Hence, we get that $Q_N/\left[m_N\,\mathrm{D_V}(G_N, G_0)\right] \to 0$ as $N \to \infty$ for almost surely $(\boldsymbol{x}, y)$. It follows from the decomposition of Q_N in eq. (18) that

$$\begin{aligned}\sum_{k=1}^{K_0}\sum_{|\boldsymbol{\ell}_1|+\ell_2=0}^{2\bar{r}(|\mathbb{A}_k|)} \tau_{\boldsymbol{\ell}_1,\ell_2}(k)\cdot \boldsymbol{x}^{\boldsymbol{\ell}_1} \exp\left(\left(\boldsymbol{\omega}^0_{1k}+\boldsymbol{t}_1\right)^\top \boldsymbol{x}\right) \frac{\partial^{\ell_2}\mathcal{N}}{\partial h_1^{\ell_2}}\left(y|\left(\boldsymbol{a}^0_k\right)^\top \boldsymbol{x} + b^0_k, \sigma^0_k\right)\\ + \sum_{k=1}^{K_0}\sum_{|\gamma|=0}^{1+\mathbf{1}_{\{|\mathbb{A}_k|>1\}}} \xi_\gamma(j)\cdot \boldsymbol{x}^\gamma \exp\left(\left(\boldsymbol{\omega}^0_{1k}+\boldsymbol{t}_1\right)^\top \boldsymbol{x}\right) p_{G_0}(y|\boldsymbol{x}) = 0,\end{aligned}$$

for almost surely $(\boldsymbol{x}, y)$, where $\tau_{\boldsymbol{\ell}_1,\ell_2}(k)$ and $\xi_\gamma(k)$ denote the limits of $T^N_{\boldsymbol{\ell}_1,\ell_2}(k)/\left[m_N\,\mathrm{D_V}(G_N, G_0)\right]$ and $S^N_\gamma(j)/\left[m_N\,\mathrm{D_V}(G_N, G_0)\right]$ as $N \to \infty$, respectively, for all $k \in [K_0], 0 \leq 2\,|\boldsymbol{\ell}_1| + \ell_2 \leq 2\bar{r}\left(|\mathbb{A}_k|\right)$ and $0 \leq |\gamma| \leq 1 + \mathbf{1}_{\{|\mathbb{A}_k|>1\}}$. By definition, at least one among $\tau_{\ell_1,\ell_2}(k)$ and $\xi_\gamma(k)$ is different from zero.

Furthermore, we denote the set $\mathcal{W}$ as follows:

$$\begin{aligned}\mathcal{W} := &\left\{\boldsymbol{x}^{\boldsymbol{\ell}_1}\exp\left(\left(\boldsymbol{\omega}^0_{1k}+\boldsymbol{t}_1\right)^\top \boldsymbol{x}\right)\frac{\partial^{\ell_2}\mathcal{N}}{\partial h_1^{\ell_2}}\left(y|\left(\boldsymbol{a}^0_k\right)^\top \boldsymbol{x}+b^0_k,\sigma^0_k\right) : k\in[K_0], 0\leq|\boldsymbol{\ell}_1|+\ell_2\leq 2\bar{r}\left(|\mathbb{A}_k|\right)\right\}\\ &\cup\left\{\boldsymbol{x}^\gamma\exp\left(\left(\boldsymbol{\omega}^0_{1k}+\boldsymbol{t}_1\right)^\top \boldsymbol{x}\right)p_{G_0}(y|\boldsymbol{x}) : k\in[K_0], 0\leq|\gamma|\leq 1+\mathbf{1}_{\{|\mathbb{A}_k|>1\}}\right\}.\end{aligned}$$

Similarly to the proof of Fact 5 in Nguyen et al., 2023a:

Fact 5 (Nguyen et al., 2023a, Lemma 2)**.** *The set $\mathcal{W}_1$ is linearly indeqendent w.r.t $\boldsymbol{x}$ and y, where $\mathcal{W}_1$ is denoted as follows:*

$$\begin{aligned}\mathcal{W}_1 := &\left\{\boldsymbol{x}^{\boldsymbol{\ell}_1}\exp\left(\left(\boldsymbol{\omega}^0_{1k}+\boldsymbol{t}_1\right)^\top \boldsymbol{x}\right)\frac{\partial^{\ell_2}\mathcal{N}}{\partial h_1^{\ell_2}}\left(y|\left(\boldsymbol{a}^0_k\right)^\top \boldsymbol{x}+b^0_k,\sigma^0_k\right) : k\in[K_0], 0\leq|\boldsymbol{\ell}_1|+\ell_2\leq 2\right\}\\ &\cup\left\{\boldsymbol{x}^\gamma\exp\left(\left(\boldsymbol{\omega}^0_{1k}+\boldsymbol{t}_1\right)^\top \boldsymbol{x}\right)p_{G_0}(y|\boldsymbol{x}) : k\in[K_0], 0\leq|\gamma|\leq 1\right\},\end{aligned}$$

the set $\mathcal{W}$ is linearly independent w.r.t $\boldsymbol{x}$ and y, it follows that

$$\tau_{\boldsymbol{\ell}_1,\ell_2}(k) = \xi_\gamma(k) = 0$$

for all $k \in [K_0], 0 \leq 2\,|\ell_1| + \ell_2 \leq 2\bar{r}\left(|\mathbb{A}_k|\right)$ and $0 \leq |\gamma| \leq 1 + \mathbf{1}_{\{|\mathbb{A}_k|>1\}}$, which is a contradiction. Hence, we achieve the eq. (17).

Global version: Hence, it is sufficient to prove its following global inequality:

$$\inf_{G\in\mathcal{O}_K:\mathrm{D_{FRA}}(G,G_0)>\varepsilon'} \mathbb{E}_{\mathbf{x}}\left[\,\mathrm{D_{TV}}(p_G(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x}))\right]/\,\mathrm{D_{FRA}}(G, G_0) > 0. \tag{23}$$

Assume by contrary that there exists a sequence $G'_N \in \mathcal{O}_K$ that satisfies

$$\begin{cases}\lim_{N\to\infty}\mathbb{E}_{\mathbf{x}}\left[\,\mathrm{D_{TV}}(p_{G'_N}(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x}))\right]/\,\mathrm{D_{FRA}}(G'_N, G_0) = 0,\\ \mathrm{D_{FRA}}(G'_N, G_0) > \varepsilon'.\end{cases}$$

Then, we get that $\mathbb{E}_{\mathbf{x}}\big[\mathrm{D_{TV}}(p_{G'_N}(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x}))\big] \to 0$ as $N \to \infty$. Since the set $\boldsymbol{\Theta}$ is compact, we can replace the sequence G'_N by its subsequence which converges to some mixing measure $G' \in \mathcal{O}_K$ such that $\mathrm{D_{FRA}}\,(G', G_0) > \varepsilon'$. Then, by the Fatou's lemma, we get

$$\lim_{N\to\infty} \mathbb{E}_{\mathbf{x}}\big[\mathrm{D_{TV}}(p_{G'_N}(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x}))\big] \geq \frac{1}{2}\int \liminf_{N\to\infty} |p_{G'_N}(y|\boldsymbol{x}) - p_{G_0}(y|\boldsymbol{x})|\mathrm{d}(\boldsymbol{x}, y).$$

It follows that

$$\int |p_{G'}(y|\boldsymbol{x}) - p_{G_0}(y|\boldsymbol{x})|\,\mathrm{d}(\boldsymbol{x}, y) = 0.$$

Thus, we obtain that $p_{G'}(y|\boldsymbol{x}) = p_{G_0}(y|\boldsymbol{x})$ for almost surely $(\boldsymbol{x}, y)$. By Fact 3, the mixing measure G' admits the form $G' = \sum_{k=1}^{K_0} \exp\left(\omega^0_{0\nu(k)} + t_0\right)\delta_{\left(\boldsymbol{\omega}^0_{1\nu(k)}+\boldsymbol{t}_1, \boldsymbol{a}^0_{\nu(k)}, b^0_{\nu(k)}, \sigma^0_{\nu(k)}\right)}$ for some $(t_0, \boldsymbol{t}_1) \in \mathbb{R}\times\mathbb{R}^D$, where ν is some permutation of the set $\{1, 2, \ldots, K_0\}$. It follows that $\mathrm{D_{FRA}}\,(G', G_0) = 0$, which contradicts the hypothesis $\mathrm{D_{FRA}}\,(G', G_0) > \varepsilon' > 0$. Hence, we obtain the inequality in eq. (16). □

Next, assume that $\widehat{G}_N \in \mathcal{E}_K$ with $K > K_0$. From Fact 4, there exists a constant $c(\boldsymbol{\Theta}, K)$ depending on $\boldsymbol{\Theta}$ and K so that on an event, we call A_N, with probability at least $1 - CN^{-c}$, we have

$$\mathbb{E}_{\mathbf{x}}[\mathrm{D_{TV}}(p_{\widehat{G}_N}(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x}))] \leq \sqrt{2}\mathbb{E}_{\mathbf{x}}[\mathrm{D^2_h}(p_{\widehat{G}_N}(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x}))] \leq c(\boldsymbol{\Theta}, K)\cdot\left(\frac{\log N}{N}\right)^{1/2}.$$

Now, we prove Theorem 1.

Proof of Theorem 1. Firstly, we prove for the over-specified case. By Lemma 1 and Theorem 5, we have the first statement.

To prove the rest, we need to consider the exact-specified case. When $\kappa' = K_0$, by definition of $\mathrm{D_{FRA}}(\widehat{G}_N^{(\kappa')}, G_0)$, we obtain that $\mathrm{D_{FRA}}(\widehat{G}_N^{(K_0)}, G_0) = \mathrm{D_E}(\widehat{G}_N^{(K_0)}, G_0)$. Hence, by Lemma 1, we get the convergence rate

$$\mathrm{D_E}(\widehat{G}_N^{(K_0)}, G_0) \lesssim \left(\frac{\log N}{N}\right)^{1/2}.$$

Assume that $\widehat{G}_N^{(K_0)} = \sum_{k=1}^{K_0}\exp(\omega^N_{0k})\delta_{(\boldsymbol{\omega}^N_{1k}, \boldsymbol{a}^N_k, b^N_k, \sigma^N_k)} \in \mathcal{E}_{K_0}$. Building on our previous work, there exist $t_0 \in \mathbb{R}$ and $\boldsymbol{t}_1 \in \mathbb{R}^D$ such that for large N enough, we get

$$|\exp(\omega^N_{0k}) - \exp(\omega^0_{0k} + t_0)| \lesssim \left(\frac{\log N}{N}\right)^{1/2},$$
$$\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}^N_{1k}, \Delta\boldsymbol{a}^N_k, \Delta b^N_k, \Delta\sigma^N_k)\| \lesssim \left(\frac{\log N}{N}\right)^{1/2},$$

for every $k \in [K_0]$, where $\Delta^N_{\boldsymbol{t}_1}\boldsymbol{\omega}^N_{1k} := \boldsymbol{\omega}^N_{1k} - \boldsymbol{\omega}^0_{1k} - \boldsymbol{t}_1$, $\Delta\boldsymbol{a}^N_k := \boldsymbol{a}^N_k - \boldsymbol{a}^0_k$, $\Delta b^N_k := b^N_k - b^0_k$ and $\Delta\sigma^N_k := \sigma^N_k - \sigma^0_k$. This implies that for every $(i, j) \in [K_0]^2$, by the triangle inequality, we have

$$\begin{aligned}\Big|\|(\boldsymbol{\omega}^N_{1i} - \boldsymbol{\omega}^N_{1j}, b^N_i - b^N_j)\| - \|(\boldsymbol{\omega}^0_{1i} - \boldsymbol{\omega}^0_{1j}, b^0_i - b^0_j)\|\Big| &\leq \|(\boldsymbol{\omega}^N_{1i} - \boldsymbol{\omega}^N_{1j} - \boldsymbol{\omega}^0_{1i} + \boldsymbol{\omega}^0_{1j}, b^N_i - b^N_j - b^0_i + b^0_j)\| \\ &\leq \|(\boldsymbol{\omega}^N_{1i} - \boldsymbol{\omega}^0_{1i} - \boldsymbol{t}_1, b^N_i - b^0_i)\| + \|(\boldsymbol{\omega}^N_{1j} - \boldsymbol{\omega}^0_{1j} - \boldsymbol{t}_1, b^N_j - b^0_j)\| \\ &\lesssim \left(\frac{\log N}{N}\right)^{1/2}.\end{aligned}$$

Similarly, we have

$$\Big|\|(\boldsymbol{a}^N_i - \boldsymbol{a}^N_j, \sigma^N_i - \sigma^N_j)\| - \|(\boldsymbol{a}^0_i - \boldsymbol{a}^0_j, \sigma^0_i - \sigma^0_j)\|\Big| \lesssim \left(\frac{\log N}{N}\right)^{1/2}.$$

Hence, we obtain that

$$\begin{aligned}
&\left| \frac{1}{\exp\left(-\omega_{0i}^N\right) + \exp\left(-\omega_{0j}^N\right)} \left(\|(\boldsymbol{\omega}_{1i}^N - \boldsymbol{\omega}_{1j}^N, b_i^N - b_j^N)\|^2 + \|(\boldsymbol{a}_i^N - \boldsymbol{a}_j^N, \sigma_i^N - \sigma_j^N)\| \right) \right. \\
&\left. - \frac{1}{\exp\left(-\omega_{0i}^0 - t_0\right) + \exp\left(-\omega_{0j}^0 - t_0\right)} \left(\|(\boldsymbol{\omega}_{1i}^0 - \boldsymbol{\omega}_{1j}^0, b_i^0 - b_j^0)\|^2 + \|(\boldsymbol{a}_i^0 - \boldsymbol{a}_j^0, \sigma_i^0 - \sigma_j^0)\| \right) \right| \\
&\lesssim \left(\frac{\log N}{N} \right)^{1/2}, \quad \forall (i,j) \in [K_0]^2. \qquad (24)
\end{aligned}$$

Hence, on A_N, the optimal choice of indices (ℓ_1, ℓ_2) to merge for $\widehat{G}_N^{(K_0)}$ will be the same as G_0 for every N large enough. It follows that we have two merged atoms are $\exp(\omega_{0*}^N)\delta_{(\boldsymbol{\omega}_{1*}^N, \boldsymbol{a}_*^N, b_*^N, \sigma_*^N)}$ and $\exp(\omega_{0*}^0)\delta_{(\boldsymbol{\omega}_{1*}^0, \boldsymbol{a}_*^0, b_*^0, \sigma_*^0)}$ denoted as follows:

$$\begin{aligned}
\omega_{0*}^N &= \log\left(\exp\omega_{0\ell_1}^N + \exp\omega_{0\ell_2}^N\right), \\
\boldsymbol{\omega}_{1*}^N &= \exp\left(\omega_{0\ell_1}^N - \omega_{0*}^N\right)\boldsymbol{\omega}_{1\ell_1}^N + \exp\left(\omega_{0\ell_2}^N - \omega_{0*}^N\right)\boldsymbol{\omega}_{1\ell_2}^N, \\
b_*^N &= \exp\left(\omega_{0\ell_1}^N - \omega_{0*}^N\right)b_{\ell_1}^N + \exp\left(\omega_{0\ell_2}^N - \omega_{0*}^N\right)b_{\ell_2}^N, \\
\boldsymbol{a}_*^N &= \frac{\exp(\omega_{0\ell_1}^N)}{\exp(\omega_{0*}^N)}\left[(\boldsymbol{\omega}_{1\ell_1}^N - \boldsymbol{\omega}_{1*}^N)(b_{\ell_1}^N - b_*^N) + \boldsymbol{a}_{\ell_1}^N\right] + \frac{\exp(\omega_{0\ell_2}^N)}{\exp(\omega_{0*}^N)}\left[(\boldsymbol{\omega}_{1\ell_2}^N - \boldsymbol{\omega}_{1*}^N)(b_{\ell_2}^N - b_*^N) + \boldsymbol{a}_{\ell_2}^N\right], \\
\sigma_*^N &= \frac{\exp(\omega_{0\ell_1}^N)}{\exp(\omega_{0*}^N)}\left[(b_{\ell_1}^N - b_*^N)^2 + \sigma_{\ell_1}^N\right] + \frac{\exp(\omega_{0\ell_2}^N)}{\exp(\omega_{0*}^N)}\left[(b_{\ell_2}^N - b_*^N)^2 + \sigma_{\ell_2}^N\right],
\end{aligned}$$

and

$$\begin{aligned}
\omega_{0*}^0 &= \log\left(\exp\omega_{0\ell_1}^0 + \exp\omega_{0\ell_2}^0\right), \\
\boldsymbol{\omega}_{1*}^0 &= \exp\left(\omega_{0\ell_1}^0 - \omega_{0*}^0\right)\boldsymbol{\omega}_{1\ell_1}^0 + \exp\left(\omega_{0\ell_2}^0 - \omega_{0*}^0\right)\boldsymbol{\omega}_{1\ell_2}^0, \\
b_*^0 &= \exp\left(\omega_{0\ell_1}^0 - \omega_{0*}^0\right)b_{\ell_1}^0 + \exp\left(\omega_{0\ell_2}^0 - \omega_{0*}^0\right)b_{\ell_2}^0, \\
\boldsymbol{a}_*^0 &= \frac{\exp(\omega_{0\ell_1}^0)}{\exp(\omega_{0*}^0)}\left[(\boldsymbol{\omega}_{1\ell_1}^0 - \boldsymbol{\omega}_{1*}^0)(b_{\ell_1}^0 - b_*^0) + \boldsymbol{a}_{\ell_1}^0\right] + \frac{\exp(\omega_{0\ell_2}^0)}{\exp(\omega_{0*}^0)}\left[(\boldsymbol{\omega}_{1\ell_2}^0 - \boldsymbol{\omega}_{1*}^0)(b_{\ell_2}^0 - b_*^0) + \boldsymbol{a}_{\ell_2}^0\right], \\
\sigma_*^0 &= \frac{\exp(\omega_{0\ell_1}^0)}{\exp(\omega_{0*}^0)}\left[(b_{\ell_1}^0 - b_*^0)^2 + \sigma_{\ell_1}^0\right] + \frac{\exp(\omega_{0\ell_2}^0)}{\exp(\omega_{0*}^0)}\left[(b_{\ell_2}^0 - b_*^0)^2 + \sigma_{\ell_2}^0\right].
\end{aligned}$$

After merging, we also have

$$\begin{aligned}
|\exp(\omega_{0*}^N) - \exp(\omega_{0*}^0 + t_0)| &= |\exp(\omega_{0\ell_1}^N) + \exp(\omega_{0\ell_2}^N) - \exp(\omega_{0\ell_1}^0 + t_0) - \exp(\omega_{0\ell_1}^0 + t_0)| \\
&\le |\exp(\omega_{0\ell_1}^N) - \exp(\omega_{0\ell_1}^0 + t_0)| + |\exp(\omega_{0\ell_1}^N) - \exp(\omega_{0\ell_1}^0 + t_0)| \\
&\lesssim \left(\frac{\log N}{N} \right)^{1/2}, \qquad (25)
\end{aligned}$$

and

$$\begin{aligned}
&\exp(\omega_{0*}^N)\|(\Delta_{\boldsymbol{t}_1}^N \boldsymbol{\omega}_{1*}^N, \Delta \boldsymbol{a}_*^N, \Delta b_*^N, \Delta \sigma_*^N)\| \\
&\quad \le \exp(\omega_{0*}^N) \times \exp\left(\omega_{0\ell_1}^N - \omega_{0*}^N\right)\|(\Delta_{\boldsymbol{t}_1}^N \boldsymbol{\omega}_{1\ell_1}^N, \Delta \boldsymbol{a}_{\ell_1}^N, \Delta b_{\ell_1}^N, \Delta \sigma_{\ell_1}^N)\| \\
&\qquad + \exp(\omega_{0*}^N) \times \exp\left(\omega_{0\ell_2}^N - \omega_{0*}^N\right)\|(\Delta_{\boldsymbol{t}_1}^N \boldsymbol{\omega}_{1\ell_2}^N, \Delta \boldsymbol{a}_{\ell_2}^N, \Delta b_{\ell_2}^N, \Delta \sigma_{\ell_2}^N)\| \\
&\quad \lesssim \left(\frac{\log N}{N} \right)^{1/2}.
\end{aligned}$$

Hence, $\mathrm{D_E}(\widehat{G}_N^{(K_0-1)}, G_0^{(K_0-1)}) \lesssim \left(\frac{\log N}{N}\right)^{1/2}$. By the induction, we have the rest statement. □

F.3 Proof of Theorem 2

For the convergence rate of the height at all levels $\kappa \geq K_0 + 1$, from Theorem 1, we have

$$\mathrm{D}_{\mathrm{FRA}}(\widehat{G}_N^{(\kappa)}, G_0) \lesssim \left(\frac{\log N}{N}\right)^{1/2}.$$

Because $\kappa \geq K_0 + 1$, by the pigeonhole principle, there exists at least two $i, j \in [\kappa]$ such that two atoms $\exp(\omega_{0i}^N)\delta_{(\boldsymbol{\omega}_{1i}^N, \boldsymbol{a}_i^N, b_i^N, \sigma_i^N)}$ and $\exp(\omega_{0j}^N)\delta_{(\boldsymbol{\omega}_{1j}^N, \boldsymbol{a}_j^N, b_j^N, \sigma_j^N)}$ belongs to a common Voronoi cell of some $\boldsymbol{\theta}_k^0$ (we suppress the dependence of i, j, and $\mathbb{A}_k$ on N for ease of notation). Hence,

$$\begin{aligned}
&\inf_{\boldsymbol{t}_1} \exp(\omega_{0i}) \left(\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1ik}, \Delta b_{ik})\|^{\bar{r}(|\mathbb{A}_k|)} + \|(\Delta \boldsymbol{a}_{ik}, \Delta\sigma_{ik})\|^{\bar{r}(|\mathbb{A}_k|)/2}\right) \\
&\quad + \exp(\omega_{0j}) \left(\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1jk}, \Delta b_{jk})\|^{\bar{r}(|\mathbb{A}_k|)} + \|(\Delta \boldsymbol{a}_{jk}, \Delta\sigma_{jk})\|^{\bar{r}(|\mathbb{A}_k|)/2}\right) \lesssim \left(\frac{\log N}{N}\right)^{1/2}.
\end{aligned}$$

Using the fact that $\min\{\exp(\omega_{0i}), \exp(\omega_{0j})\} \geq \dfrac{1}{\exp(-\omega_{0i}) + \exp(-\omega_{0j})}$, $\bar{r}(\widehat{G}_N) \geq \bar{r}(|\mathbb{A}_k|) \geq \bar{r}(2) = 4$, and using the Hölder's inequality, for every $\boldsymbol{t}_1 \in \mathbb{R}^D$ we have

$$\begin{aligned}
&\exp(\omega_{0i}) \left(\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1ik}, \Delta b_{ik})\|^{\bar{r}(|\mathbb{A}_k|)} + \|(\Delta \boldsymbol{a}_{ik}, \Delta\sigma_{ik})\|^{\bar{r}(|\mathbb{A}_k|)/2}\right) \\
&\quad + \exp(\omega_{0j}) \left(\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1jk}, \Delta b_{jk})\|^{\bar{r}(|\mathbb{A}_k|)} + \|(\Delta \boldsymbol{a}_{jk}, \Delta\sigma_{jk})\|^{\bar{r}(|\mathbb{A}_k|)/2}\right) \\
&\geq \frac{1}{\exp(-\omega_{0i}) + \exp(-\omega_{0j})} \Bigg[\left(\|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1ik}, \Delta b_{ik})\|^{\bar{r}(|\mathbb{A}_k|)} + \|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1jk}, \Delta b_{jk})\|^{\bar{r}(|\mathbb{A}_k|)}\right) \\
&\quad + \left(\|(\Delta \boldsymbol{a}_{ik}, \Delta\sigma_{ik})\|^{\bar{r}(|\mathbb{A}_k|)/2} + \|(\Delta \boldsymbol{a}_{jk}, \Delta\sigma_{jk})\|^{\bar{r}(|\mathbb{A}_k|)/2}\right) \Bigg] \\
&\gtrsim \frac{1}{\exp(-\omega_{0i}) + \exp(-\omega_{0j})} \left(\|(\boldsymbol{\omega}_{1i} - \boldsymbol{\omega}_{1j}, b_i - b_j)\|^{\bar{r}(|\mathbb{A}_k|)} + \|(\boldsymbol{a}_i - \boldsymbol{a}_j, \sigma_i - \sigma_j)\|^{\bar{r}(|\mathbb{A}_k|)/2}\right) \\
&\gtrsim \left(\frac{1}{\exp(-\omega_{0i}) + \exp(-\omega_{0j})} \left(\|(\boldsymbol{\omega}_{1i} - \boldsymbol{\omega}_{1j}, b_i - b_j)\|^2 + \|(\boldsymbol{a}_i - \boldsymbol{a}_j, \sigma_i - \sigma_j)\|\right)\right)^{\bar{r}(\widehat{G}_N)/2}.
\end{aligned}$$

Since the height of the dendrogram is the minimum of d over all pairs (i, j), we obtain that

$$\mathsf{h}_N^{(\kappa)} \lesssim \frac{1}{\exp(-\omega_{0i}) + \exp(-\omega_{0j})} \left(\|(\boldsymbol{\omega}_{1i} - \boldsymbol{\omega}_{1j}, b_i - b_j)\|^2 + \|(\boldsymbol{a}_i - \boldsymbol{a}_j, \sigma_i - \sigma_j)\|\right) \lesssim \left(\frac{\log N}{N}\right)^{1/\bar{r}(\widehat{G}_N)},$$

for all $\kappa \geq K_0 + 1$.

When $\kappa \leq K_0$, the conclusion follows from inequality in eq. (24) in the proof of Theorem 1.

F.4 Proof of Theorem 3

Before we prove Theorem 3, we revisit preliminary on empirical process theory and connection between the Hellinger distance and the Wasserstein metric.

Preliminary on Empirical Process Theory. Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_N \sim P_{G_0}$. Denote $P_N := \frac{1}{N}\sum_{n=1}^N \delta_{\mathbf{x}_n}$ is the empirical measure. Denote the empirical process for G:

$$\nu_N(G) := \sqrt{N}(P_N - P_{G_0}) \log \frac{\bar{p}_G}{p_{G_0}}.$$

The following results is important in proof below.

Fact 6 (van de Geer, 2000, Theorem 5.11). *Let positive numbers R, C, C_1, a satisfy:*

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

and

$$a \geq \sqrt{C^2(C_1+1)} \left(\int_{a/(2^6\sqrt{N})}^{R} H_B^{1/2} \left(\frac{u}{\sqrt{2}}, \{p_G : G \in \mathcal{O}_K, \mathrm{D}_{\mathrm{h}}^2(p_G, p_{G_0}) \leq R\}, \nu \right) \mathrm{d}u \vee R \right),$$

then

$$\mathbb{P}_{G_0} \left(\sup_{G \in \mathcal{O}_K, \mathrm{D}_{\mathrm{h}}^2(p_G, p_{G_0}) \leq R} |\nu_N(G)| \geq a \right) \leq C \exp\left(-\frac{a^2}{C^2(C_1+1)R^2} \right).$$

Connection between the Hellinger distance and the Wasserstein metric. We introduce the Wasserstein distances to measure the difference between two measures. For two mixing measure $G = \sum_{k=1}^{K} p_k \delta_{\theta_k}$ and $G' = \sum_{\ell=1}^{K'} p'_\ell \delta_{\theta'_\ell}$, the Wasserstein-$r$ distance (for $r \geq 1$) between G and G' is defined as

$$W_r(G, G') := \left(\inf_{\boldsymbol{q} \in \Pi(\boldsymbol{p}, \boldsymbol{p}')} \sum_{k,\ell=1}^{K,K'} q_{k\ell} \left\| \theta_k - \theta'_\ell \right\|^r \right)^{1/r}, \tag{26}$$

where $\Pi(\boldsymbol{p}, \boldsymbol{p}')$ is the set of all couplings between $\boldsymbol{p} = (p_1, \ldots, p_K)$ and $\boldsymbol{p}' = (p'_1, \ldots, p'_{K'})$, i.e, $\Pi(\boldsymbol{p}, \boldsymbol{p}') = \left\{ \boldsymbol{q} \in \mathbb{R}_+^{K \times K'} : \sum_{k=1}^{K} q_{k\ell} = p'_\ell, \sum_{l=1}^{K'} q_{k\ell} = p_k, \forall k \in [K], \ell \in [K'] \right\}$. Fix $G_0 = \sum_{k=1}^{K_0} \pi_k^0 \delta_{\theta_k^0} \in \mathcal{E}_{K_0}$, and consider $G = \sum_{\ell=1}^{K} \pi_\ell \delta_{\theta_\ell}$ such that $W_r(G, G_0) \to 0$, we obtain that

$$W_r^r(G, G_0) \asymp \sum_{k=1}^{K_0} \left(\left| \sum_{\ell \in \mathbb{A}_k(G)} \pi_\ell - \pi_k^0 \right| + \sum_{\ell \in \mathbb{A}_k(G)} \pi_\ell \left\| \theta_\ell - \theta_k^0 \right\|^r \right).$$

Now, we remind Lemma 1 in Ho & Nguyen (2016a).

Fact 7 (Ho & Nguyen, 2016a, Lemma 1). *Let $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$ denote a discrete probability measure and $p_G(x) = \sum_{i=1}^{k} p_i f(x|\theta_i)$ be the mixture density. According to the Lemma 1: Let $G, G' \in \mathcal{O}_k(\Theta)$ such that both $\rho_\phi(p_G, p_{G'})$ and $d_{\rho_\phi}(G, G')$ are finite for some convex function ϕ. Then, $\rho_\phi(p_G, p_{G'}) \leq d_{\rho_\phi}(G, G')$.*

By Fact 7, we can compare the expectation of Hellinger distance between $p_G(y|\boldsymbol{x})$ and $p_{G'}(y|\boldsymbol{x})$ with the Wasserstein metric between G and G' following:

$$\mathbb{E}_{\mathbf{x}}(\mathrm{D}_{\mathrm{h}}^2(p_G(\cdot|\boldsymbol{x}), p_{G'}(\cdot|\boldsymbol{x})) \lesssim W_2(G, G').$$

Now, we are going to prove Theorem 3.

Proof of Theorem 3. Firstly, we recall the empirical average log-likelihood and population average log-likelihood as follows:

$$\bar{\ell}_N(p_G) = \frac{1}{N} \sum_{n=1}^{N} \log p_G(y_n|\mathbf{x}_n) =: P_N \log p_G,$$

$$\mathcal{L}(p_G) = \mathbb{E}_{(\mathbf{x},y)\sim P_{G_0}} \left[\log p_G(y|\boldsymbol{x}) \right] = \int \log p_G(y|\boldsymbol{x}) \, dP_{G_0}(\boldsymbol{x}, y) =: P_{G_0} \log p_G,$$

where $P_N := \frac{1}{N} \sum_{n=1}^{N} \delta_{(\mathbf{x}_n, y_n)}$ is the empirical measure from data, and the joint distribution P_{G_0} over $(\boldsymbol{x}, y)$ is then constructed by first sampling $\boldsymbol{x} \sim P_{\mathbf{x}}$ and then $y|\boldsymbol{x} \sim p_{G_0}(y|\boldsymbol{x})$.

We divide into three cases.

Case 1: $\kappa \geq K_0$. For any G, we denote P_G by the distribution of p_G. By the concavity of log function, we have

$$\frac{1}{2} \log \frac{p_G}{p_{G_0}} \leq \log \frac{p_G + p_{G_0}}{2 p_{G_0}} = \log \frac{\bar{p}_G}{p_{G_0}}, \quad \forall G \in \mathcal{O}_K.$$

Therefore, for all $\kappa > K_0$ we have

$$\begin{aligned}\frac{1}{2}P_N \log \frac{p_{\widehat{G}_N^{(\kappa)}}}{p_{G_0}} &\leq P_N \log \frac{\bar{p}_{\widehat{G}_N^{(\kappa)}}}{p_{G_0}} \\ &= (P_N - P_{G_0}) \log \frac{\bar{p}_{\widehat{G}_N^{(\kappa)}}}{p_{G_0}} - \mathrm{KL}(p_{G_0} \| \bar{p}_{\widehat{G}_N^{(\kappa)}}) \\ &\leq (P_N - P_{G_0}) \log \frac{\bar{p}_{\widehat{G}_N^{(\kappa)}}}{p_{G_0}}.\end{aligned}$$

Hence,

$$\begin{aligned}P_N \log p_{\widehat{G}_N^{(\kappa)}} - P_{G_0} \log p_{G_0} &= P_N \log \frac{p_{\widehat{G}_N^{(\kappa)}}}{p_{G_0}} + (P_N - P_{G_0}) \log p_{G_0} \\ &\leq 2(P_N - P_{G_0}) \log \frac{\bar{p}_{\widehat{G}_N^{(\kappa)}}}{p_{G_0}} + (P_N - P_{G_0}) \log p_{G_0}.\end{aligned}$$

By Theorem 1, we obtain that $\mathrm{D}_{\mathrm{FRA}}(\widehat{G}_N^{(\kappa)}, G_0) \lesssim (\log N/N)^{1/2}$, and obviously we have

$$\inf_{t_0, \boldsymbol{t}_1} W_{\bar{r}(\widehat{G}_N)}^{\bar{r}(\widehat{G}_N)}(\widehat{G}_N^{(\kappa)}, G_{0,t_0,\boldsymbol{t}_1}) \leq \mathrm{D}_{\mathrm{FRA}}(\widehat{G}_N^{(\kappa)}, G_0),$$

where $G_{0,t_0,\boldsymbol{t}_1} = \sum_{k=1}^{K_0} \exp(\omega_{0k}^0 + t_0)\delta_{(\boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1, \boldsymbol{a}_k^0, b_k^0, \sigma_k^0)}$, so there exists a constant D such that

$$\mathbb{P}_{G_0}\left(\inf_{t_0, \boldsymbol{t}_1} W_{\bar{r}(\widehat{G}_N)}(\widehat{G}_N^{(\kappa)}, G_{0,t_0,\boldsymbol{t}_1}) \leq D\left(\frac{\log N}{N}\right)^{1/2\bar{r}(\widehat{G}_N)}\right) \geq 1 - c_1 N^{-c_2}, \quad \kappa \in [K_0, K].$$

Now, we compare Wasserstein metrics W_2 and $W_{\bar{r}(\widehat{G}_N)}$. Since $2/\bar{r}(\widehat{G}_N) \leq 2/4 < 1$, with a note that for a probability $q_{i,j}$, we have $q_{i,j} \leq q_{i,j}^{2/\bar{r}(\widehat{G}_N)}$. Combining with all norms on finite space is equivalent, we obtain that

$$\left(\sum_{i,j} q_{i,j}\|\theta_k - \theta'_\ell\|^2\right)^{1/2} \leq \left(\sum_{i,j} q_{i,j}^{2/\bar{r}(\widehat{G}_N)}\|\theta_k - \theta'_\ell\|^2\right)^{1/2} \lesssim \left(\sum_{i,j} q_{i,j}\|\theta_k - \theta'_\ell\|^{\bar{r}(\widehat{G}_N)}\right)^{1/\bar{r}(\widehat{G}_N)}.$$

Then, we get $W_2 \lesssim W_{\bar{r}(\widehat{G}_N)}$. Using the fact that $\mathbb{E}_{\mathbf{x}}(\mathrm{D}_{\mathrm{h}}^2(p_G(\cdot|\boldsymbol{x}), p_{G'}(\cdot|\boldsymbol{x})) \lesssim W_2(G, G')$ and $p_{G_0} = p_{G_{0,t_0,\boldsymbol{t}_1}}, \forall (t_0, \boldsymbol{t}_1) \in \mathbb{R} \times \mathbb{R}^D$, we also have

$$\begin{aligned}&\mathbb{P}_{G_0}\left(\mathbb{E}(\mathrm{D}_{\mathrm{h}}^2(p_{\widehat{G}_N^{(\kappa)}}(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x})) \leq D\left(\frac{\log N}{N}\right)^{1/2\bar{r}(\widehat{G}_N)}\right) \\ &= \mathbb{P}_{G_0}\left(\inf_{t_0, \boldsymbol{t}_1} \mathbb{E}_{\mathbf{x}}(\mathrm{D}_{\mathrm{h}}^2(p_{\widehat{G}_N^{(\kappa)}}(\cdot|\boldsymbol{x}), p_{G_{0,t_0,\boldsymbol{t}_1}}(\cdot|\boldsymbol{x})) \leq D\left(\frac{\log N}{N}\right)^{1/2\bar{r}(\widehat{G}_N)}\right) \\ &\geq \mathbb{P}_{G_0}\left(\inf_{t_0, \boldsymbol{t}_1} W_{\bar{r}(\widehat{G}_N)}(\widehat{G}_N^{(\kappa)}, G_{0,t_0,\boldsymbol{t}_1}) \leq D\left(\frac{\log N}{N}\right)^{1/2\bar{r}(\widehat{G}_N)}\right) \geq 1 - c_1 N^{-c_2}, \quad \kappa \in [K_0, K].\end{aligned}$$

Let $\mathcal{P}_K(\Theta) := \{p_G(y|\boldsymbol{x}) : G \in \mathcal{O}_K(\Theta)\}$ and $H_B(\varepsilon, \mathcal{P}_K(\Theta), h)$ denotes the bracketing entropy of $\mathcal{P}_K(\Theta)$ under the Hellinger distance. By the Lemma 3 in Nguyen et al. (2023a), there is a constant $C > 0$ such that $H_B(\varepsilon, \mathcal{P}_K(\Theta), h) \lesssim \log(1/\varepsilon)$ for any $0 \leq \varepsilon \leq 1/2$.

Define, $\alpha := 1/2\bar{r}(\widehat{G}_N) \leq 1/4$, substitute $R = D\left(\frac{\log N}{N}\right)^{\alpha}$, $a = D\frac{\log^{\alpha+1/2} N}{N^{\alpha}}$, then for any positive number $\varepsilon < R$, we have $0 \leq \varepsilon \leq 1/e < 1/2$ and $\log(1/\varepsilon) > 1$ for large N enough. Therefore, for large N enough, we obtain that

$a \leq \sqrt{N}R^2 \leq \sqrt{N}R$ and

$$\begin{aligned} a \geq R\left(\log\left(\frac{2^6\sqrt{N}}{a}\right)\right) &\geq \int_{a/(2^6\sqrt{N})}^{R} \log\frac{1}{\varepsilon}d\varepsilon \\ &\geq \int_{a/(2^6\sqrt{N})}^{R} \log^{1/2}\frac{1}{\varepsilon}d\varepsilon \\ &\geq \int_{a/(2^6\sqrt{N})}^{R} H_B^{1/2}(\varepsilon, \mathcal{P}_K(\Theta), h)d\varepsilon \\ &\geq \int_{a/(2^6\sqrt{N})}^{R} H_B^{1/2}(\varepsilon, \{p_G : G \in \mathcal{O}_K(\Theta), \mathbb{E}_{\mathbf{x}}(\mathrm{D}_{\mathrm{h}}^2(p_G(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x})) \leq R\}, \nu)d\varepsilon. \end{aligned}$$

By Fact 6, we get

$$\mathbb{P}_{G_0}\left(\sup_{\mathbb{E}_{\mathbf{x}}(\mathrm{D}_{\mathrm{h}}^2(p_G(\cdot|\boldsymbol{x}),p_{G_0}(\cdot|\boldsymbol{x}))\leq D(\log N/N)^\alpha}\left|\sqrt{N}(P_N - P_{G_0})\log\frac{\bar{p}_G}{p_{G_0}}\right| \geq D\frac{\log^{\alpha+1/2}N}{N^\alpha}\right) \leq N^{-c_2}.$$

Combining with the bound on Hellinger distance, we have

$$\begin{aligned} &\mathbb{P}_{G_0}\left(\left|(P_N - P_{G_0})\log\frac{\bar{p}_{\widehat{G}_N^{(\kappa)}}}{p_{G_0}}\right| \geq D\frac{\log^{\alpha+1/2}N}{N^{\alpha+1/2}}\right) \\ &\leq \mathbb{P}_{G_0}\left(\mathbb{E}_{\mathbf{x}}(\mathrm{D}_{\mathrm{h}}^2(p_{\widehat{G}_N^{(\kappa)}}(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x})) \geq D(\log N/N)^\alpha\right) \\ &\quad + \mathbb{P}_{G_0}\left(\left|(P_N - P_{G_0})\log\frac{\bar{p}_{\widehat{G}_N^{(\kappa)}}}{p_{G_0}}\right| \geq D\frac{\log^{\alpha+1/2}N}{N^{\alpha+1/2}},\ \mathbb{E}_{\mathbf{x}}(\mathrm{D}_{\mathrm{h}}^2(p_{\widehat{G}_N^{(\kappa)}}(\cdot|\boldsymbol{x}), p_{G_0}(\cdot|\boldsymbol{x})) \leq D(\log N/N)^\alpha\right) \\ &\leq c_1 N^{-c_2} + \mathbb{P}_{G_0}\left(\sup_{\mathbb{E}_{\mathbf{x}}(\mathrm{D}_{\mathrm{h}}^2(p_G(\cdot|\boldsymbol{x}),p_{G_0}(\cdot|\boldsymbol{x}))\leq D(\log N/N)^\alpha}\left|\sqrt{N}(P_N - P_{G_0})\log\frac{\bar{p}_G}{p_{G_0}}\right| \geq D\frac{\log^{\alpha+1/2}N}{N^\alpha}\right) \\ &\leq c_1' N^{-c_2}. \end{aligned}$$

For the second term, by the Chebyshev inequality, we have

$$\mathbb{P}_{G_0}(|(P_N - P_{G_0})\log p_{G_0}| \geq t) \leq \frac{\mathrm{Var}(\log p_{G_0})}{Nt^2}. \tag{27}$$

Choose $t = (\log N/N)^\alpha$, we have

$$\mathbb{P}_{G_0}(|(P_N - P_{G_0})\log p_{G_0}| \leq (\log N/N)^\alpha) \geq 1 - c_1 N^{-c_2}.$$

Hence, we conclude that

$$\mathbb{P}_{G_0}\left(\bar{\ell}_N(\widehat{G}_N^{(\kappa)}) - \mathcal{L}(p_{G_0}) \leq \left(\frac{\log N}{N}\right)^{1/2\bar{r}(\widehat{G}_N)}\right) \geq 1 - c_1 N^{-c_2}.$$

Case 2: $\kappa = K_0$. By the Theorem 1, we have

$$\mathrm{D}_{\mathrm{E}}(\widehat{G}_N^{(K_0)}, G_0) \lesssim \left(\frac{\log N}{N}\right)^{1/2}.$$

Assume that $\widehat{G}_N^{(K_0)} = \sum_{k=1}^{K_0}\exp(\omega_{0k}^N)\delta_{(\boldsymbol{\omega}_{1k}^N, \boldsymbol{a}_k^N, b_k^N, \sigma_k^N)}$, since $\mathrm{D}_{\mathrm{E}}(\widehat{G}_N^{(K_0)}, G_0) \to 0$ as $N \to \infty$, the Voronoi cell $\mathbb{A}_k$ has only one element for any $k \in [K_0]$. WLOG, we suppose that $\mathbb{A}_k = \{k\}$ for all $k \in [K_0]$. Moreover, there exist

$t_0 \in \mathbb{R}$ and $\boldsymbol{t}_1 \in \mathbb{R}^D$ independent of N such that $\exp(\omega_{0k}^N) \to \exp(\omega_{0k}^0 + t_0)$ and $\boldsymbol{\omega}_{1k}^N \to \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1$ as $N \to \infty$ for all $k \in [K_0]$. By the definition of $\mathrm{D_E}(\widehat{G}_N^{(K_0)}, G_0)$, we get for large N enough

$$|\exp(\omega_{0k}^N) - \exp(\omega_{0k}^0 + t_0)| \lesssim \left(\frac{\log N}{N}\right)^{1/2}, \quad \|(\Delta_{\boldsymbol{t}_1}\boldsymbol{\omega}_{1k}^N, \Delta\boldsymbol{a}_k^N, \Delta b_k^N, \Delta\sigma_k^N)\| \lesssim \left(\frac{\log N}{N}\right)^{1/2},$$

for every $k \in [K_0]$, where $\Delta_{\boldsymbol{t}_1}^N\boldsymbol{\omega}_{1k}^N := \boldsymbol{\omega}_{1k}^N - \boldsymbol{\omega}_{1k}^0 - \boldsymbol{t}_1$, $\Delta\boldsymbol{a}_k^N := \boldsymbol{a}_k^N - \boldsymbol{a}_k^0$, $\Delta b_k^N := b_k^N - b_k^0$ and $\Delta\sigma_k^N := \sigma_k^N - \sigma_k^0$.

Because the function $f(\boldsymbol{x}, y|\boldsymbol{\theta}) = u(y|\boldsymbol{x}; \boldsymbol{\omega}_1, \boldsymbol{a}, b, \sigma)$ satisfies Condition K (see Lemma 2), let $\epsilon_N = (\log N/N)^{1/2} \to 0$, from condition K, there exist c_α and c_ω such that

$$u(y|\boldsymbol{x}; \boldsymbol{\omega}_{1k}^N, \boldsymbol{a}_k^N, b_k^N, \sigma_k^N) \geq (u(y|\boldsymbol{x}; \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1, \boldsymbol{a}_k^0, b_k^0, \sigma_k^0))^{(1+c_\omega\epsilon_N)} e^{-c_\alpha\epsilon_N}, \quad \forall k \in [K_0].$$

Besides, we can find constant $c_q > 0$ and $c_p > 0$ such that

$$\exp(\omega_{0k}^N) \geq (1 - c_p\epsilon_N)\exp(\omega_{0k}^0 + t_0), \qquad \forall k \in [K_0],$$

$$\sum_{k=1}^{K_0} \exp((\boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1)^\top \boldsymbol{x} + \omega_{0k}^0 + t_0) \geq (1 - c_q\epsilon_N)\sum_{k=1}^{K_0} \exp((\boldsymbol{\omega}_{1k}^N)^\top \boldsymbol{x} + \omega_{0k}^N), \qquad \forall k \in [K_0].$$

Hence, we have

$$\begin{aligned}
&\left[\sum_{k=1}^{K_0} \exp((\boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1)^\top \boldsymbol{x} + \omega_{0k}^0 + t_0)\right] \cdot p_{\widehat{G}_N^{(K_0)}}(y|\boldsymbol{x}) \\
&= \frac{\sum_{k=1}^{K_0} \exp((\boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1)^\top \boldsymbol{x} + \omega_{0k}^0 + t_0)}{\sum_{k=1}^{K_0} \exp((\boldsymbol{\omega}_{1k}^N)^\top \boldsymbol{x} + \omega_{0k}^N)} \cdot \sum_{k=1}^{K_0} \exp(\omega_{0k}^N) u(y|\boldsymbol{x}; \boldsymbol{\omega}_{1k}^N, \boldsymbol{a}_k^N, b_k^N, \sigma_k^N) \\
&\geq (1 - c_q\epsilon)\sum_{k=1}^{K_0} (1 - c_p\epsilon)\exp(\omega_{0k}^0 + t_0)(u(y|\boldsymbol{x}; \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1, \boldsymbol{a}_k^0, b_k^0, \sigma_k^0))^{(1+c_\omega\epsilon_N)} e^{-c_\alpha\epsilon_N}.
\end{aligned}$$

With the fact that $g(t) = t^{1+c_\omega\epsilon_N}$ is a convex function, we get

$$\begin{aligned}
p_{\widehat{G}_N^{(K_0)}}(y|\boldsymbol{x}) &\geq (1 - c_q\epsilon)(1 - c_p\epsilon)\frac{1}{\sum_{k=1}^{K_0} \exp((\boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1)^\top \boldsymbol{x} + \omega_{0k}^0 + t_0)} \\
&\quad \times \sum_{k=1}^{K_0} \exp(\omega_{0k}^0 + t_0)(u(y|\boldsymbol{x}; \boldsymbol{\omega}_{1k}^0 + \boldsymbol{t}_1, \boldsymbol{a}_k^0, b_k^0, \sigma_k^0))^{(1+c_\omega\epsilon_N)} e^{-c_\alpha\epsilon_N} \\
&\geq (1 - c_q\epsilon)(1 - c_p\epsilon)e^{-c_\alpha\epsilon_N} \sum_{k=1}^{K_0} \frac{\exp((\omega_{1k}^0)^\top \boldsymbol{x} + \omega_{0k}^0)}{\sum_{j=1}^{K_0} \exp((\omega_{1j}^0)^\top \boldsymbol{x} + \omega_{0j}^0)} \cdot \mathcal{N}(y|\boldsymbol{a}_k^0\boldsymbol{x} + b_k^0, \sigma_k^0)^{(1+c_\omega\epsilon_N)} \\
&\geq (1 - c_q\epsilon)(1 - c_p\epsilon)e^{-c_\alpha\epsilon_N} p_{G_0}(y|\boldsymbol{x})^{(1+c_\omega\epsilon_N)}.
\end{aligned}$$

Therefore, we have

$$\frac{1}{N}\sum_{i=1}^{N} \log \frac{p_{\widehat{G}_N^{(K_0)}}}{p_{G_0}}(y_i|\boldsymbol{x}_i) \geq \log((1 - c_q\epsilon)(1 - c_p\epsilon)) - (c_\alpha\epsilon_N) + (c_\omega\epsilon_N)\frac{1}{N}\sum_{i=1}^{N} \log p_{G_0}(y_i|\boldsymbol{x}_i).$$

Hence

$$\bar{\ell}(p_{\widehat{G}_N^{(K_0)}}) - \mathcal{L}(p_{G_0}) \geq \log((1 - c_q\epsilon)(1 - c_p\epsilon)) - (c_\alpha\epsilon_N) + (c_\omega\epsilon_N)P_{G_0}\log p_{G_0} + (1 + c_\omega\epsilon_N)(P_N - P_{G_0})\log p_{G_0}. \tag{28}$$

Now, we will bound the right-hand side of above equation, from Chebyshev inequality from eq. (27), choose $t = (\log N/N)^{1/2}$, we get that

$$\mathbb{P}_{G_0}\left(|(P_N - P_{G_0})\log p_{G_0}| \geq \left(\frac{\log N}{N}\right)^{1/2}\right) \leq \frac{\mathrm{Var}(\log p_{G_0})}{\log N}.$$

Obviously, the terms $|\log((1-c_q\epsilon)(1-c_p\epsilon)) - (c_\alpha\epsilon_N) + (c_\omega\epsilon_N)P_{G_0}\log p_{G_0}| \lesssim \epsilon_N = (\log N/N)^{1/2}$, thus there exist a constant $C > 0$ such that $\log((1-c_q\epsilon)(1-c_p\epsilon)) - (c_\alpha\epsilon_N) + (c_\omega\epsilon_N)P_{G_0}\log p_{G_0} \geq -C(\log N/N)^{1/2}$. Then, for some constant $C_e > 0$, we have

$$\mathbb{P}_{G_0}\left(\text{RHS of eq. (28)} \geq -C_e\left(\frac{\log N}{N}\right)^{1/2}\right) \geq 1 - \frac{\mathrm{Var}(\log p_{G_0})}{\log N}.$$

Call the event under above case is B, then we obtain that

$$\begin{aligned}\mathbb{P}_{G_0}\left(\bar{\ell}(p_{\widehat{G}_N^{(K_0)}}) - \mathcal{L}(p_{G_0}) \geq -C_e\left(\frac{\log N}{N}\right)^{1/2}\right) &\geq \mathbb{P}_{G_0}(A_N \cap B) = \mathbb{P}_{G_0}(B) - \mathbb{P}_{G_0}(B \cap A_N^c) \\ &\geq \mathbb{P}_{G_0}(B) - \mathbb{P}_{G_0}(A_N^c) = 1 - \frac{\mathrm{Var}(\log p_{G_0})}{\log N} - c_1 N^{-c_2}\end{aligned}$$

approach 1 when $N \to \infty$, where A_N is defined in Appendix F.2. Therefore, combine both results, we can conclude that

$$|\bar{\ell}(p_{\widehat{G}_N^{(K_0)}}) - \mathcal{L}(p_{G_0})| \lesssim \left(\frac{\log N}{N}\right)^{1/2\bar{r}(\widehat{G}_N)}.$$

Case 3: $\kappa < K_0$. Since $|\log p_G(y|\boldsymbol{x})| \leq m(y|\boldsymbol{x})$ for a measurable function m for all $G \in \mathcal{O}_\kappa$, we can use uniform law of large number to get that

$$\sup_{G\in\mathcal{O}_\kappa} \left|\bar{\ell}_N(G) - P_{G_0}\log p_G\right| \xrightarrow{\mathbb{P}} 0,$$

where $\xrightarrow{\mathbb{P}}$ means convergence in probability. Therefore,

$$\left|\bar{\ell}_N(\widehat{G}_N^{(\kappa)}) - P_{G_0}\log p_{\widehat{G}_N^{(\kappa)}}\right| \xrightarrow{\mathbb{P}} 0.$$

We know that $\log p_{\widehat{G}_N^{(\kappa)}} \to \log p_{G_0^{(\kappa)}}$ in probability, by application of Dominated Convergence theorem, we obtain

$$P_{G_0}\log p_{\widehat{G}_N^{(\kappa)}} \xrightarrow{\mathbb{P}} P_{G_0}\log p_{G_0^{(\kappa)}}.$$

Combining the above results together, we get

$$\bar{\ell}_N(\widehat{G}_N^{(\kappa)}) \xrightarrow{\mathbb{P}} P_{G_0}\log p_{G_0^{(\kappa)}} = \mathcal{L}(\log P_{G_0^{(\kappa)}}).$$

□

Checking condition K. Finally, we check condition K for the function $f(\boldsymbol{x}, y|\boldsymbol{\theta}) := \exp\left(\boldsymbol{\omega}_1^\top\boldsymbol{x}\right)\mathcal{N}\left(y|\boldsymbol{a}^\top\boldsymbol{x}+b,\sigma\right)$.

Lemma 2. *The condition K is satisfied for $f(\boldsymbol{x}, y|\boldsymbol{\theta}) := \exp\left(\boldsymbol{\omega}_1^\top\boldsymbol{x}\right)\mathcal{N}\left(y|\boldsymbol{a}^\top\boldsymbol{x}+b,\sigma\right)$, where $\boldsymbol{\theta} = (\boldsymbol{\omega}_1, \boldsymbol{a}, b, \sigma) \in \mathbb{R}^D \times \mathbb{R}^D \times \mathbb{R} \times \mathbb{R}$ and $\mathcal{X}$ are bounded as from the initial setup, and the eigenvalues of σ are bounded below and above by the positive constants $\sigma_{\min}$ and $\sigma_{\max}$.*

Proof of Lemma 2. When $\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\| \leq \epsilon$ with $\boldsymbol{\theta}^0 = (\boldsymbol{\omega}_1^0, \boldsymbol{a}^0, b^0, \sigma^0)$, by the equivalence of the norm, we can consider the cases where $\left\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_1^0\right\|, \left\|\boldsymbol{a} - \boldsymbol{a}^0\right\|, \left\|b - b^0\right\|, \left\|\sigma - \sigma^0\right\| \leq \epsilon$. We aim to show that for sufficiently small ϵ, there exist $c_\alpha, c_\beta > 0$ such that

$$\log\left(\exp(\boldsymbol{\omega}_1^\top\boldsymbol{x})\mathcal{N}(y|\boldsymbol{a}^\top\boldsymbol{x}+b,\sigma)\right) \geq (1+c_\beta\epsilon)\log\left(\exp((\boldsymbol{\omega}_1^0)^\top\boldsymbol{x})\mathcal{N}(y|(\boldsymbol{a}^0)^\top\boldsymbol{x}+b^0,\sigma^0)\right) - c_\alpha\epsilon.$$

which is equivalent to

$$\begin{aligned}&\left[(1+c_\beta\epsilon)\left(\boldsymbol{\omega}_1^0\right)^\top\boldsymbol{x} - (\boldsymbol{\omega}_1)^\top\boldsymbol{x}\right] + \left[(1+c_\beta\epsilon)\log\left(\left|\sigma^0\right|\right) - \log(|\sigma|)\right] \\ &+\left[(1+c_\beta\epsilon)\left(y - (\boldsymbol{a}^0)^\top\boldsymbol{x} - b^0\right)^\top\left(\sigma^0\right)^{-1}\left(y - (\boldsymbol{a}^0)^\top\boldsymbol{x} - b^0\right) - (y - \boldsymbol{a}^\top\boldsymbol{x} - b)^\top(\sigma)^{-1}(y - \boldsymbol{a}^\top\boldsymbol{x} - b)\right] + c_\alpha\epsilon \geq 0.\end{aligned}$$

Firstly, since $\mathcal{X}$ is bounded, we can omit the term $\left[(1+c_\beta\epsilon)\left(\boldsymbol{\omega}_1^0\right)^\top \boldsymbol{x} - (\boldsymbol{\omega}_1)^\top \boldsymbol{x}\right]$. Next, we note that

$$\frac{\mathrm{d}\log(|\sigma|)}{\mathrm{d}\sigma} = \sigma^{-1}$$

and if $\|\sigma\|$ is bounded above and below far from 0 (which satisfies because σ is positive definite), then the map $\sigma \mapsto \log(|\sigma|)$ is Lipschitz; that is, there exists a constant c_σ such that

$$\left|\log\left(\left|\sigma^0\right|\right) - \log(|\sigma|)\right| \leq c_\sigma \left\|\sigma^0 - \sigma\right\|.$$

Furthermore, we have $|\sigma| \geq \sigma_{\min}$. Hence, for all $c_\beta > \frac{c_\sigma}{\log(\sigma_{\min})}$, then we have

$$c_\beta\epsilon \log\left(\left|\sigma^0\right|\right) \geq c_\sigma\epsilon \geq c_\sigma \left\|\sigma - \sigma^0\right\| \geq \left|\log(|\sigma|) - \log\left(\left|\sigma^0\right|\right)\right|.$$

So that

$$(1+c_\beta\epsilon)\log\left(\left|\sigma^0\right|\right) \geq \log(|\sigma|).$$

We want to choose $c_\alpha > 0$ such that

$$\left[(1+c_\beta\epsilon)\left(y - (\boldsymbol{a}^0)^\top\boldsymbol{x} - b^0\right)^\top \left(\sigma^0\right)^{-1}\left(y - (\boldsymbol{a}^0)^\top\boldsymbol{x} - b^0\right) - (y - \boldsymbol{a}^\top\boldsymbol{x} - b)^\top(\sigma)^{-1}(y - \boldsymbol{a}^\top\boldsymbol{x} - b)\right] + c_\alpha\epsilon \geq 0.$$

Let $u := y - (\boldsymbol{a}^0)^\top\boldsymbol{x} - b^0, \Delta u := (\boldsymbol{a}^0)^\top\boldsymbol{x} + b^0 - [\boldsymbol{a}^\top\boldsymbol{x} + b]$, using the boundedness of σ, there exist c_σ such that

$$(\sigma^0)^{-1} \geq c_\sigma\sigma^{-1}.$$

Hence, we only need to prove

$$(1+c_\beta\epsilon)\, c_\sigma u^\top\sigma^{-1}u - (u+\Delta u)^\top\sigma^{-1}(u+\Delta u) + c_\alpha\epsilon \geq 0,$$

which is equivalent to

$$\begin{aligned} & c_\beta\epsilon c_\sigma u^\top\sigma^{-1}u - u^\top\sigma^{-1}\Delta u - (\Delta u)^\top\sigma^{-1}u - (\Delta u)^\top\sigma^{-1}\Delta u + c_\alpha\epsilon \geq 0 \\ \Leftrightarrow \quad & \epsilon c_\beta c_\sigma \left(u - \frac{\Delta u}{\epsilon c_\beta c_\sigma}\right)^\top \sigma^{-1}\left(u - \frac{\Delta u}{\epsilon c_\beta c_\sigma}\right) + c_\alpha\epsilon \geq \left(1 + \frac{1}{\epsilon c_\beta c_\sigma}\right)(\Delta u)^\top\sigma^{-1}(\Delta u). \end{aligned}$$

We can bound the right-hand side of above equation as follow

$$\left(1 + \frac{1}{\epsilon c_\beta c_\sigma}\right)(\Delta u)^\top\sigma^{-1}(\Delta u) \leq \left(1 + \frac{1}{\epsilon c_\beta c_\sigma}\right)\frac{\|\Delta u\|^2}{\sigma_{\min}} \leq \left(1 + \frac{1}{\epsilon c_\beta c_\sigma}\right)\frac{\epsilon^2}{\sigma_{\min}}.$$

Hence, it is sufficient to choose c_α such that

$$c_\alpha \geq \left(1 + \frac{1}{\epsilon c_\beta c_\sigma}\right)\frac{\epsilon}{\sigma_{\min}} = \frac{\epsilon}{\sigma_{\min}} + \frac{1}{c_\beta c_\sigma \sigma_{\min}}.$$

Then $(1+c_\beta\epsilon)\left(y - (\boldsymbol{a}^0)^\top\boldsymbol{x} - b^0\right)^\top \left(\sigma^0\right)^{-1}\left(y - (\boldsymbol{a}^0)^\top\boldsymbol{x} - b^0\right) - (y - \boldsymbol{a}^\top\boldsymbol{x} - b)^\top(\sigma)^{-1}(y - \boldsymbol{a}^\top\boldsymbol{x} - b) + c_\alpha\epsilon \geq 0$.
Therefore, we complete the proof. □

F.5 Proof of Theorem 4

Define $\mathrm{DSC}_N^{(\kappa)} = -\big(\mathsf{h}_N^{(\kappa)} + \epsilon_N\, \bar{\ell}_N(p_{\widehat{G}_N^{(\kappa)}})\big)$ with $1 \ll \epsilon_N \ll (N/\log N)^{1/(2\bar{r}(\widehat{G}_N))}$ (e.g., $\epsilon_N = \log N$). For $\kappa > K_0$, $\mathsf{h}_N^{(\kappa)}$ shrinks at order $(\log N/N)^{1/\bar{r}(\widehat{G}_N)}$ while the likelihood term cannot compensate at that scale given the chosen ϵ_N, so $\mathrm{DSC}_N^{(\kappa)}$ is suboptimal. For $\kappa < K_0$, the (under-fit) likelihood gap dominates and $\mathrm{DSC}_N^{(\kappa)}$ is worse than at $\kappa = K_0$. Hence $\widehat{K}_N = \arg\min_\kappa \mathrm{DSC}_N^{(\kappa)} \to K_0$ in probability. We will give a more detailed proof below.

Proof of Theorem 4. Note that entropy $H\left(p_{G_0}\right) = -\mathcal{L}\left(p_{G_0}\right)$. We have

$$\mathsf{h}_N^{(\kappa)} = \begin{cases} O\left(\left(\dfrac{\log N}{N}\right)^{1/\bar{r}\left(\widehat{G}_N\right)}\right), & \text{if } \kappa > K_0 \\ \mathsf{h}_0^{(\kappa)} + O\left(\left(\dfrac{\log N}{N}\right)^{1/2}\right), & \text{if } \kappa \leq K_0 \end{cases}$$

and in the proof of Theorem 3, we get

$$\begin{cases} \bar{\ell}_N^{(\kappa)} \leq -H\left(p_{G_0}\right) + O\left(\left(\dfrac{\log N}{N}\right)^{1/2\bar{r}\left(\widehat{G}_N\right)}\right), & \text{if } \kappa > K_0 \\ \bar{\ell}_N^{(\kappa)} = -H\left(p_{G_0}\right) + O\left(\left(\dfrac{\log N}{N}\right)^{1/2\bar{r}\left(\widehat{G}_N\right)}\right), & \text{if } \kappa = K_0 \\ \bar{\ell}_N^{(\kappa)} = -H\left(p_{G_0}\right) - \mathrm{KL}\left(p_{G_0} \| p_{G_0}^{(\kappa)}\right) + o(1), & \text{if } \kappa < K_0 \end{cases}$$

Then we have

$$\begin{cases} \mathrm{DSC}_N^{(\kappa)} \geq \epsilon_N H\left(p_{G_0}\right) + O\left(\epsilon_N \left(\frac{\log N}{N}\right)^{1/2\bar{r}\left(\widehat{G}_N\right)}\right), & \text{if } \kappa > K_0 \\ \mathrm{DSC}_N^{(\kappa)} = \epsilon_N H\left(p_{G_0}\right) - \mathsf{h}_0^{(\kappa)} + O\left(\epsilon_N \left(\frac{\log N}{N}\right)^{1/2\bar{r}\left(\widehat{G}_N\right)}\right), & \text{if } \kappa = K_0 \\ \mathrm{DSC}_N^{(\kappa)} = \epsilon_N H\left(p_{G_0}\right) + \epsilon_N \mathrm{KL}\left(p_{G_0} \| p_{G_0}^{(\kappa)}\right) - \mathsf{h}_0^{(\kappa)} + o\left(\epsilon_N\right), & \text{if } \kappa < K_0 \end{cases}$$

Since $\epsilon_N \to \infty, \epsilon_N (\log N/N)^{1/2} \bar{r}\left(\widehat{G}_N\right) \to 0$ and $\mathrm{KL}\left(p_{G_0} \| p_{G_0}^{(\kappa)}\right) > 0$, then as $N \to \infty, \mathrm{DSC}_N^{K_0}$ is the smallest number. Hence, $\mathbb{P}_{p_{G_0}}\left(\widehat{K}_N = K_0\right) \geq \mathbb{P}_{p_{G_0}}\left(A_N\right) \to 1$ as $N \to \infty$, or $\widehat{K}_N \to K_0$ in probability. □